D6.3

Second generation of Human- and Society-centered AI algorithms

Project Title	AI4Media - A European Excellence Centre for Media, Society and Democracy
Contract No.	951911
Instrument	Research and Innovation Action
Thematic Priority	H2020-EU.2.1.1. - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial technologies - Information and Communication Technologies (ICT) / ICT-48-2020 - Towards a vibrant European network of AI excellence centres
Start of Project	1 September 2020
Duration	48 months

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

Deliverable title	Second generation of Human- and Society-centered AI algorithms
Deliverable number	D6.3
Deliverable version	1.0
Previous version(s)	-
Contractual date of delivery	August 31, 2023
Actual date of delivery	September 22, 2023
Deliverable filename	AI4Media_D6_3.pdf
Nature of deliverable	Report
Dissemination level	Public
Number of pages	182
Work Package	WP6
Task(s)	T6.2-T6.7
Partner responsible	UPB
Editor	Mihai Gabriel Constantin
Officer	Evangelia Markidou

Abstract	This document presents the intermediate outcomes of the research on human- and society-centered AI algorithms, reporting the progress in tasks T6.2, T6.3, T6.4, T6.5, T6.6, and T6.7 in the period M17-M36. Specifically, this document builds upon and presents the updates of the work presented in deliverable D6.1. For each task we present the technological and research advances, as well as relevant publications and published software, datasets, repositories, or other resources. Finally, we discuss the limitations of current technologies, as well as our plans for further development and research.
Keywords	Artificial intelligence, media, content moderation, synthetic content creation, synthetic content detection, deepfake, manipulation detection, online political debate, privacy-aware recommendation, private content sharing, hyper-local news, user perception measurement

Copyright

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 951911

Authors & Contributors

Name	Organization
Niki Maria Foteinopoulou	QMUL
Davide Alessandro Coccomini	CNR
Fabrizio Falchi	CNR
Claudio Gennaro	CNR
Giuseppe Amato	CNR
Roberto Caldelli	UNIFI
Luca Cuccovillo	FhG
Thomas Köllmer	FhG
Ioanna Koroni	AUTH
Mihai Gabriel Constantin	UPB
Cristian Stanciu	UPB
Mathias-Felipe de-Lima-Santos	UvA
Anna Schjøtt Hansen	UvA
Tobias Blanke	UvA
Natali Helberger	UvA
Elisa Ricci	UNITN
Nicu Sebe	UNITN
Symeon Papadopoulos	CERTH
Yiannis Kompatsiaris	CERTH
Nikos Giatsoglou	CERTH
Christoforos Papastergiopoulos	CERTH
Daniel Gatica-Perez	IDIAP
David Alonso del Barrio	IDIAP
Victor Bros	IDIAP
Haeeun Kim	IDIAP
Pierpaolo Goffredo	UCA
Mariana Chaves	UCA
Serena Villata	UCA

Peer Reviews

Name	Organization
Rasa Bočytė	NISV
Silvi Rusi	UM

Revision History

Version	Date	Reviewer	Modifications
0.1	16.06.2023	Mihai Gabriel Constantin	First draft with contributions from all partners
0.2	15.08.2023	Mihai Gabriel Constantin	Second draft with contributions from all partners
0.3	16.08.2023	Mihai Gabriel Constantin	Draft sent to internal reviewers
0.4	08.09.2023	Mihai Gabriel Constantin	Updated version based on internal reviews
1.0	22.09.2023	Mihai Gabriel Constantin	Final version

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

Table of Abbreviations and Acronyms

Abbreviation	Meaning
1D	Uni-dimensional
2D	Two-dimensional
3D	Three-dimensional
3DCNN	Three-Dimensional Convolutional Neural Network
Acc	Accuracy
ADD	Average Detection Distance
AE	Auto-Encoder
AI	Artificial Intelligence
API	Application Programming Interface
ARI	Automated Readability Index
ARs	Assessment Reports
ASR	Automatic Speech Recognition
ASV	Automatic Speaker Verification
AUC	Area Under the Curve
Audio DAE	Audio Denoising Autoencoder
BCE	Binary Cross Entropy
BiLSTM	Bidirectional Long Short-Term Memory
CAFE	CoArse-to-FinE
CapsNet	Capsule Networks
CFM	Coarse-to-Fine Module
CLI	Coleman-Liau Index
CLS	Conditional Latent Space
CM	CounterMeasure system
CNN	Convolutional Neural Network
CPRA	California Privacy Rights Act
CRF	Conditional Random Field
DnCNN	Denoising CNN
DNN	Deep Neural Network
DSA	Digital Services Act
DSP	Digital Signal Processing
EC	European Commission
EER	Equal Error Rate
EKG-RM	Explainable Knowledge Graph-based Recommendation Model
EU	European Union
F-K GL	Flesch-Kincaid Grade level
F1	F1-score
FAR	False Acceptance Rate

Abbreviation	Meaning
FFN	Feed-Forward Networks
FFT	Fast Fourier Transform
FG	Feature Generation
FID	Fréchet Inception Distance
FL	Federated Learning
FoR	Fake or Real
FPS	Frames Per Second
FRE	Flesch Reading Ease
FRR	False Rejection Rate
FSM	Face Stacked Manipulation
FVD	Fréchet Video Distance
GAM	Gaze Animation Module
GAN	Generative Adversarial Network
GCM	Gaze Correction Module
GDPR	General Data Protection Regulation
GFI	Gunning-Fog Index
GNeRF	Generative Neural Radiance Field
GNN	Graph Neural Network
IPCC	Intergovernmental Panel on Climate Change
K-L	Kullback-Leibler divergence
KD	Knowledge Distillation
KG	Knowledge Graph
LLM	Large Language Model
LPIPS	Learned Perceptual Image Patch Similarity
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MCA	Microphone Classification Accuracy
MDR	Missing Detection Rate
MFCC	Mel-Frequency Cepstral Coefficients
MINTIME	Multi-Identity size-iNvariant TIMEsformer
MLM	Masked Language Modeling
MLP	MultiLayer Perceptron
MSE	Mean Squared Error
MSSSIM	Multi-Scale Structural Similarity
MTCNN	Multi-Task Cascaded Convolutional Neural Networks
NDCG	Normalized Discounted Cumulative Gain
NeRF	Neural Radiance Field
NLP	Natural Language Processing
NPMI	Normalized Pointwise Mutual Information
OSN	Online Social Network

Abbreviation	Meaning
P	Precision
PCM	Pulse-Code Modulation
PE	Playable Environment
PGPR	Policy-Guided Path Reasoning
PPE	Predict, Prevent, and Evaluate
PSNR	Peak Signal-to-Noise Ratio
PVG	Playable Video Generation
PVM	Predicting Video Memorability
QC	Quality Calculation
R	Recall
RBO	Ranked-Based Overlap
RL	Reinforcement Learning
RMSE	Root-Mean-Square Error
ROC curve	Receiver Operating Characteristic curve
SGAN	Semantic-Guided Generative model
SGD	Stochastic Gradient Descent
SLC	Sentence-Level Classification
SNR	Signal to Noise Ratio
SOTA	State-Of-The-Art
SSH	Social Sciences and Humanities
SSIM	Structural Similarity Index Measure
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
t-DCF	t-Detection Cost Function
TCPE	Temporal-Coherent Positional Embedding
TSC	Target-dependent Sentiment Classification
TTS	Text-to-Speech
UC	Use Case
URL	Uniform Resource Locator
V	Vocoding
VAD	Valence, Arousal, Domination
VAE	Variational Auto-Encoder
VC	Voice Conversion
ViT	Vision Transformer
WAV	WAVeform audio file format
WP	Work Package
XAI	eXplainable AI

Contents

List of Tables

List of Figures

1 Executive summary

This deliverable shows the current status of Work Package 6 (WP6): *Human and Society-centered AI*, presenting the results attained and the work published between months 17 (January 2022) and 36 (August 2023) of the AI4Media project. WP6 has an integral role in the AI4Media project, creating methods and algorithms that can then be exploited by the use cases developed in WP8. This deliverable presents progress in Tasks 6.2-6.7, covering a number of different topics, ranging from deepfake detection to social media analysis to user perception of media and more.

Sections 3 and 4 cover Task 6.2: *Manipulation and synthetic content detection in multimedia* from two perspectives. The first presents our work in synthetic content creation, including work in gaze correction, 3D-aware generative models, playable environments, and text-driven image manipulation (Section 3). The second perspective deals with synthetic and manipulated content detection, presenting techniques for different modalities, i.e. for video- and image-based detection, audio-based detection, and text-based detection (Section 4).

Section 5 presents the work covered in Task 6.3: *Hybrid, privacy-enhanced recommendation*, describing a knowledge graph-based method for news recommendation that aims to provide personalized news recommendations and tries to explain why an item is recommended to a particular user.

Section 6 describes the work done in Task 6.4: *AI for Healthier Political Debate*, focusing on the development of AI models to analyse political debate. The presented techniques explore a variety of topics, including sentiment analysis for public opinion polling and detection of politically charged tweets, classification and detection of special patterns like fallacious arguments, propaganda messages, and political debate concepts, as well as the study of subjective and objective understanding of political news and a preliminary study on the ephemerality of discourse during the COVID pandemic.

Section 7 corresponds to Task 6.5: *Perceptions of hyper-local news*, targeting the analysis of local news and the understanding of their perception both by people and machines. This section presents work related to the analysis of health news, European news, anti-vax news, as well as an analysis of French-speaking local news and referencing in YouTube videos for knowledge communication.

Progress in Task 6.6: *Measuring and predicting user perception of social media* is presented in Section 8. The works presented here include methods and analyses for several perception-related concepts like video memorability, media content affective effects, as well as the perception of disinformation echo-chambers.

Section 9 presents the work done in Task 6.7: *Real-life effects of private content sharing*, presenting AI models and techniques for predicting the effect of sharing private data like photos in social media, considering a diverse set of real-life scenarios like applying for a job or a loan.

Finally, Section 10 presents additional work done in WP6, consisting of several activities related to the organization of special issues in journals, workshops, and benchmarking initiatives.

The impact of our work in WP6 is reflected by the fact that a large number of the works presented in D6.3 have resulted in publications in prestigious international journals and conferences in the field. More specifically, work in WP6 during this period has resulted in more than 20 publications, which have been uploaded to open repositories. Beyond this, and to increase the impact in the field, several of the works provide open software or datasets. We make explicit references to the corresponding publications and/or software/datasets provided by each partner and we also establish connections of the presented work with the WP8 Use Cases and media-related applications in general.

2 Introduction

The goal of WP6 is to investigate the societal impact of AI technologies with focus on topics which are important for the classical and social media, such as: synthetic content generation and detection, content recommendation, analysis of the political debate, perception of local news, analysis of the perception of social media by users, and effects of private content sharing. Below, we briefly discuss the main challenges which are addressed in relation to each of these research topics.

Generative AI technologies make it possible to produce synthetic content at unprecedented scale. Such content can be used as a complement to or instead of real training data whenever the later are limited or unavailable, with applications in domains such as medicine, security, and the media. Advances in generative AI are of utmost importance for media professionals since they can have a strong impact on future journalism. For instance, they can be used to assist journalists with text writing and summarisation, or to create suitable illustrations for multimedia content. However, the same techniques can be deployed by malevolent entities to create deepfakes and fuel disinformation campaigns. It is consequently important for media professionals, but also for the general public, to have **synthetic content detection tools** available in order to counter such disinformation campaigns in an efficient way. The current contributions of AI4Media regarding synthetic data creation and data manipulation are presented in Section 3, while the automatic detection of such data is discussed in Section 4. In total, seventeen contributions were accepted to peer-reviewed conferences, seven were accepted to peer-reviewed journals, and one is currently under review.

Recommender systems shape the way users access online content since they are a core component of online social networks and of news aggregators. Their functioning is most often opaque and it is important to introduce explanation mechanisms which give users feedback about why content is recommended to them. The current contributions of AI4Media regarding privacy-enhanced recommenders are discussed in Section 5. They have resulted in a conference paper for explainable personalized news recommendations.

A **healthy political debate** is a needed for the good functioning of democratic societies. The advent of powerful AI technologies comes with both opportunities and challenges related to political debates. On the one hand, they lead to unprecedented challenges associated with the spread of misinformation, the occurrence of echo chambers, the creation of biased and unfair representations of political events, and the polarization of political discussions. These phenomena have deleterious effects in society, and media professionals need analysis tools to understand and counter them. On the other hand, AI technologies can be used for a fine grained understanding of the large amounts of content produced during political debates. For instance, sentiment analysis can be deployed to characterize the subjective part of political debates, fallacious argument classification can expose problematic parts of the political discourse, and temporal statistical analysis can be used to analyse the dynamics of the debate around impactful topics, such as COVID-19. The contributions of AI4Media regarding the use of AI to promote a healthier political debate are discussed in Section 6. In total, three works were accepted to peer-reviewed conferences, one was accepted to a peer-reviewed journal, and four are currently under review.

News play an important role in shaping our understanding of the world. While there are many research efforts which focus on automatic news analysis, local news are currently understudied. This is problematic insofar as local news provide up-to-date critical information about communities' life. Their analysis in Europe is challenging because local news are produced by hundreds of sources in many languages. Equally important, the language used in news varies with the domain and the type of media which publishes them. The advent of large language models, which cover different languages and can be easily fine-tuned for specific tasks, offers the opportunity to analyse local news along different dimensions which are important for professional and general public stakeholders. The current contributions of AI4Media regarding the **perception of hyper-local news** are presented in Section 6. In total, five contributions were accepted to peer-reviewed conferences while several datasets were created.

The **measurement and prediction of the perception of social media content** by users is a very challenging area of research. Challenges arise from the limited amount of annotated data, the subjective nature of individual perceptions, the evolving perception of the same content over time, and the increasing difficulty of accessing social media data for research purposes. To address the this research area properly, inputs from different disciplines including computer vision, linguistics, psychology, and sociology are required. They were integrated in the proposed works which address content memorability, human affect analysis and echo chambers on social media. Media professionals can use the proposed components to increase the impact of the content they produce, and to counter coordinated disinformation campaigns. The current contributions of AI4Media regarding the measuring and the prediction of user perception of social media are presented in Section 8. Three contributions were accepted to peer-reviewed conferences and three chapters were accepted in peer-reviewed books.

Social media incentivise their users to share data in exchange for free access to their services. The success of this business model is conditioned by the level of detail of user profiles which can be derived from personal data shared online. The advent of powerful AI technologies which automate the analysis of users' behaviour and of their multimedia content led to the ability to create very detailed profiles. However, this progress in profiling has not been accompanied by sufficient efforts to make the process more explainable. Users are entitled to understand how their data could be used by OSNs and associated third parties in contexts which are unforeseen at sharing time. Since direct access to OSN data is very difficult, a modelling of impactful situations is proposed to provide feedback about the **potential effects of data sharing**. The current contributions of AI4Media in this area are discussed in Section 9. They led to one publication in a peer-reviewed conference.

A part of the contributions proposed during the second period build on and extend the works discussed in D6.1, but also encompass new areas of interest covered by WP6. For instance, partners continued their work on the detection of manipulated content for video and audio modalities, but also proposed a new method for detecting synthetic short texts. The analysis of political debate was enriched with new dimensions, for instance related to fallacious argument classification, propagandist message detection, or predicting political debates in complex settings.

A part of WP6 contributions are standalone, while others are driven by upstream work done in other WPs of the project. For instance, robustness tools proposed in WP4 and language analysis components from WP5 were used in tasks T6.2 and T6.4, respectively. Downstream, the components are offered for integration in project use cases. For instance, tools for manipulated content detection (T6.2) are already integrated in UC1, while components which analyse the political debate (T6.4) are currently integrated in UC2 and UC4.

To summarize, the works presented in this deliverable highlight the role of AI technologies in society, but also question their societal implications. They propose an innovative take at hot topics, such as deepfakes, the analysis of political debates, or the perception of social media. These contributions led to a significant number of publications in leading journals and conferences in the respective areas of research.

3 Manipulation and synthetic content detection in multimedia (T6.2) - Data creation and manipulation

Contributing partners: CERTH, CEA, FhG-IDMT, CNR, QMUL, UPB, UNIFI, UNITN

The topic of manipulated or synthetically generated multimodal content is of paramount importance in today's news and social media environments. Synthetic data creation with the help of generative neural networks are applied to data augmentation for training DNNs, in domains where enough data may not be available, like medicine, security, and anomaly detection AI models, as well as contributing to the creation of artificial personal assistants, drug creation, and language translation models. However, some of the concerns regarding the use are related to their use in the creation and spreading of deepfakes, fake news and misinformation, bias and discrimination, and intellectual property issues [1], [2]. This Section describes the work done in AI4Media in content creation and manipulation, as part of Task 6.2 - *Manipulation and synthetic content detection in multimedia*. This Section analyzes the first part of this Task, focusing specifically on new content generation methods, while the following Section 4 will present new techniques for manipulated or synthetic content detection.

Section 3.1 presents work done in high-resolution portrait gaze correction and animation, aiming to manipulate the gaze direction of a face image with respect to a specific target direction. Section 3.2 proposes a two-step 3D-aware human generation process that produces high-quality realistic images of human bodies and faces. A set of text-driven image manipulation methods are presented in Section 3.3 that allow producing facial images with specific facial characteristics. Finally, a method for spatio-temporal video manipulation is presented in Section 3.4, which enables 3D- and action-aware video editing, camera trajectory manipulation, changing the action sequence, the agents and their styles, or continuing the video in time beyond the observed footage.

3.1 Unsupervised High-Resolution Portrait Gaze Correction and Animation

Contributing partner: UNITN

Gaze correction is manipulating the gaze direction of a face image with respect to a specific target direction. The main application of this task is altering the eye appearance so that the person's gaze is directed into the camera. For example, shooting a good portrait is challenging as the subjects may be too nervous to stare at the camera. Another scenario is videoconferencing, where eye contact is very important. The gaze can express attributes such as attentiveness and confidence. Unfortunately, eye contact is frequently lost during a video conference, as the participants look at the monitors and not directly into the camera. Moreover, some works use gaze redirection to improve few-shot gaze estimation task [3], [4].

In this research, we extend GazeGAN [5] to work also with higher resolution portrait images. Specifically, we first create a new dataset, CelebHQGaze, containing 512 × 512 high-resolution portrait images. Second, we propose a novel Gaze Correction Module (GCM) and a Gaze Animation Module (GAM) integrated with a coarse-to-fine module (CFM) (see Fig. 1). In more detail, CFM first allows the inpainting model to be trained using low-resolution images for coarse-grained image generation. Then it uses a global nonparametric model, Laplacian Reconstruction, and a local parametric model, Local-Refinement Autoencoder, to compensate for the high-frequency information loss and to remove possible artifacts for the eye region. Utilizing this new architecture, we can avoid training each module using high-resolution images. CFM speeds up both the training and the inference process while obtaining high-quality results, comparable with directly training with high-resolution images.

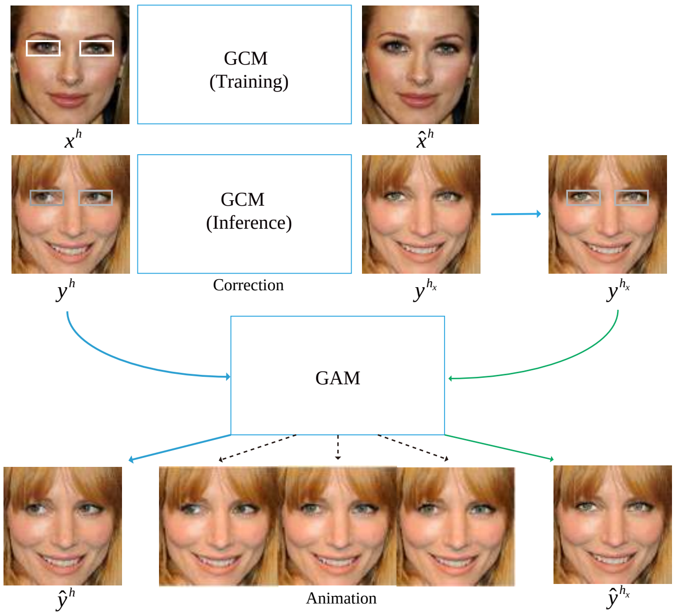


Figure 1. Overview of the proposed architecture. We have two main modules: Gaze Correction Module for performing gaze correction (GCM) and Gaze Animation Module for performing gaze animation (GAM). Moreover, we propose to use the gaze-corrected samples from GCM to train GAM (Synthesis-as-Training). The trained GAM can achieve gaze animation by interpolating the latent feature. The white boxes are the eye mask to remove the eye region. The gray boxes represent the cropping of the eye region.

Similar to GazeGAN [5], an autoencoder is pretrained using self-supervised mirror learning (PAM), where the bottleneck features are used as an extra input to the dual inpainting model to preserve the identity of the corrected results. Moreover, global and local discriminators are used to improve the visual quality of the generated samples. Finally, our qualitative and quantitative evaluations show that our method generates higher-quality results with respect to the state-of-the-arts in both the gaze correction and the gaze animation tasks.

3.1.1 Experiments

This section introduces the details of our datasets, our network training, and baseline models. Then, we compare the proposed method with the state-of-the-art methods of gaze correction in the wild using both qualitative and quantitative evaluations. Next, we demonstrate the effectiveness of the proposed method on gaze animation with various outputs by interpolating and extrapolating in the latent space. For brevity, we refer to the method presented in [5] as **GazeGAN** and the extended version introduced in this work as **GazeGANV2**. Note that we do not use any post-processing step for GazeGAN and GazeGANV2.

3.1.1.1 Datasets Most of the existing benchmarks [9]–[12] do not contain enough image variability (e.g., a wide gaze-direction range, various head poses, and different illumination conditions) for our gaze correction task in the wild. Recently, [13] presented a large-scale gaze tracking dataset, called Gaze360, for robust 3D gaze estimation in unconstrained images. Although this dataset has been labeled with a 3D gaze direction with a wide range of angles and head poses, it still lacks high-resolution images for face and eye regions. Moreover, this dataset does not provide annotations of the eye gaze staring at the camera, which is required in our domain set X. More recently, [14] proposed a large scale (over 1 million samples) of high-resolution images for gaze estimation.

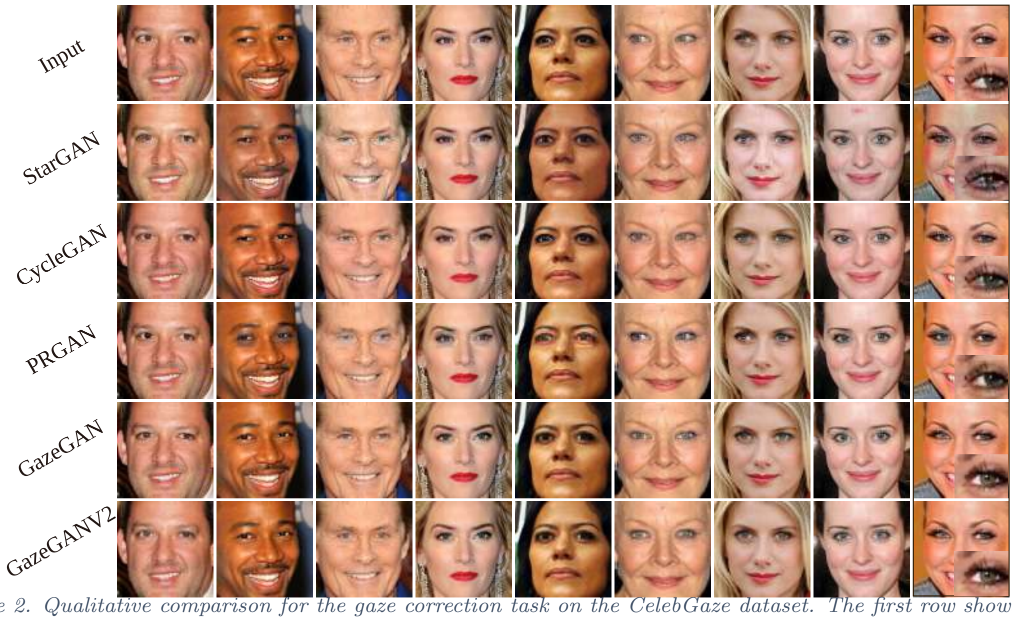


Figure 2. Qualitative comparison for the gaze correction task on the CelebGaze dataset. The first row shows the input images, and the following rows show the gaze correction results of StarGAN [6], CycleGAN [7], PRGAN [8], GazeGAN and GazeGANV2. Magnified left eyes are shown in the last column. Zoom in for the best of view.

However, these images are collected in laboratory conditions and are not suitable for our gaze correction task in the wild. To remedy this problem, we propose collecting new datasets consisting of lots of high-resolution portraits without labelling head poses and gaze information. In detail, five volunteers are asked to divide the row data (face) into two domains according to whether the face eyes are staring at the camera. The gaze and head estimation model can automate 'Staring at the camera' annotation. However, the existing state-of-the-art methods [13], [15] cannot achieve accurate gaze estimation for CelebHQGaze, as an overlarge domain shift exists between training data and test data.

CelebGaze. CelebGaze consists of 25,283 celebrity images, most of which have been collected from CelebA [16] and a minority from the Internet. There are 21,832 face images with the eyes staring at the camera and 3,451 face images with the eyes looking somewhere else. We crop all the images to 256 × 256 and compute the eye mask region using Dlib [17]. Specifically, we use Dlib to extract 68 facial landmarks, and we compute the mean of 6 points near the eye region, which is the center point of the mask. The size of the mask is fixed to 30 × 50. We randomly select 300 samples from domain Y and 100 samples from domain X as their corresponding test sets, and we use the remaining images for the training set. Note that this dataset is unpaired and not labeled with the specific eye angle or the head pose information.

CelebHQGaze. CelebHQGaze consists of 29,255 high-resolution celebrity images that are collected from CelebA-HQ [18]. It consists of 21,005 face images with the eyes staring at the camera and 8,250 face images with eyes looking somewhere else. Similarly to CelebGaze, we extract facial landmarks and generate the mask. All images are cropped to 512 × 512, and the mask size is fixed to 46 × 80. Similar to the CelebGaze dataset, also for the CelebHQGaze, we randomly select 300 samples from domain Y and 100 samples from domain X for the test set, and we use all the remaining images for the training set.

3.1.1.2 Baseline Models

Gaze Correction. PRGAN [8] achieved state-of-the-art gaze redirection results on the Columbia

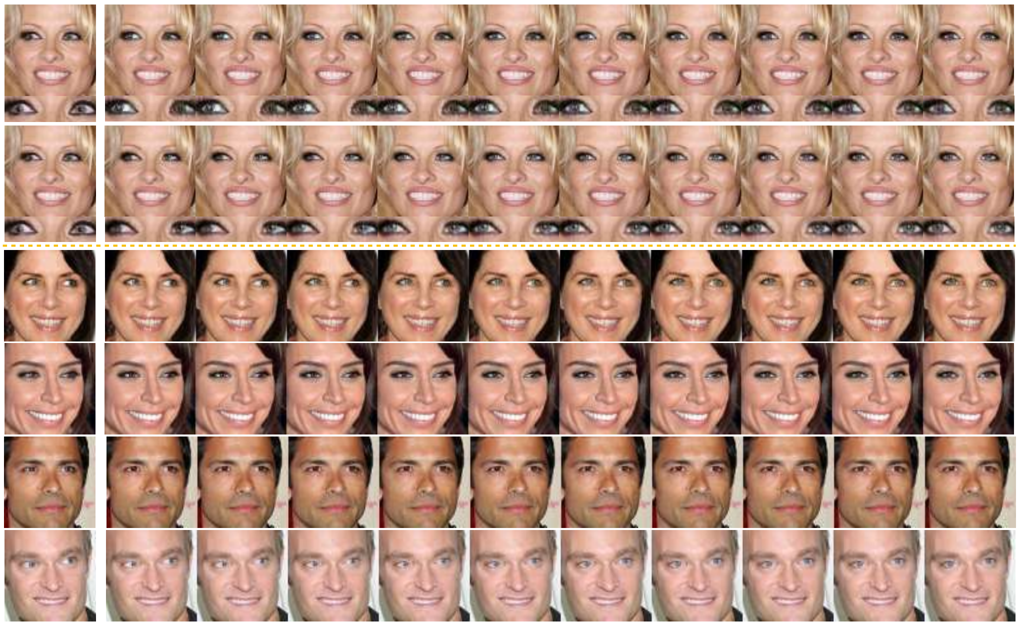


Figure 3. Gaze animation results using the interpolation of the latent features r on the CelebGaze dataset. The top two rows show the images generated by GazeGAN and GazeGANV2, respectively, jointly with the eye regions. The other rows show the gaze animation results of GazeGANV2. The first and the last columns show the input images and the gaze-corrected results, respectively. The middle columns show the interpolated images.

Method	CelebGaze						CelebHQGaze					
	MSSSIM ↑	LPIPS ↓	FID ↓	US ↑	Params ↓	FPS ↓	MSSSIM ↑	LPIPS ↓	FID ↓	US ↑	Params ↓	FPS ↓
Other	-	-	-	24.20%	-	-	-	-	-	23.20%	-	-
StarGAN [6]	0.96	0.073	82.49	3.400%	-	-	0.94	0.084	185.47	4.400%	-	-
CycleGAN [7]	0.99	0.026	70.12	15.00%	-	-	0.98	0.028	**53.690**	8.670%	-	-
PRGAN [8]	1.00	0.000	84.61	8.330%	-	-	1.00	0.000	106.79	22.40%	-	-
GazeGAN	**1.00**	**0.000**	62.12	22.40%	73.26M	30.29	**1.00**	**0.000**	60.520	25.50%	183.2M	23.20
GazeGANV2	**1.00**	**0.000**	**56.37**	**32.40%**	**47.20M**	**38.40**	**1.00**	**0.000**	63.590	**27.30%**	**84.18M**	**27.70**
GT	1.00	0.000	-	100%	-	-	1.00	0.000	-	100%	-	-

Table 2. Quantitative results on both the CelebGaze and the CelebHQGaze dataset. The higher is better for MSSSIM and the user study; the lower is better for LPIPS and FID. The columns Params and FPS report the corresponding network parameters and frame per second at test time, respectively. US: user studies.

gaze dataset [10] based on a single encoder-decoder network with adversarial learning, similarly to the StarGAN architecture [6]. The original PRGAN is trained on paired samples with labeled angles. To train PRGAN on the proposed CelebGaze and CelebHQGaze datasets, we remove the VGG perceptual loss of PRGAN, and learn the gaze redirection task between domain X and Y. We train PRGAN only with the local eye region, the same way as the original paper.

Facial Attribute Manipulation. Gaze correction and animation can be regarded as a subtask of facial attribute manipulation. Recently, StarGAN [6] achieved very high-quality results in facial attribute manipulation. We train StarGAN on the CelebGaze dataset to learn the translation mapping between domain X and domain Y.

Moreover, gaze correction can be considered as an image translation task. Thus, we adopt CycleGAN as another baseline for our experiments. Note that we do not compare GazeGAN with AttGAN [19], STGAN [20], RelGAN [21], CAFE-GAN [22], SSCGAN [23] as they have a performance very close to StarGAN in the facial attribute manipulation task. We use the public code of StarGAN [1], CycleGAN [2] and PRGAN [3].

3.1.1.3 Gaze Correction

This section qualitatively and quantitatively compares the proposed method with state-of-the-arts on both CelebGaze and CelebHQGaze datasets for the gaze correction task.

Qualitative Results. As shown in the last row of Fig. 2, GazeGANV2 can correct the eyes to look at the camera while preserving the identity information such as the eye shape and the iris color, validating the effectiveness of the proposed method. The 2nd row of the figure shows the results of StarGAN [6]. We note that StarGAN could not produce precise gazes staring at the camera, and it suffers from a low-quality generation with lots of artifacts (Zoom in for the best of view). The results of CycleGAN are shown in the 3rd row. Although the results of CycleGAN are very realistic and with few artifacts in the eye region, this method does not produce a precise correction of the gaze direction (e.g., see the magnified eye regions of Fig. 2). We explain that both StarGAN and CycleGAN use the cycle-consistency loss, which requires that the mapping between X and Y be continuous and invertible. According to the invariance of the Domain Theorem[4], the intrinsic dimensions of the two domains should be the same. However, the intrinsic dimension of Y is much larger than X, as Y has more variations for the gaze angle than X. Moreover, we compare GazeGANV2 with PRGAN [8]. PRGAN is trained using only local eye regions (same as in the original paper), which may help focus on the translation of the eye region. The results of PRGAN are shown in the 4th row of Fig. 2. Compared with GazeGANV2, PRGAN does not produce precise and realistic correction results. Additionally, PRGAN suffers from the boundary mismatch problem between the local eye region and the global face.

Finally, as shown in the last row of Fig. 2, comparing GazeGANV2 with GazeGAN, we observe that both models can produce realistic and faithful results.

Quantitative Evaluation Protocol. The qualitative evaluation has validated the effectiveness and the superiority of our proposed GazeGANV2 in the gaze correction task. To further support the previous evaluation with quantitative results, we use the MSSSIM [24] and the LPIPS [25] metrics to measure the preservation ability of the *irrelevant regions*, i.e.,, the whole image except the eye region ($M(y^h)$). Specifically, we compute the mean MSSSIM and LPIPS scores between $M(y^h)$ and $M(\hat{y}^h)$ across all the *test* data of Y^h. Moreover, the Fréchet Inception Distance (FID) [26] has been shown to correlate well with the human judgment and has become a popular metric for GAN-based methods. We use it to evaluate the quality of the generated eye region for the gaze correction and the gaze animation tasks.

In addition to the aforementioned automatic metrics, we conduct a user study to compare the results of the gaze correction task of different models. In detail, given an input face image of the CelebGaze or CelebHQGaze test dataset (extracted from Y), we show the gaze-corrected results produced by different models to 30 respondents, who were asked to select the best image based on the perceptual realism and the precision of the gaze correction. They also can select “Other”, which means that the results of all the models are not satisfactory enough. This study is based on 50 questions (i.e., 50 randomly sampled images) for each respondent.

Quantitative Results. The first two columns of the left part (CelebGaze) and the right part (CelebHQGaze) of Table 2 show the MSSSIM and LPIPS scores evaluating the preservation ability of the corrected images using different methods. GazeGANV2 and PRGAN obtain the best results, with 1.0 for MSSSIM and 0.0 for LPIPS. The original irrelevant regions are integrated with the generated eye region in both models using binary masks. StarGAN and CycleGAN get

[1] https://github.com/yunjey/StarGAN

[2] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

[3] https://github.com/HzDmS/gaze_redirection

[4] https://en.wikipedia.org/wiki/Invariance_of_domain

the worse irrelevant region preservation scores. The FID scores of the eye regions are reported in the 3rd column. In the CelebGaze dataset, GazeGANV2 and GazeGAN outperform all the other methods, reaching comparable scores on the CelebHQGaze dataset. Though CycleGAN has the best FID scores, it fails to generate precise gaze correction results. The penultimate column of both parts in Table 2 shows the evaluation results of the user study. For the CelebGaze dataset, the average vote for GazeGANV2 is higher than for all the other methods. The same conclusions can be drawn with the CelebHQGaze dataset. Importantly, GazeGANV2 achieves a performance very close to GazeGAN. However, it has fewer parameters and higher FPS, as shown in the last two columns of Table 2. Overall, the qualitative and quantitative evaluations demonstrate the effectiveness and superiority of our approach.

3.1.1.4 Gaze Animation The bottom row of Fig. 3 shows gaze animation results using input images with various gaze directions. The latent-space interpolation results are smooth and plausible in each row. Each column has a different gaze direction angle, but the identity information is overall preserved (e.g., the eye shape, the iris color, etc.).

The top row of Fig. 3 shows the gaze animation comparison between GazeGAN and GazeGANV2. GazeGANV2 can produce more realistic images with fewer artifacts than GazeGAN on the CelebGaze dataset, while they have comparable performance on the CelebHQGaze dataset.

3.1.2 Conclusions

In this research we introduce an unsupervised inpainting architecture for high-resolution gaze correction and animation and propose a novel CFM module that can alleviate both the memory and the computational costs in the training and inference stages while achieving high-quality results comparable with training with high-resolution facial images and a GAM module and a Synthesis-As-Training method to generate gaze-correction results with variable angles. We also make publicly available the CelebHQGaze dataset for the research community interested in gaze correction and animation.

3.1.3 Relevant publications

- J. Zhang, J. Chen, H. Tang, E. Sangineto, P. Wu, Y. Yan, N. Sebe, and Wei Wang, Unsupervised High-Resolution Portrait Gaze Correction and Animation, IEEE Transactions on Image Processing, 31(7):5272-5286, July 2022. [27].
  Zenodo record: https://zenodo.org/record/7100432.

3.1.4 Relevant software, datasets and other resources

- The Pytorch of GazeGANV2 implementation and the CelebHQGaze dataset can be found at https://github.com/zhangqianhui/GazeAnimationV2.

3.1.5 Relevance to AI4media use cases and media industry applications

Our gaze correction tool is generic and can be applied to media industry applications that are using face images (e.g., anchor person in news). Concretely, our approach could be useful to (a) UC3 “AI in Vision - High Quality Video Production and Content Automation” and requirement 3A3-11 “Visual indexing and search”, and (b) UC7 “AI for (Re-)organisation and Content Moderation” and requirement 7A3 “(Re)organisation of visual content”, by supporting the efficient training and organization of image and video collections.

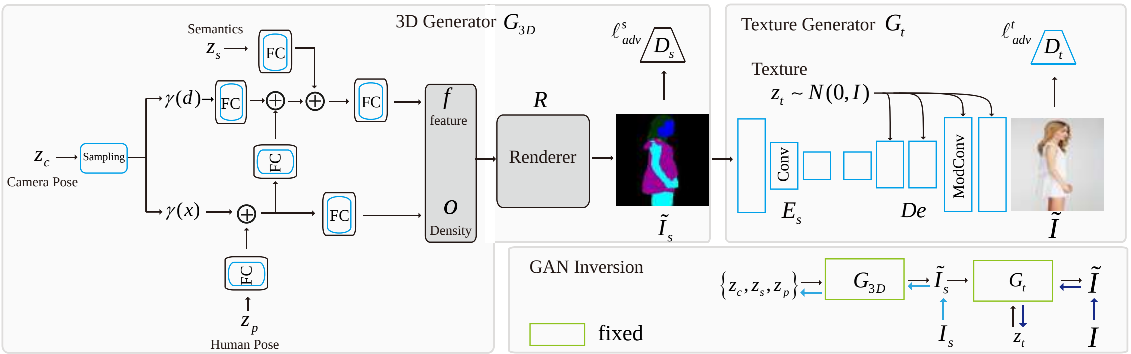


Figure 4. An overview of the proposed 3D-SGAN architecture, composed of two main generators. G_{3D} (on the left) follows a GNeRF structure, with a NeRF kernel used to represent implicit 3D information, latent codes governing different appearance variations and a discriminator (D_s) which is used for adversarial training. The output of G_{3D} is the semantic masks $\tilde{I}_s$ (middle). The second generator (G_t, right) translates the semantic masks into a photo-realistic image $\tilde{I}$. Also G_t is trained adversarially (see top right, the second discriminator D_t). The human generation process can be controlled by interpolating different latent codes: the semantics code $\boldsymbol{z}_s$, the pose code $\boldsymbol{z}_p$, the camera code $\boldsymbol{z}_c$, and the texture code $\boldsymbol{z}_t$. The bottom of the figure shows the GAN inversion scheme.

3.2 3D-Aware Semantic-Guided Generative Model for Human Synthesis

Contributing partner: UNITN

Recent deep generative models can generate and manipulate high-quality images. Specifically, Generative Adversarial Networks (GANs) [28], have been applied to different tasks, such as image-to-image translation [29]–[31], portrait editing [32]–[35], and semantic image synthesis [36], to mention a few. However, most state-of-the-art GAN models [37]–[43] are trained using 2D images, operate in the 2D domain and ignore the 3D nature of the world. Thus, they often struggle to disentangle the underlying 3D factors.

Recently, different 3D-aware generative models [44]–[46] have been proposed to solve this problem. Since most of these methods do not need 3D annotations, they can create 3D content while reducing the hardware costs of common computer graphics alternatives. Differently from generating 3D untextured shapes [46], [47], some of these methods [44], [45], [48]–[50] focus on 3D-aware realistic image generation and controllability. Generally speaking, these models mimic the traditional computer graphics rendering pipeline: they first model the 3D structure, then they use a (differentiable) projection module to project the 3D structure into 2D images. The latter may be a depth map [49], a sketch [48] or a feature map [44] which is finally mapped into the real image by a rendering module. During training, some methods require 3D data [48], [49], and some [44], [45], [50] can learn a 3D representation directly from raw images.

An important class of *implicit* 3D representations is composed of Neural Radiance Fields (NeRFs), which can generate high-quality novel views of complex scenes [51]–[57]. Generative NeRFs (GNeRFs) combine NeRFs with GANs in order to condition the generation process with a latent code governing the object's appearance or shape [57]–[59]. However, these methods [57]–[59] focus on relatively simple and "rigid" objects, such as cars and faces, and they usually struggle to generate non-rigid objects such as the human body. This is likely due to the fact that the human body appearance is highly variable because of both its articulated poses and the variability of the clothes texture, these two factors being entangled with each other. Thus, adversarially learning the data distribution modeling all those factors, is a hard task, especially when the training set is relatively small.

To mitigate this problem, we propose to *split* the human generation process in two separate steps and use intermediate segmentation masks as the bridge of these two stages. Specifically, our 3D-

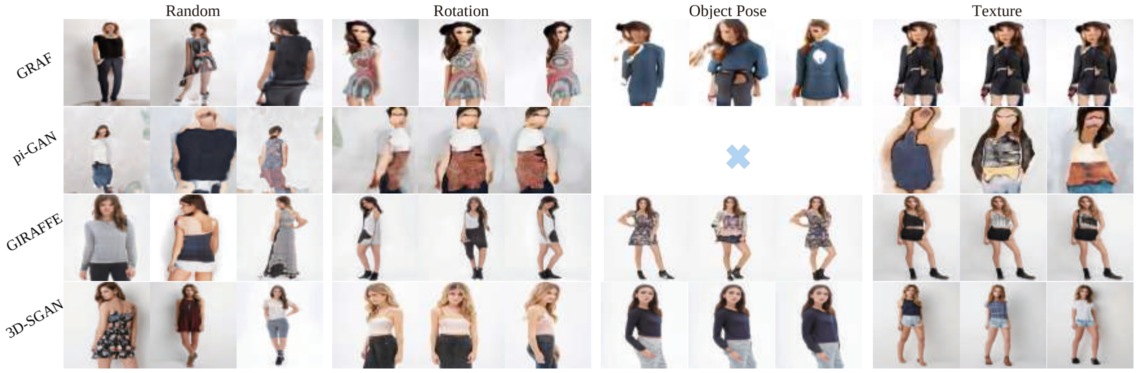


Figure 5. Qualitative comparison. 'Random' means that the results are generated by random sampling the latent codes from the corresponding learned marginal distributions. The other 3 columns show controllable person generation with respect to the rotation, the human pose, and the texture attributes. Note that pi-GAN lacks of the results for 'Object Pose' because it does not include a latent pose code.

Method	Random	Rotation	Object Pose	Texture
GRAF [58]	52.68	176.9	57.76	220.9
pi-GAN [57]	137.6	213.7	-	135.23
GIRAFFE [59]	42.73	123.4	82.61	98.41
3D-SGAN (ours)	**8.240**	**117.3**	**54.00**	**60.63**

Table 3. Quantitative comparison using FID scores (↓).

aware Semantic-Guided Generative model (3D-SGAN) is composed of two generators: a GNeRF model and a texture generator (see Fig. 4). The GNeRF model learns the 3D structure of the human body and produces a semantic segmentation of the main body components, which is largely invariant to the surface texture. The texture generator translates the previous segmentation output into a photo-realistic image. To control the texture style, a Variational AutoEncoder (VAE [60]) is used to modulate the final decoding process. We empirically show that splitting the human generation process into these two stages brings the following three advantages. First, the GNeRF model is able to learn the intrinsic 3D geometry of the human body, even when trained with a small dataset (e.g., DeepFashion [61]). Second, the texture generator can successfully translate semantic information into a textured object. Third, both generators can be controlled by explicitly varying their respective conditioning latent codes. Moreover, we propose two consistency losses to further disentangle the latent codes representing the garment type (which we call "semantic" code) and the human pose. Experiments conducted on the DeepFashion dataset [61] show that 3D-SGAN can generate high-quality person images significantly outperforming state-of-the-art approaches.

3.2.1 Experiments

Datasets. We evaluate 3D-SGAN on the DeepFashion In-shop Clothes Retrieval benchmark [61], which consists of 52,712 high-resolution person images ($1,101\times750$ resolution) with various appearances and poses. This dataset has been widely used in pose transfer tasks. We use the following preprocessing. First, we remove overly cropped images, such as incomplete images of humans. Then, the remaining 42,978 images are resized into a 256×256 resolution, and are divided into 41,001 training and 1,976 test images.

Baselines. We compare 3D-SGAN with three state-of-the-art 3D-aware generative approaches, i.e., GRAF [58], pi-GAN [57] and GIRAFFE [59]. For each baseline, we use the corresponding publicly available code with minor adaptations for the DeepFashion dataset. For a fair comparison, we train all the methods on the DeepFashion dataset.

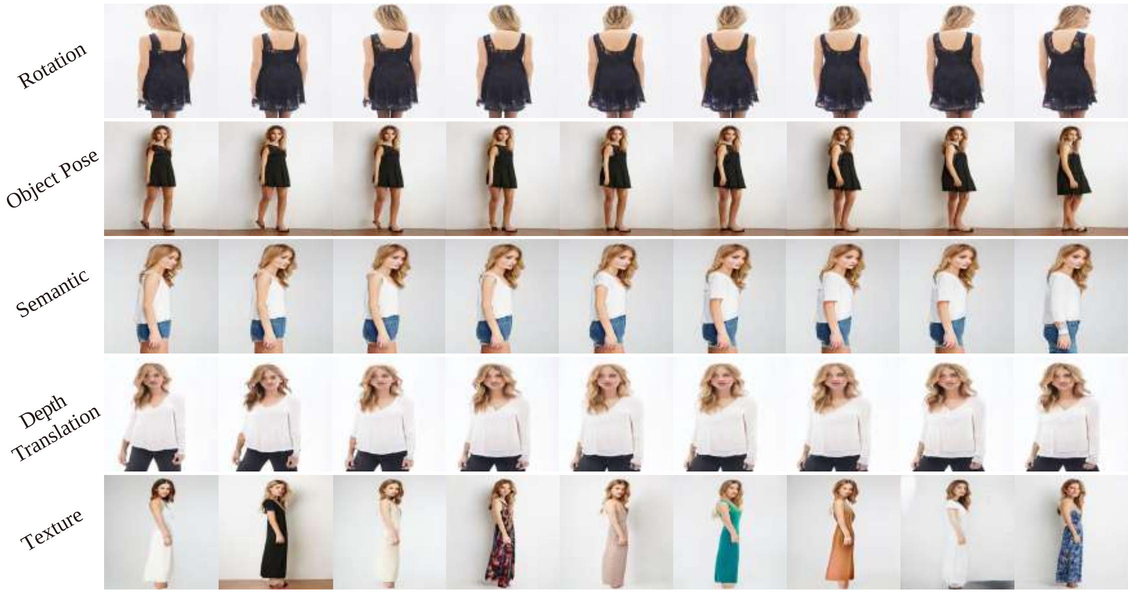


Figure 6. Controllable person generation by interpolating latent codes (Rows 1-4). The 5-th row shows texture generation results obtained randomly sampling $\boldsymbol{z}_t$.

3.2.1.1 Comparisons with state-of-the-art methods

Unconditioned human generation. Fig. 5 ("Random" column) shows a qualitative comparison between image samples generated by all the models. Both GRAF [58] and pi-GAN [57] fail to generate realistic human images. On the other hand, GIRAFFE [59] generates very reasonable human images, but it still suffers from visual artifacts and texture blurs. Compared with these baselines, 3D-SGAN synthesizes much better and more photo-realistic results.

In the first column of Table 3, we provide a quantitative evaluation using FID [62] scores, which are computed using 5,000 randomly sampled images, following standard practice. We observe that 3D-SGAN significantly outperforms all the other baselines, quantitatively confirming the qualitative analysis in Fig. 5.

Controllable human generation. We analyse the representation controllability of all the models. The representation controllability reflects the ability of a model to disentangle different attributes from each other. We achieve this by manipulating a single latent code while fixing the others.

Fig. 5 (columns "Rotation", "Object Pose" and "Texture") shows a qualitative comparison by varying only a single latent code. We observe that all the models can rotate the camera viewpoint. However, GRAF [58] fails to disentangle object pose and texture, showing that for GRAF [58] it is hard to model complex pose variations. Moreover, pi-GAN [57] also suffers from the same problem, since it uses one single latent code to model both texture and pose.

On the other hand, both GIRAFFE [59] and 3D-SGAN can effectively disentangle the different variation factors. However, GIRAFFE [59] suffers from multi-view inconsistencies and mode collapse for texture generation. Compared with the baselines, our model has a better view consistency and a more realistic generation. Table 3 quantitatively confirms of the aforementioned observations.

Fig. 6 shows additional controllable human image generation results obtained with 3D-SGAN. The generated images are realistic and, most of the time, the attributes are effectively disentangled. Specifically, Fig. 6 (1-st row) shows camera rotation results. The images generated by interpolating the camera pose parameter are consistent, and the transition from one image to the next is smooth,

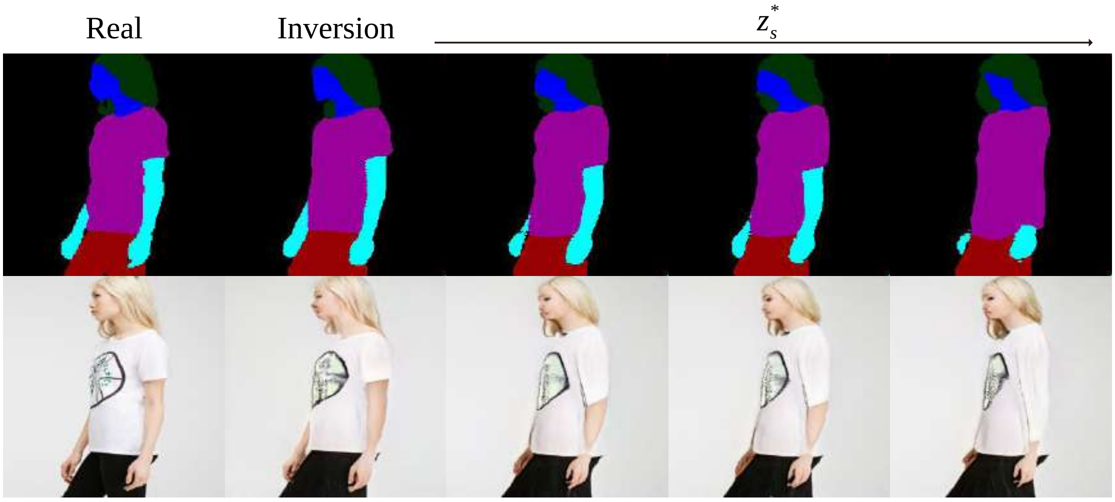


Figure 7. Real data semantic editing results using GAN inversion. The optimal semantic code z_s^*, searching by GAN inversion, can be manipulated to achieve human semantic editing.

while simultaneously preserving the other attributes such as the texture and the pose. On the other hand, Fig. 6 (2-nd row) shows images generated by interpolating the pose code. We again observe that human identity has been well preserved. Additionally, Fig. 6 (3-rd row) shows that the head poses from left to right have slight changes. i.e., face frontalization. It can be explained that the limited training data and the special distribution of fashion images give rise to data bias. More results can be found in [63].

3.2.1.2 Real human image editing

In this section, we use GAN inversion for real data editing tasks. The 2-nd column of Fig. 7 shows that the optimal latent code values $(\boldsymbol{z}_c^*, \boldsymbol{z}_s^*, \boldsymbol{z}_p^*, \boldsymbol{z}_t^*)$, lead to an effective reconstruction of the real input data (1-st column). In other columns, we linearly manipulate the semantic code $\boldsymbol{z}_s^*$, while keeping fixed the other codes. Specifically, the second row of Fig. 7 shows the photo-realistic final images corresponding to the semantic masks in the first row. These results demonstrate the effectiveness of the GAN inversion mechanism and the possibility to apply our model to a wide range of human image editing tasks.

3.2.2 Conclusions

We propose a novel 3D-aware Semantic-Guided Generative model (3D-SGAN) for human synthesis. Specifically, we use a generative NeRF to implicitly represent the 3D human body and render 3D representation into 2D semantic masks. Then, the semantic masks are mapped into the final photo-realistic images using a VAE-conditioned texture generator. Moreover, we propose two consistency losses further to disentangle the geometry pose and the semantics factors. Our experiments show that the proposed approach generates human images which are more realistic and more controllable than state-of-the-art methods.

Limitations. The results from our model are not always perfect, such as low-quality generation and unqualified disentanglement in some cases. As for the reasons, the limited training data and the special distribution of fashion data give rise to data bias and make the model struggle sometimes to disentangle the multiple factors in human person generation. Additionally, the GAN inversion method that we use compromises the reconstruction and the editing ability. A more effective disentanglement and GAN inversion method will be explored in the future.

3.2.3 Relevant publications

- J. Zhang, E. Sangineto, H. Tang, A. Siarohin, Z. Zhong, N. Sebe, and W. Wang, 3D-Aware Semantic-Guided Generative Model for Human Synthesis, European Conference on Computer Vision (ECCV), 2022 [63].
  Zenodo record: https://zenodo.org/record/7525413.
- H. Tang, S. Bai, P. Torr, and N. Sebe, Bipartite Graph Reasoning GANs for Person Pose and Facial Image Synthesis, International Journal of Computer Vision, vol. 131(3): 644-658, March 2023. [64].
  Zenodo record: https://zenodo.org/record/7858398.

3.2.4 Relevant software, datasets and other resources

- The Pytorch implementation for 3D-Aware Semantic-Guided Generative Model for Human Synthesis can be found in
  https://github.com/zhangqianhui/3DSGAN.
- The Pytorch implementation of Bipartite Graph Reasoning GANs for Person Pose and Facial Image Synthesis can be found in
  https://github.com/Ha0Tang/BiGraphGAN.

3.2.5 Relevance to AI4media use cases and media industry applications

Our tools for Human Synthesis can be applied in image editing tasks for the generation of new views of faces and people with controlled features. This could be useful in media industry applications such as movie and advertisement production. As in the other situations where image editing is involved, our approach could be useful to (a) UC3 “AI in Vision - High Quality Video Production and Content Automation” and requirement 3A3-11 “Visual indexing and search”, and (b) UC7 “AI for (Re-)organisation and Content Moderation” and requirement 7A3 “(Re)organisation of visual content”, by supporting editing the content of image and video collections.

3.3 Predict, Prevent, and Evaluate: Disentangled Text-Driven Image Manipulation Empowered by Pre-Trained Vision-Language Model

Contributing partner: UNITN

Disentangled image manipulation [33], [65]–[73] aiming at changing the desired attributes of the image while keeping the other attributes unchanged, has long been studied for its research significance and application value. Reaching this target is not easy, especially when attributes naturally entangle in the real world. Therefore, concrete attribute annotations are of vital importance, making disentangled image manipulation a labor-consuming task.

Several works [65], [68]–[70] use an encoder-decoder architecture and need manual annotations on multiple attributes of images. The models encode the original image and the manipulating attribute, then decode the manipulated image. Specifically, they use an attribute-specific loss to encourage the manipulation of a specific attribute while discouraging the others. The loss comes from pre-trained classifiers for all annotated attributes. Many recent works focus on latent space image manipulation since large-scale pre-trained GANs, e.g., StyleGANs [74], [75], can generate high-quality images from well-disentangled latent spaces. Despite the convenience of directly using the pre-trained GANs to generate images, all these methods need human annotations [33], [66], [67], [71]–[73]. Moreover, the available manipulating attributes are limited to the annotated set.

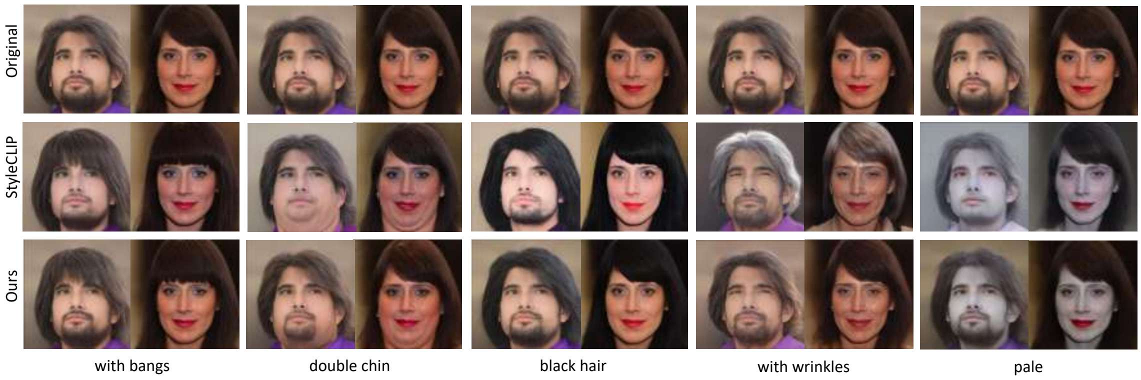


Figure 8. Comparisons on disentangled image manipulation between the StyleCLIP [76] baseline and our Predict, Prevent, and Evaluate (PPE). Ours manages to manipulate only the command-attribute (as indicated under each column) while remaining unchanged to the others.

Recently, the rise of the large-scale pre-trained vision-language model CLIP [77] has brought a new insight. Since CLIP provides effective signals about the semantic similarity of image and text, various manipulations [76], [78], [79] can be performed with a text command and a CLIP-based loss, instead of exhaustive human annotations. Nevertheless, achieving disentangled image manipulation is still tricky. For instance, StyleCLIP [76] introduces three methods: latent optimization and latent mapper take no consideration of achieving disentangled results; global direction, which is based on the more disentangled $\mathcal{S}$ latent space [80], needs human trials-and-errors to find appropriate parameters in each case to reach the expected effects. To only manipulate a desired attribute, TediGAN [81], [82] merely revise the latent vectors of layers corresponding to that attribute. Yet, they have to figure out in advance the relations between attributes and layers in StyleGAN.

In this work, we explore achieving disentangled image manipulation with as less human labor as possible. We propose a novel framework, i.e., Predict, Prevent, and Evaluate (PPE), to approach the target by leveraging the power of CLIP in depth. Firstly, we propose to **Predict** the possibly entangled attributes for given text commands. We assume that the entanglements result from the distributions of attributes in the real world. Therefore, we draw support from CLIP to find the attributes that appear most frequently in the command-related images, then regard the attributes of high co-occurrence frequency as the possibly entangled attributes. Secondly, we introduce a novel entanglement loss to **Prevent** entanglements during training. The loss punishes the changes of the possibly entangled attributes before and after the manipulation, so as to enforce the model to find a less disentangled manipulating direction. Lastly, based on the predicted entangled attributes, we introduce a new evaluation metric to simultaneously **Evaluate** the manipulation effect and the entanglement condition. The manipulation effect is measured based on the change of command-attribute while the entanglement condition is based on the change of the entangled attributes, before and after manipulation. All the changes are estimated according to the CLIP distance between the texts of attributes and the images.

To evaluate, we implement our method based on the simple and versatile latent mapper from StyleCLIP and conduct experiments on the challenging face editing task, using the large-scale human face dataset CelebA-HQ [38], [83]. Qualitative and quantitative results indicate that we achieve superior disentangled performance compared to the StyleCLIP baselines (see Fig. 8). Meanwhile, we show that our results present a better linear consistency.

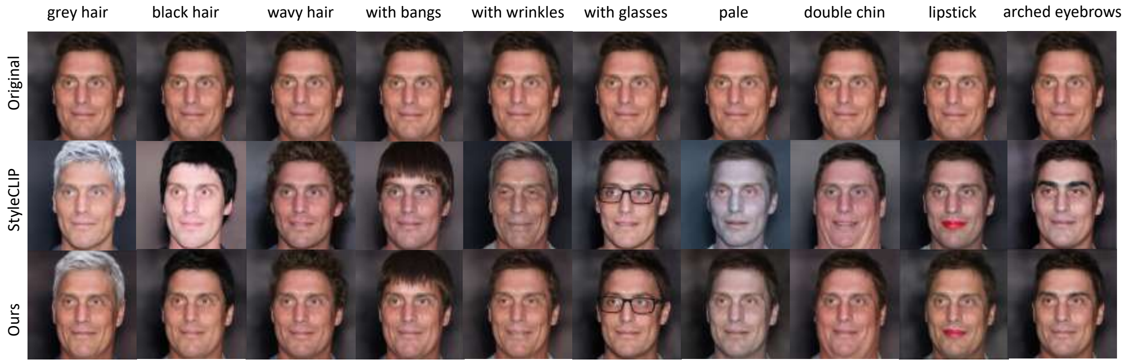


Figure 9. Qualitative comparison with StyleCLIP [76] using different text commands (indicated on the top). Ours achieves more disentangled manipulation results as only the desired attribute is manipulated while others are maintained well.

3.3.1 Experiments

3.3.1.1 Qualitative Results

Direct Manipulation Outputs. We firstly compare the directly outputted manipulation results from the trained models, without changing the manipulation strengths. In Fig. 9, we illustrate the comparing qualitative results on multiple text commands. As can be seen in the manipulation results of "StyleCLIP", it not only manipulates the required attributes, but also manipulates other attributes. Take text command *"grey hair"* as an example; the manipulated face gets grey hair, while it gets whiter skin and grey eyes simultaneously. Similarly, for the text command *"with wrinkles"*, the manipulated face gets wrinkles, grey hair, and more closed eyes. Other manipulation results are obtained in similar conditions.

By contrast, "Ours" achieves more ideal manipulation results, where almost only the desired attribute is manipulated while other attributes of are well preserved. For example, for *"wavy hair"*, "Ours" hair becomes wavy while the hair length is close to the original one and the skin color does not become whiter; for *"double chin"*, "Ours" gets double chin while the eye color remains light brown, skin color is kept well, and mouth does not open much. In addition, it is worth mentioning that the qualitative results are quite consistent with the quantitative results, indicating that the proposed evaluation metrics are effective for the disentangled image manipulation task.

Strength-Adjusted Manipulation Outputs. We further compare the manipulation results with gradually increasing manipulation strength. To illustrate, we show two groups of comparing results in Fig. 10 (more results in the published article). In each group, we present the manipulation results for male and female, respectively. We observe that our method learns more disentangled manipulation directions compared to StyleCLIP. For StyleCLIP, when the manipulation strength increases, the desired attribute becomes more and more obvious, as well as the entangled attributes. As the male-case in Fig. 10 (top), from left to right, the eyes become increasingly blue while they also become wider, the face becomes whiter, and the hair color becomes lighter. Contrarily, our method presents better manipulation consistency. When the manipulation strength increases, the target attribute gradually turns more prominent while others remain almost unchanged.

3.3.2 Conclusions

We propose Predict, Prevent, and Evaluate (PPE) to achieve disentangled image manipulation with little manual effort by deeply exploiting the powerful large-scale pre-trained vision-language model CLIP. CLIP is leveraged to 1) **Predict** the entangled attributes given textual manipulation

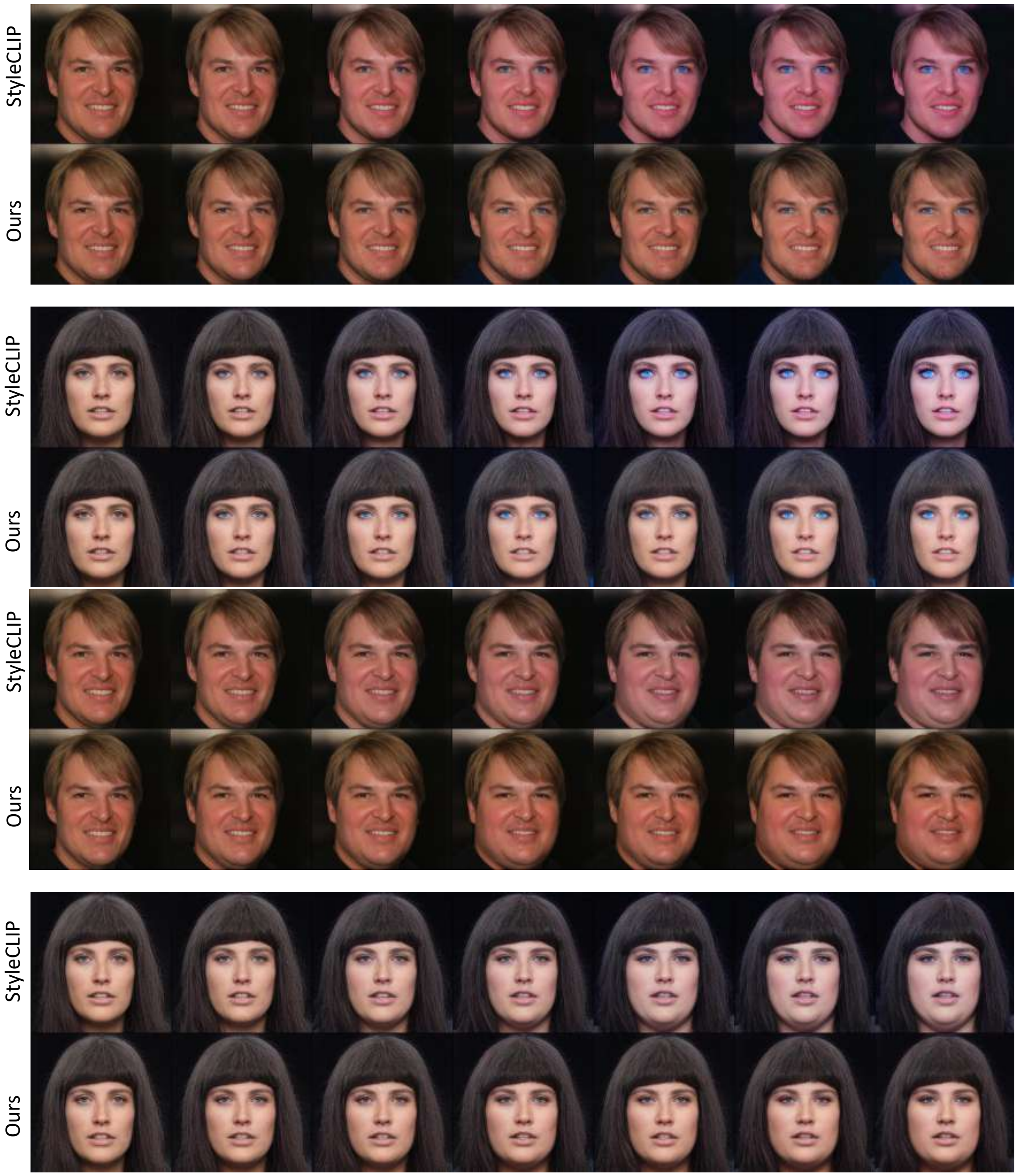


Figure 10. Image manipulation results from StyleCLIP [76] and ours, using gradually increasing manipulation strengths (blue eyes (top), chubby (bottom). Ours present better fore-and-aft consistency along the change of manipulation strength.

command, 2) **Prevent** the model from finding entangled manipulating latent directions through a novel entanglement loss, which punishes the directions that lead to the change of predicted entangled attributes, and 3) establish a new evaluation metric that can simultaneously **Evaluate** the change of the command-attribute and entangled attributes in image manipulation. PPE is tested

on the challenging face editing task and is proven better than the StyleCLIP baseline according to quantitative and qualitative results.
Limitations. The limitations of the proposed PPE method are as follows: 1) Similar to StyleCLIP, the command out of the domain of CLIP and StyleGAN may not obtain ideal manipulation results. Nevertheless, the problem can be partially solved by recent works that study domain adaptation of image generators, like StyleGAN-NADA [78]. 2) The extent of disentanglement in the manipulation results depends on the extent of disentanglement in the latent space of StyleGAN. Since we study latent space image manipulation, the best our method can do is to find the most disentangled latent path in the latent space of pre-trained generator. If the attributes are originally entangled for the generator, PPE is unable to achieve completely disentangled manipulation results.

3.3.3 Relevant publications

- Z. Xu, T. Lin, H. Tang, F. Li, D. He, N. Sebe, R. Timofte, L. Van Gool, and E. Ding, "Predict, Prevent, and Evaluate: Disentangled Text-Driven Image Manipulation Empowered by Pre-Trained Vision-Language Model," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022 [84].
  Zenodo record: `https://zenodo.org/record/7100266`.

3.3.4 Relevant software, datasets and other resources

- The Pytorch implementation can be found in `https://github.com/zipengxuc/PPE`.

3.3.5 Relevance to AI4media use cases and media industry applications

Our PPE approach can be applied in image editing tasks for the manipulation of facial attributes. This could be useful in media industry applications such as movie and advertisement production. As in the other situations where image editing is involved, our approach could be useful (a) UC3 "AI in Vision - High Quality Video Production and Content Automation" and requirement 3A3-11 "Visual indexing and search", and (b) UC7 "AI for (Re-)organisation and Content Moderation" and requirement 7A3 "(Re)organisation of visual content", by supporting editing the content of image and video collections.

3.4 Playable Environments: Video Manipulation in Space and Time

Contributing partner: UNITN

What would you change in the last tennis match you saw? The actions of the player? The style of the field, or, perhaps, the camera trajectory to observe a highlight more dramatically? To do so interactively, the geometry and the style of the field and the players need to be reconstructed. Players' actions need to be understood and the outcomes of future actions anticipated. To enable these features one needs to reconstruct the observed *environment* in 3D and provide simple and intuitive interaction, offering an experience similar to *playing* a video game. We call these representations Playable Environments (PE).

Such a representation enables multiple creative applications, such as 3D- and action-aware video editing, camera trajectory manipulation, changing the action sequence, the agents and their styles, or continuing the video in time, beyond the observed footage. Fig. 11 shows a playable environment for tennis matches. In it, the user specifies actions to move the players, controls the viewpoint and changes the style of the players and the field. The environment can be played, akin to a video game, but with real objects.

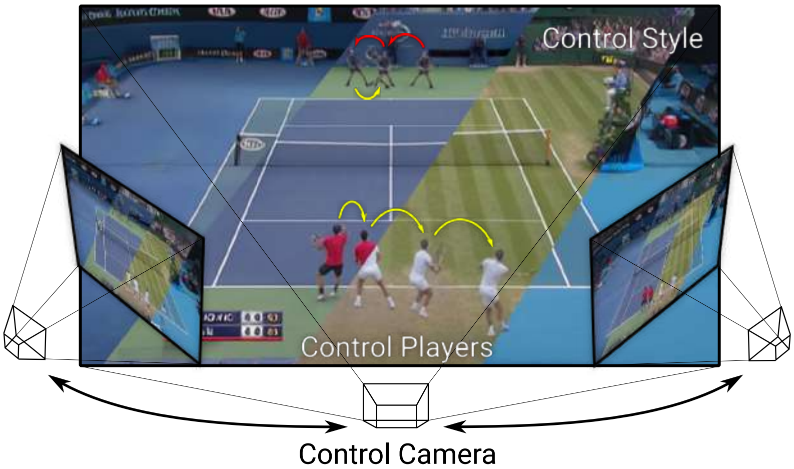


Figure 11. Given a single initial frame, our method creates playable environments *that allow the user to interactively generate different videos by specifying discrete actions to control players, manipulating camera trajectory and indicating the style for each object in the scene.*

	Name	Description
⟨**1**⟩	*Playability*	The user can control generation with discrete actions.
⟨**2**⟩	*Camera control*	The camera pose is explicitly controlled at test time.
⟨**3**⟩	*Multi-object*	Each object is explicitly modeled.
⟨**4**⟩	*Deformable objects*	The model handles deformable object such as human bodies
⟨**5**⟩	*Appearance changes*	The model handles objects whose appearance is not constant is the training set
⟨**6**⟩	*Robustness*	The model is robust to calibration and localization errors.

Table 4. Characteristics of our method for Playable Environments. Each row is referred in the text with ⟨·⟩ *symbols.*

In this work, we propose a method to construct PEs of complex scenes that supports a large set of interactive manipulations. Trained on a dataset of monocular videos, our method presents six core characteristics listed in Table 4 that enable the creation of such PEs. Our framework allows the user to interactively generate videos by providing discrete actions ⟨**1**⟩ and controlling the camera pose ⟨**2**⟩. Furthermore, it can represent environments with multiple objects ⟨**3**⟩ with varying poses ⟨**4**⟩ and appearances ⟨**5**⟩ and is robust to imprecise inputs ⟨**6**⟩. In particular, we do not require ground-truth camera intrinsics and extrinsincs, but assume they can be estimated for each frame. Neither do we assume ground-truth object locations, but rely on an off-the-shelf object detector [85] to locate the agents in 2D, such as both tennis players. No other supervision is required.

Playable Environments encapsulate and extend representations built by several prior image or video manipulation methods. Novel view synthesis and volumetric rendering methods support re-rendering of static scenes. However, while some methods support moving or articulated objects [86]–[89], it is challenging for them to handle dynamic environments and they do not allow user interaction, making them undesirable for modeling compelling environments. Video synthesis methods manipulate videos by predicting future frames [90]–[93], animating [94]–[96] or playing videos [97], but environments modeled with such methods typically lack camera control and multi-object support. Consequently, these methods limit interactivity as they do not take into account the 3D nature of the environment.

	LPIPS↓	FID↓	FVD↓	Δ-*MSE*↓	Δ-*Acc*↑	ADD↓	MDR↓
MoCoGAN [92]	0.266	132	3400	101	26.4	28.5	20.2
MoCoGAN+	0.166	56.8	1410	103	28.3	48.2	27.0
SAVP [90]	0.245	156	3270	112	19.6	10.7	19.7
SAVP+	0.104	25.2	**223**	116	33.1	13.4	19.2
CADDY [97]	0.102	**13.7**	239	72.2	45.5	**8.85**	1.01
(Ours)	**0.089**	15.3	237	**32.8**	**68.1**	9.47	**0.15**

Table 5. Comparison with PVG state of the art on the Static Tennis dataset of [97]. Δ-MSE, Δ-Acc *and MDR in %, ADD in pixels.*

Our method consists of two components. The first one is the synthesis module. It extracts the state of the environment—location, style and non-rigid pose of each object—and renders the state back to the image space. Recently introduced Neural Radiance Fields (NeRFs) [51] represent an attractive tool for their ability to render novel views. In this work, we introduce a style-based modification of NeRF to support objects of different appearances. Furthermore, we propose a compositional non-rigid volumetric rendering approach handling the rigid parts of the scene and non-rigid objects. The second component—the action module—enables playability. It takes two consecutive states of the environment and predicts an action with respect to the camera orientation. We train our framework using reconstruction losses in the image space and the state space, and a novel loss for action consistency. Finally, to improve temporal dynamics, we introduce a temporal discriminator that operates on sequences of environment states.

To thoroughly evaluate ⟨**1**−**6**⟩, we introduce two complementary large-scale datasets for the training of playable environments, a synthetic and a real one. The first is intended to evaluate ⟨**1**−**5**⟩, with a particular focus on camera control thanks to the synthetic ground truth, the second to evaluate ⟨**1**−**6**⟩, with a particular focus on ⟨**4**−**6**⟩ given the high diversity present in this dataset. We propose an extensive evaluation of our method with several baselines derived from existing NeRF and video generation methods. These experiments show that our method is able to generate high-quality videos and outperforms all baselines in terms of playability, camera control and video quality.

3.4.1 Experiments

Datasets. Evaluating ⟨**1-6**⟩ is challenging and requires video datasets featuring camera motion ⟨**2**⟩, multiple playable objects ⟨**1**,**3**⟩, deforming objects ⟨**4**⟩ and varied appearance ⟨**5**⟩. For this reason, we collect three datasets:

- *Minecraft* dataset. We collect a synthetic video dataset with duration of 1h with two sparring *Minecraft* [98] players. Wide camera movement and diverse, deforming players allow the evaluation of ⟨**1**−**5**⟩.
- *Minecraft Camera* dataset. We collect *Minecraft* [98] sequences where the camera is moved in the neighborhood of a starting position. We use these frames as a synthetic ground truth for the evaluation of camera control ⟨**2**⟩.
- *Tennis* dataset. We collect a large-scale dataset of 43 broadcast tennis matches totalling 12h of videos for the evaluation of ⟨**1-6**⟩. The dataset features challenging player poses ⟨**5**⟩, high variability in tennis fields and players ⟨**4**⟩ and noise in camera estimation and player localizaton ⟨**6**⟩.

To allow comparison with playable video generation methods under their simplifying assumptions, we adopt the *Tennis* dataset of [97], referred to as *Static Tennis.* The dataset features limited camera movement, each video is cropped to depict only a single player, only one field is present and players have uniform appearance, thus only ⟨**1**,**4**⟩ are evaluated.

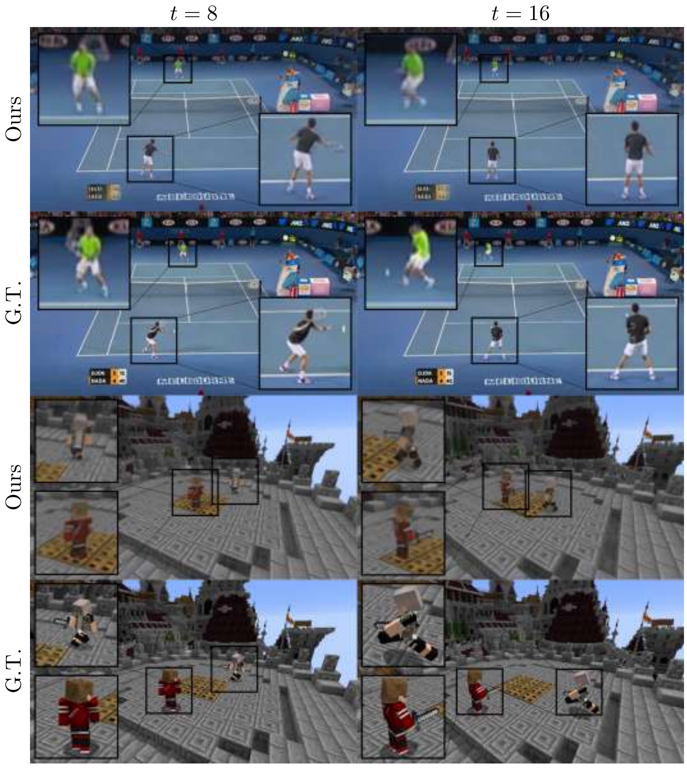


Figure 12. Qualitative reconstruction results produced by our method on the Tennis *and* Minecraft *datasets. In the reconstructed sequence, playable objects move according to the ground truth sequence and are rendered in realistic poses.*

Evaluation Protocol. We perform a separate evaluation of the synthesis ⟨**2-6**⟩ and the action modules ⟨**1**⟩ using similar evaluation protocols. For the former, we reconstruct each test sequence by extracting the environment state of each frame and rendering the original frame back. For the action module, we follow the evaluation protocol of [97]. In particular, we consider a test sequence and extract the environment state of the first frame, then we use the action network to extract the sequence of discrete actions present in the sequence and reconstruct each frame starting from the first environment state.

As video quality metrics ⟨**2**,**4-6**⟩ we adopt *LPIPS* [99], *FID* [100] and *FVD* [101] computed between the test sequences and the reconstructed sequences. For evaluation of the action space ⟨**1**,**3**⟩, following [97], we define Δ as the difference in position of an object between two given frames and use the following metrics:

• Δ *Mean Squared Error (Δ-MSE)*: The expected error in terms of MSE in the regression of Δ from a discrete action. For each action, the average Δ is used as the optimal estimator. The metric is normalized by the variance of Δ.

• Δ*-based Action Accuracy (Δ-Acc)*: The accuracy with which a discrete action can be regressed from Δ.

• *Average Detection Distance (ADD)*: The average Euclidean distance between the bounding box centers of corresponding objects in the test and reconstructed frames.

• *Missing Detection Rate (MDR)*: The portion of detections that are present in the test sequences but that are not matched by any detection in the reconstructed sequences.

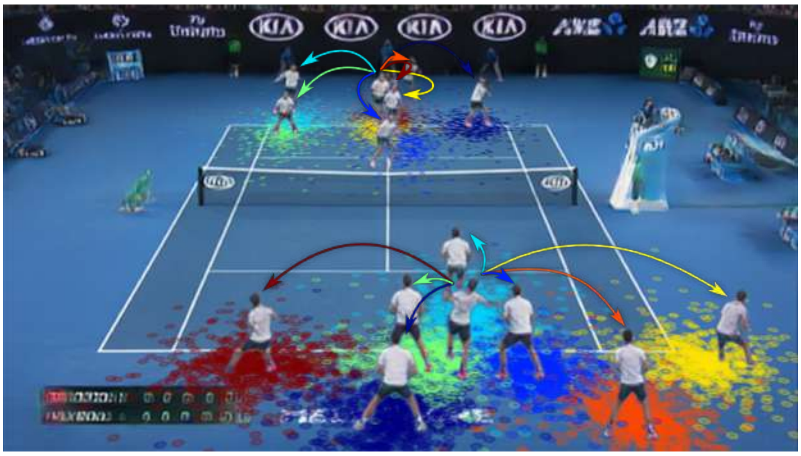


Figure 13. Action space learned by our method on the Tennis *dataset. Each color represents a learned action and each arrow shows the effects of applying the respective action six times to the initial player. The overlay on the floor shows the distribution of possible ending positions after the application of each action.*

3.4.1.1 Comparison on Playable Video Generation In this section, we evaluate the action-modeling capabilities of our method by comparing against the state of the art in the related problem of Playable Video Generation (PVG) [97] where the objective is to learn a set of discrete action labels in an unsupervised fashion to condition video generation. Differently from our setting, in PVG no explicit camera control is required. Moreover, existing PVG methods assume a single user-controllable object and that camera motion is limited. To satisfy these simplifying assumptions, we adopt the *Static Tennis* dataset of [97]. Tab. 5 shows the results. Our method substantially improves the Δ-MSE and Δ-Acc action quality metrics suggesting that the learned actions are better correlated with player movement. In addition, the reduced LPIPS and MDR indicate an improvement in the quality of the generated reconstruction.

3.4.1.2 Comparison with Previous Methods We propose to build baselines for the creation of PEs from state of the art methods in the related problem of PVG. We make use of the following set of versions of CADDY [97] which are modified to account for multiple playable objects and for camera motion: (i) the action network produces a distinct output for each dynamic object in the environment; (ii) (i) + the action and dynamics networks are conditioned on bounding box and camera information; (iii) (ii) + output resolution is increased to match our method; (iv) (ii) + $\mathcal{L}_\Delta$; (v) (iii) + $\mathcal{L}_\Delta$.

Playability and camera control evaluation

Fig. 12 shows qualitative reconstruction results for our method. As suggested by the MDR and ADD scores, the model correctly synthesizes both players and is able to reconstruct the player movements of the ground truth sequence using only a sequence of discrete actions. In addition, a visualization of the learned action space (see Fig. 13) shows that the model learns a set of diverse discrete actions that correspond to the main movement directions.

In Fig. 14 we show qualitative camera and style manipulation results for our method on the *Tennis* dataset. Our model can synthesize the scene under novel views and correctly alter the style of the field and players to the one of a target image.

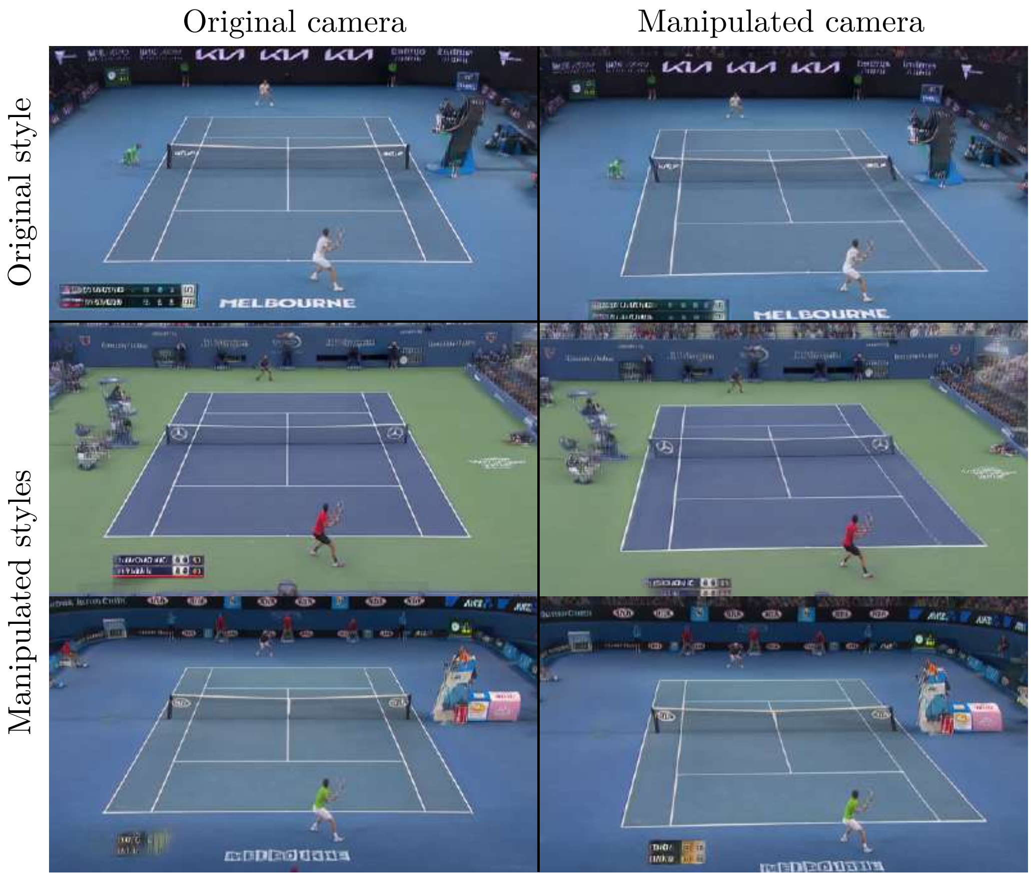


Figure 14. Camera and style manipulation results on the Tennis *dataset. The original image is rendered under a novel camera perspective using varying styles for the field and players.*

3.4.2 Conclusions

In summary, the primary contributions of this work are as follows: **a new framework** for the creation of compelling Playable Environments with the characteristics in Tab. 4, featuring **a new compositional NeRF** that handles deformable objects with different visual styles and an **action module** that operates in the latent space of our NeRF model; **two challenging large-scale datasets** for training and evaluating PEs to stimulate future research in this area.

3.4.3 Relevant publications

- W. Menapace, A. Siarohin, C. Theobalt, V. Golyanik, S. Tulyakov, S. Lathuilière, E. Ricci, “Playable Environments: Video Manipulation in Space and Time”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’22), June 2022. [102].
  Zenodo record: https://zenodo.org/record/7075454.

3.4.4 Relevant software, datasets and other resources

- The Pytorch implementation can be found in
  https://github.com/willi-menapace/PlayableEnvironments.

3.4.5 Relevance to AI4media use cases and media industry applications

Our video generation approach can be applied to creative media industries where content manipulation is important. This can be extended to the gaming industry by providing realistic games based on real world footage, thus making our approach relevant to UC5 “AI for games”. As in the other situations where content editing is involved, our approach could be also useful to (a) UC3 “AI in Vision - High Quality Video Production and Content Automation” and requirement 3A3-11 “Visual indexing and search”, and (b) UC7 “AI for (Re-)organisation and Content Moderation” and requirement 7A3 “(Re)organisation of visual content”, by supporting editing the video content.

4 Manipulation and synthetic content detection in multimedia (T6.2) - Manipulated content detection

Contributing partners: CERTH, CEA, FhG-IDMT, CNR, QMUL, UPB, UNIFI, UNITN

Unlike Section 3, which presents methods for content creation, this Section shows the research carried out in AI4Media for the detection of such synthetically generated or manipulated content, by developing novel detection techniques for different modalities. Given some of the concerns listed in the previous Section, related to the creation and spreading of fake news, deepfakes, and misinformation, AI methods that target the detection of this type of content are important in creating trustworthy and fair media spaces. Furthermore, this Section tackles important issues regarding the performance of manipulated content detection systems, looking into their generalization capabilities, data augmentation techniques, noisy data, and AI model robustness and reliability.

More specifically, Section 4.1 focuses on techniques for video- and image-based detection, Section 4.2 on audio-based detection, and finally, Section 4.3 on text-based detection.

4.1 Video- and image-based synthetic content detection

This Section focuses on the detection of video and image synthetic content, created via popular deepfake creation methods. It addresses some of the most important topics and limitations of current technologies, including the analysis of spatial and temporal limitations of detection technologies, important metrics of performance related to reliability and generalization capabilities of AI detector models, as well as data augmentation for deepfake detection. Thus, the research presented here shows an interest to go beyond just creating synthetic content detectors and look into topics related to the robustness of these systems.

4.1.1 MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection

Contributing partners: CNR, UNIFI, CERTH

In recent years, there has been a notable increase in the proliferation of deepfake images and videos. These manipulations are becoming increasingly credible and lifelike, thanks to the continuous advancements in generative models like Generative Adversarial Networks (GANs). As a result, verifying the authenticity and truthfulness of such media has become a challenging task. In response to this uncontrollable evolution, various efforts have been made to combat deepfakes by developing numerous deepfake detection systems based on different approaches.

To overcome deepfake detection challenges, the method denominated Multi-Identity size-iNvariant TIMEsformer (MINTIME) for video deepfake detection has been developed. The method's pipeline is sketched in Figure 15.

According to Figure 15, the preprocessing pipeline (left) starts with the detection of faces in the video, follows with the clustering of identities and then creates the input sequence. The sequence of faces is converted into features by the convolutional backbone (right), which once converted into tokens and concatenated to the embeddings, pass into the TimeSformer, and finally into the MLP Head for the final classification.

In order to be able to capture both the spatial and temporal aspects, while also handling multiple identities, the architecture is based on a modified TimeSformer so that spatiotemporal attention is calculated separately for each identity and fed into a common CLS. This novel attention mechanism has been called Identity-aware Spatio-Temporal Attention. To ensure consistency between tokens, positional embedding was also evolved by introducing Temporal-Coherent Posi-

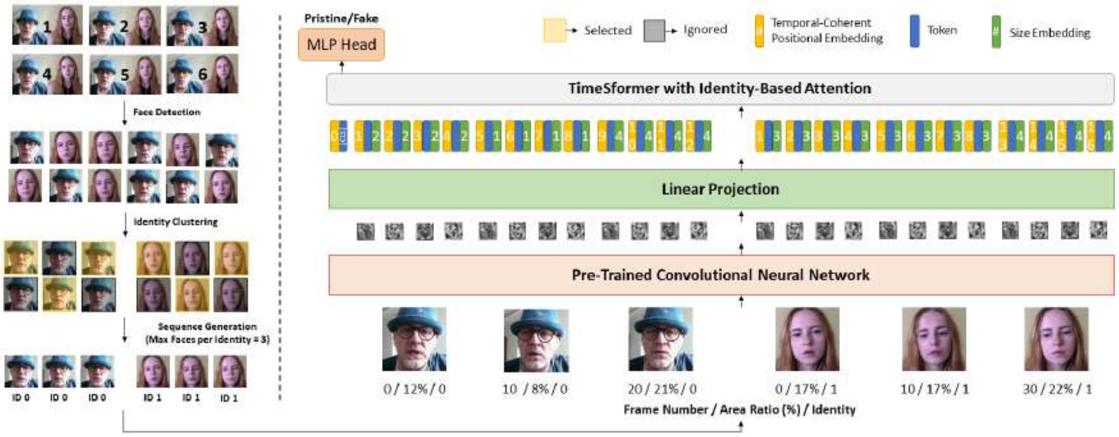

Pristine/Fake
MLP Head
Selected
Ignored
Temporal-Coherent Positional Embedding
Token
Size Embedding
TimeSformer with Identity-Based Attention
Linear Projection
Pre-Trained Convolutional Neural Network
0 / 12% / 0
10 / 8% / 0
20 / 21% / 0
0 / 17% / 1
10 / 17% / 1
30 / 22% / 1
Frame Number / Area Ratio (%) / Identity
Face Detection
Identity Clustering
Sequence Generation (Max Faces per Identity = 3)
ID 0
ID 0
ID 0
ID 1
ID 1
ID 1

Figure 15. MINTIME pipeline

tional Embedding (TCPE) capable of numbering face tokens according to both their position in space and time.

This work has been conducted in collaboration between CNR, CERTH and UNIFI during the AI4Media Junior Fellow Exchange program.

4.1.1.1 Experiments According to Table 6, MINTIME-XC (based on XceptionNet) outperforms the state-of-the-art on the ForgeryNet [103] validation set in terms of AUC (Area Under Curve) and is quite similar to what was obtained with a SlowFast in terms of accuracy, which is, however, limited to considering a single identity in the classification phase.

Model	#IDs	Acc	AUC
SlowFast R-50† [104]	1	82.59	90.86
SlowFast R-50* [104]	1	**88.78**	93.88
X3D-M* [105]	1	87.93	93.75
MINTIME-EF	1	81.92	90.13
MINTIME-EF	2	82.28	90.45
MINTIME-EF	3	82.05	90.28
MINTIME-XC	1	85.96	93.20
MINTIME-XC	2	87.64	**94.25**
MINTIME-XC	3	86.98	94.10

Table 6. Video-Level Evaluation on ForgeryNet Validation Set. The identities column is the number of considered identities for the inference. † *Indicates that the model has been trained in our setup.* * *Indicates that the result is taken from [103].*

All MINTIME models also prove to be particularly robust at analyzing videos considering a variable number of identities without a significant loss of overall accuracy. By testing the trained models considering only videos containing more than one person in the scene, MINTIME-XC significantly outperforms the trained state-of-the-art models by correctly classifying most of the videos considered as shown in Table 7. Interestingly, SlowFast R-50 trained considering only one identity per video performs poorly on multi-identity videos as it is heavily influenced by the choice of this single identity to be analyzed. This is certainly one of the most interesting results obtained as

Model	#IDs	Acc	AUC
SlowFast R-50 [104]	1	72.63	80.92
MINTIME-EF	2	81.21	89.56
MINTIME-XC	2	**86.68**	**94.12**

Table 7. Evaluation on multi-identity only videos of ForgeryNet Validation Set. The models are all trained in our setup.

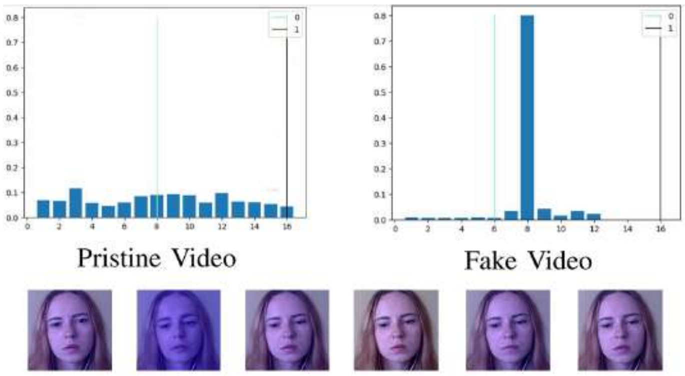


Figure 16. Attention values on the sixteen input faces for a pristine video (left) and a fake video (right). The vertical coloured lines separate the two identities. The faces on the bottom are the ones extracted from the fake video, coloured based on the intensity of attention on each of them.

it concretely highlights how necessary it is to create models capable of handling multiple identities in the same video and not simply rely on selection criteria for these.

Qualitative Evaluation

All our models have been trained to perform binary classification of the entire video but, in the case of deepfake videos, a hypothetical final user might be interested not only in knowing whether the video has been manipulated but also at what instant and if there is more than one tampered person. These are typical requirements when we want to expose these systems to end users (e.g. journalists) [106].

We can retrieve this information by analyzing the attention values obtained on the various faces which compose the input sequence. Indeed, it has been empirically shown that when the video is pristine, there are no relevant alarms of detection, as shown in Figure 16 (left), while in the presence of a deepfake video, the model pays more attention on the faces containing the trace, as shown in Figure 16 (right). By analyzing the attention values, it is easy to trace which identity has been manipulated and at what instant(s) the trace is present.

Examples of outcomes from the model are shown in Figure 17 and it can be seen that in all cases the proposed model is able to identify the fake identity, if any, even in crowd situations. Interestingly is the case (Figure 17 bottom-left) in which a cartoon face picture is detected, the model still manages to realize that the manipulated face is that of the man.

4.1.1.2 Conclusions The main novelties presented by the MINTIME approach are the following:

- Ability to identify inconsistencies in video in both space and time through the combination of a Spatio-Temporal Transformer and a Convolutional Neural Network (CNN).
- Ability to handle multiple people in the same video through an Identity-aware attention

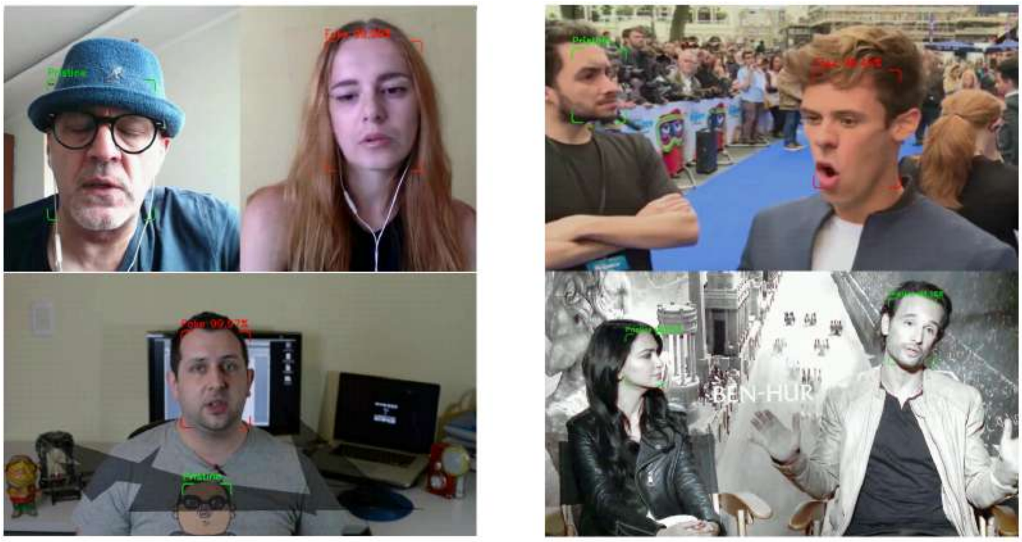


Figure 17. Shots of outcomes obtained with MINTIME in different multi-identity contexts.

mechanism, capable of keeping track of the identity to which each face detected in the video refers, combined with a positional embedding technique, namely Temporal Coherent Positional Embedding, which can maintain both spatial and temporal coherence.

- Ability to handle variations in the face-frame area ratio through the introduction of size embeddings that keep track of the ratio between the detected face area and the entire frame at each instant of time.
- Being unaffected by aggregation strategies thanks to an internal aggregation obtained by analyzing the entire video in a single sequence, letting the network infer the video-level prediction by handling appropriately multi-identity videos in a single forward pass. This way, the model directly returns a single prediction for the entire video without requiring additional post-processing.

It is important to emphasise that it is generally necessary for the future to also create datasets that are suitable for handling the problem of real-world situations, as deepfake detectors trained and validated on datasets that are too far removed from reality and too standardised are likely to be useless when used in the real world.

4.1.1.3 Relevant publications

- Coccomini, D.A., Zilos, G.K., Amato, G., Caldelli, R., Falchi, F., Papadopoulos, S., and Gennaro, C. (2022). MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection. ArXiv, abs/2211.10996.

4.1.1.4 Relevant software, datasets and other resources

- The Pytorch implementation can be found in `https://github.com/davide-coccomini/MINTIME-Multi-Identity-size-iNvariant-TIMEsformer-for-Video-Deepfake-Detection`.

4.1.1.5 Relevance to AI4media use cases and media industry applications The MINTIME solution contributes to UC1 (AI for Social Media and Against Disinformation), and specifically to feature 1A (Detection/Verification of Synthetic Media). Such systems could be adopted by journalists who use AI-based tools to verify the authenticity of multimedia content in a real-world scenario and by social media for content moderation. In fact, in these cases, it is important to catch deepfakes even in borderline and very specific situations that may not be so common in datasets but are the normality in real-world scenarios, such as multi-identity videos or rapid face movements and size variations.

4.1.2 On the Generalization of Deep Learning Models in Video Deepfake Detection

Contributing partners: CNR, UNIFI

The present work is the result of a joint and fruitful collaboration between the partners CNR and UNIFI. Deep Learning has had a profound impact on society, driving remarkable progress in various fields. However, its application can also yield negative consequences, exemplified by the emergence of deepfakes. Deepfakes refer to manipulated images or videos that portray subjects in ways they never actually acted or spoke.

To address this issue, researchers have devised deepfake detection techniques, primarily based on deep learning. These methods aim to identify traces left behind during the manipulation process, The goal is to create powerful deepfake detectors capable of generalizing and identifying deepfakes, regardless of the manipulation technique employed, even if it's novel and not present in the training data. During training, a diverse set of data is required to expose the models to numerous forms of deepfakes, encouraging them to abstract and generalize effectively.

In this study, we compared different deep learning architectures to assess their ability to generalize against deepfake videos generated with multiple methods. Three types of network architectures were compared: a convolutional network, such as "EfficientNet V2," a standard "Vision Transformer," and a "Swin Transformer," inspired by the hierarchical approach of convolutional neural networks. The experiments revealed that the Vision Transformer outperformed other models in terms of generalization ability when evaluated in a cross-forgery context. However, the Swin Transformer showed better performance in cross-dataset experiments. Vision Transformer seems to be able to generalize better only when a large and diverse training dataset is available. In contrast, the EfficientNet-V2 and Swin Transformer performed satisfactorily even with limited training data, whereas the Vision Transformer struggled to learn under such constraints.

To validate the neural network's ability to detect deepfakes generated by methods not used in its training set, a dataset containing a variety of deepfake generation methods and labels is needed. The chosen dataset for this purpose is ForgeryNet [103], which is one of the most comprehensive deepfake datasets available, containing 2.9 million images and 220,000 video clips. The fake images are manipulated using 15 different manipulations while the videos are manipulated using only 8 of them [107]–[116]. To each image and video, more than 36 mix-perturbations are randomly applied on more than 4300 distinct subjects. Examples of applied perturbations are optical distortion, multiplicative noise, random compression, blur and many others shown in more detail in the ForgeryNet paper [103]. Furthermore, the different manipulations applied can be grouped into two macro-categories, *ID-remained* and *ID-replaced*. The first category involves manipulations of the subject's face without changing their identity, while the second category involves replacing the subject's face with a different one. These two categories are further divided into four sub-categories: all the videos falling under the ID-remained category are manipulated with Face Reenactment methods, while the ID-replaced class is divided into Face Transfer, Face Swap, and Face Stacked Manipulation (FSM). These sub-categories collectively make up a significant portion of the deepfake generation techniques currently known. The ForgeryNet dataset includes people in various settings and situations.

The extracted frames are pre-processed, as done in many other deepfake detection methods [106], [117]–[120] by introducing a face extraction step using the state-of-the-art face detector, MTCNN [121]. The models are trained and evaluated on a per-face basis and data augmentation was performed, similar to [117], [119], [122]. However, we extracted the faces to be squared and with an additional 30% padding in order to catch also a portion of the background behind the person. We exploited the Albumentations library [123] and applied common transformations randomly during training. Whenever an image is an input to the network during training, it is randomly resized using three types of isotropic resize with different interpolation methods (area, cubic, or linear). Afterwards, random transformations such as image compression, Gaussian noise, horizontal flip, brightness or saturation distortion, gray-scale conversion, shift, rotation, or scaling, are applied.

The present paper is derived from another joint work between CNR and UNIFI [124] and presented in MAD'22 Workshop where we already conducted a similar cross-forgery analysis on the part of the dataset consisting of still images, in this case, we performed on videos and, in particular, we have made a specific comparison among different kinds of architectures. It is important to carry out this analysis on videos because the anomalies that are introduced in videos can also differ greatly from those that may result from the manipulation of a single image. Therefore, the behaviour of the various deep learning methods can also change greatly. In the ForgeryNet dataset, there is a label assigned to each video indicating whether it has been manipulated or not. Additionally, the label specifies the method employed to perform the manipulation. Among the methods used, FaceShifter and ATVG-Net manipulate all frames of the video, while the other methods partially manipulate the video frames, providing information on which frames have been manipulated and which ones are left unaltered.

To perform this comparative analysis on cross-forgery generalization capability we have considered three kinds of network architectures. Convolutional Neural Networks (CNNs), a widely used type of neural network in computer vision, and Vision Transformers (ViTs) [125], a newer, highly competitive deep learning model. As the representative of the CNN category, we have taken an *EfficientNetV2-M* [126], which is a newer and more advanced version of the well-known EfficientNet. EfficientNets are widely used in deepfake detection and remain a cornerstone of many state-of-the-art methods on leading datasets. On the contrary, for the Vision Transformers, we have used the *ViT-Base*, a ViT of similar dimensions to the CNN which is one of the first versions introduced. However, a further third architecture, such as the Swin Transformer [127], has been taken into account; this has been done because this type of Transformers is particularly interesting for our analysis in that although it is attention-based, the computation of attention takes place in a hierarchical manner emulating the convolutional layers of CNNs. Swin Transformer is an architecture for image classification that improves the traditional transformer approach by using hierarchical feature representations and a window-based attention mechanism. It divides the input image into patches and transforms them into low-dimensional feature vectors using a learnable projection. These vectors are then passed through a series of transformer blocks, organized into stages, to capture spatial and channel-wise dependencies. Finally, the output is passed through a classification head to produce the class probabilities. Swin Transformer achieves state-of-the-art performance while being computationally efficient and scalable to larger image sizes. The Swin Transformer selected for our experiments is the *Swin-Small*.

All the models were pretrained on ImageNet-21k and fine-tuned on sub-datasets from ForgeryNet, which were constructed with a nearly equal balance of fake and real images. To reduce false detections only faces with a confidence level above 95% were included. All networks were trained by keeping freeze a number of layers such that the trained parameters correspond to approximately 45M. In particular, only the last layers of the models considered were made trainable so that the number of parameters was always comparable between the various experiments, while the other layers' weights remained at the values based on the pretraining.

4.1.2.1 Experiments **Single Method Training** Figure 18 shows the accuracy achieved by the three considered models trained in the *Single Method Training* setup with respect to each of the methods comprised within the test set. Looking at the accuracy of the three models, it can be pointed out that the EfficientNetV2-M and the Swin-Small maintain results often above 80% in correspondence of test frames manipulated with the same methods used in the training set (as expected) and, at the same time, obtain a certain degree of generalization. In fact, the same models sometimes succeed in detecting frames manipulated with methods unseen during training, anyway reaching values of accuracy that are quite limited. The case of method number 5 (*MMReplacement*) is rather anomalous, though the detection percentage is often very high indeed; this behaviour is probably induced by the low number of available examples.

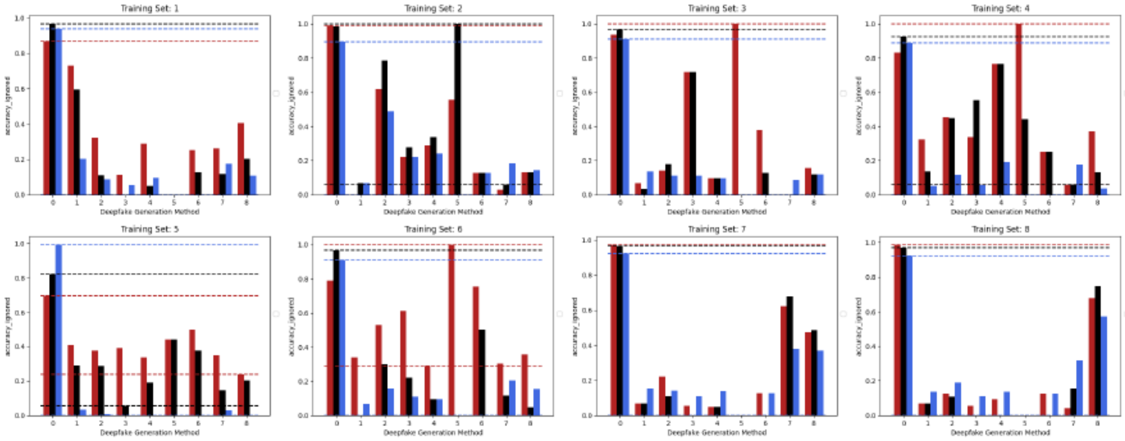


Figure 18. The performance in terms of accuracy achieved by each of the three considered models with respect to the 8 different training sets following the Single Method Training setup: EfficientNetV2-M (red), Swin-Small (black) and ViT-Base (blue).

On the contrary, it can be easily noticed that, in all the cases, the ViT-Base is substantially unable to learn in the presence of relatively few training images. In fact, for instance, by training the model on methods 3, 4, 5 and 6 and then testing it on the test set, it is evident that the model is substantially underfitting and practically unusable compared to the two others taken into consideration. Interestingly, the Swin Transformer, although also based on the attention mechanism, is not particularly affected by this phenomenon and instead succeeds in obtaining competitive results in all contexts. This probably lies in the hierarchical nature that emulates the convolutional layers of traditional CNNs and thus allows it to exploit implicit inductive biases better. Good performances are anyway preserved, in any case, with respect to pristine frame detection. In this setup, the architecture based on a convolutional network seems to prove more capable of generalization. The accuracies obtained from the three models are also shown in the confusion matrices in Figure 19 where all previously commented trends are reconfirmed again.

Multiple Methods Training The behaviour of the three networks is now analysed in the second considered setup, namely *Multiple Methods Training*, and corresponding results are shown in Figure 20. In this case, the datasets are composed of frames extracted from videos manipulated by not only one method, so the models will have more difficulty focusing on specific artifacts and be forced to generalize. In this setup, the situation is significantly different from the previous one. Surprisingly, the classic Vision Transformer, which previously struggled to train effectively, is now

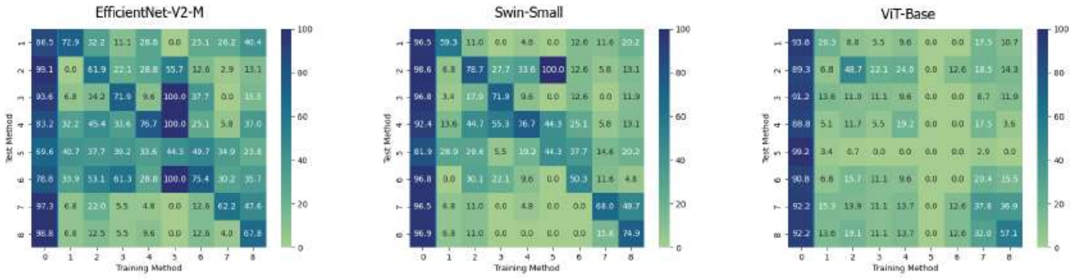


Figure 19. Confusion matrices of the frame-level accuracy values for the three models under consideration trained in the Single Method Training setup and tested on frames manipulated with all the available methods respectively.

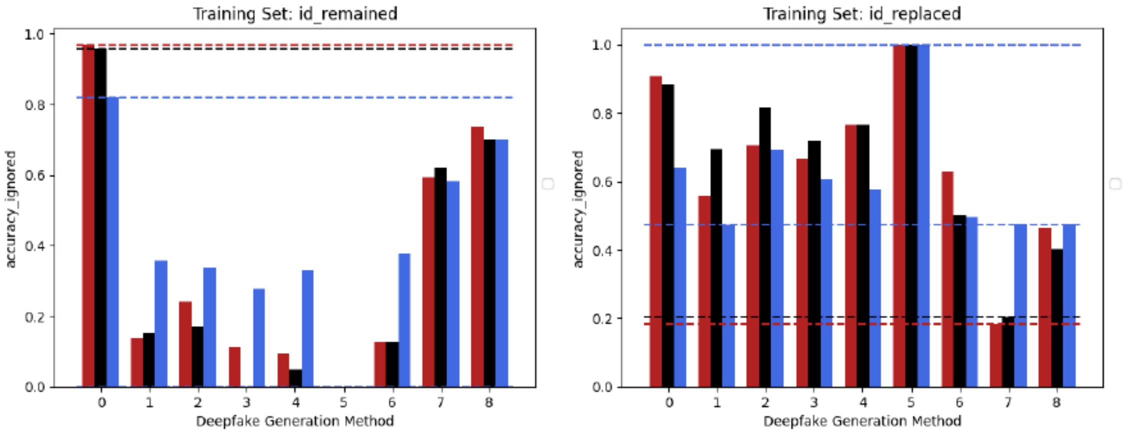


Figure 20. Accuracy performances achieved by each of the models considered in the two different training sets constructed following the Multiple Methods Training setup: EfficientNetV2-M (red), Swin-Small (black) and ViT-Base (blue). ID-Replaced methods (1-6), ID-Remained methods (7-8) and Pristine (0).

the only model capable of generalizing well to frames that have been manipulated using techniques that were not present in the training data. This result probably stems from the fact that the training set consists of significantly more images than in the previous setup and it is strongly in line with what is presented in [124]. This particular architecture shows in many contexts a major need for data and resources which, when available, enable it to achieve very competitive results. In this case, the confusion matrices (see Figure 21) clearly show the greater generalization capacity of the Vision Transformer although at the expense of more false positives. In fact, the 'pristine' class is less accurately classified by this latter. This may be a problem since in a real-world context we may want to lower down as much as possible the number of false alarms, in particular, if the system is fully automatized.

4.1.2.2 Conclusions In this study, we investigated the generalization capabilities for detecting deepfake videos of various deep learning architectures. Our findings suggest that in real-world scenarios where large, diverse deepfake detection datasets are available and generalization is critical, the Vision Transformer may be the optimal choice for detecting deepfakes. However, in cases where the training data is limited, a convolutional network such as EfficientNet-V2 may be more suitable and be considered a good enough alternative. The Swin Transformer provides a good balance between the two in terms of generalization and performance demonstrating a good generalization capability in all the considered contexts and a pretty low false positive rate. It also

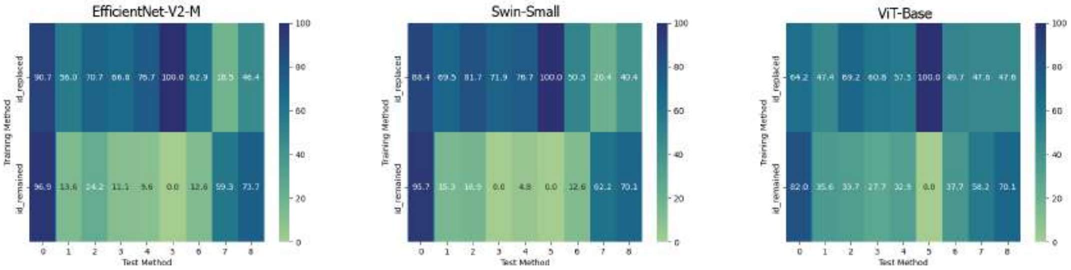


Figure 21. Confusion matrices of the frame-level accuracy of the three models trained in the Multiple Method Training setup and tested on frames manipulated with all available methods.

Model	**Train set**	**AUC**
Face X-ray[128]	FF++	65.5
Patch-based**patch**	FF++	65.6
DSP-FWA**Li˙2019˙CVPR˙Workshops**	FF++	67.3
CSN[129]	FF++	68.1
Multi-Task**multitask**	FF++	68.1
CNN-GRU**cnngru**	FF++	68.9
Xception[130]	FF++	70.9
CNN-aug**cnn-aug**	FF++	72.1
LipForensics**lip**	FF++	73.5
FTCN[118]	FF++	74.0
RealForensics[129]	FF++	75.9
RealForensics[129]	FF++	75.9
iCaps-Dfake**icaps**	FF++	76.8
MINTIME-XC[117]	ForgeryNet (All)	77.9
EfficientNet-V2-M	ForgeryNet (ID-Remained)	50.0
	ForgeryNet (ID-Replaced)	50.1
ViT-Base	ForgeryNet (ID-Remained)	51.0
	ForgeryNet (ID-Replaced)	57.2
Swin-Small	ForgeryNet (ID-Remained)	58.7
	ForgeryNet (ID-Replaced)	71.2

Table 8. Cross-Dataset comparison of video-level AUC on DFDC Preview test set.

results in significantly more capable of generalising the concept of deepfake when tested in a cross-dataset scenario. This suggests that probably the attention mechanisms may enable the models to generalize better the concept of deepfakes but only when enough data are provided. Overall, our study highlights the significance of considering the specific characteristics of the dataset and deep learning architecture when detecting deepfakes to be able to create a detector which may be applied in the real world. Future work will be further dedicated to analyse and possibly improve the level of generalization of such AI-based instruments.

4.1.2.3 Relevant publications

- Coccomini, D.A.; Caldelli, R.; Falchi, F.; Gennaro, C. On the Generalization of Deep Learning Models in Video Deepfake Detection. J. Imaging 2023, 9, 89. https://doi.org/10.3390/jimaging9050089

4.1.2.4 Relevant software and/or external resources

- The Pytorch implementation can be found in `https://github.com/davide-coccomini/Cross-Forgery-Video-Deepfake-Detection`.

4.1.2.5 Relevance to AI4media use cases and media industry applications The methods developed in this work contribute to UC1 (AI for Social Media and Against Disinformation), and specifically, Feature 1A (Detection/Verification of Synthetic Media). Such systems could be adopted by journalists which use AI-based tools to verify the authenticity of multimedia contents in a real-world scenario. In fact, in this case it is important to establish the degree of generalization provided by such AI-based instruments when they are called to operate in a context that is rather different with respect to that of the training phase and, above all, against possible forgeries not seen during training.

4.1.3 Autoencoder-based Data Augmentation for Deepfake Detection

Contributing partner: UPB

Generative models have evolved a lot in the last few years. Less than a decade ago, it was not possible to generate an image realistic enough to not be easily distinguished with the naked eye, and now the technology allows us to generate realistic images depicting anything, using models like Generative Adversarial Networks (GAN) [28] or Stable Diffusion [131]. With these generative models came a new threat to society: deepfakes. Deepfakes are deep learning generated images or videos that usually portray humans and can easily mislead or disinform the people that are viewing them.

With the rise of deepfakes, there is a constant battle between image generators and fake detectors. As generators get better and images get more realistic, deepfake detectors must keep up and help erase disinformation.

Deepfake detection methods are very varied, ranging from Convolutional Neural Networks, using image features like the frequency spectrum [132]–[134] to detect some ”generator fingerprint” or using temporal neural networks like 3DCNN, LSTM or Transformers [135]–[137], and much more. The majority of those methods achieve a good performance on the datasets they were tested on, some of them having an AUC % or over 99%. While that proves that deepfake detection methods are very efficient, this is not the full story. When testing on other datasets or real-world data, the majority of the approaches presented above experience a big loss in performance. This is due to the fact that usually, deepfake detectors would be trained on one dataset, containing real samples and fakes, generated with some generator architectures. The detectors would learn the "fingerprints" of those generators, but that would not be efficient when testing on other deepfakes, as they are generated using other models. Therefore, we can say that deepfake detectors have a generalization problem.

There have been a few approaches aimed at generalization like LipForensics [136] or Self-Blended Image Augmenatation [138] which improved performance versus unseen datasets, but the problem is still not solved. More than that, when images are uploaded to a website in real life, they could be affected by changes like noisy compression, resolution changes and so on.

To help overcome this challenge, we propose an Augmentation method using Autoencoders. The augmentation technique is based on trying to eliminate or change the generator-specific artefacts found in images and in their frequency fingerprints. More than that, we try to introduce new types of artefacts in the images. By doing that, we can train the detectors on new possible image perturbations. Our approach is presented in Figure 22, and it is composed of the following elements:

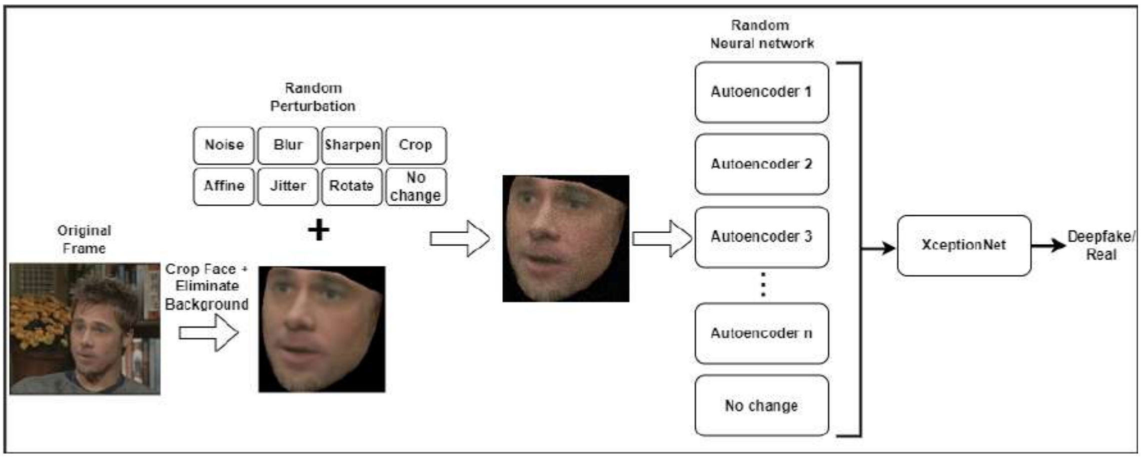


Figure 22. The Training Pipeline using our proposed Autoencoder Data Augmentation for deepfake detection

- Starting from a video frame, either real or fake, the facial region is cropped and the background is eliminated.
- Random perturbations are added to the image to increase robustness.
- More than 80 autoencoders and U-Net models with different architectures, sizes, kernel dimensions and upsampling techniques are used to generate a similar image. The autoencoders were trained beforehand on real videos and their role is to reproduce the exact image that is passed as an input, or as close to it as possible, but to insert their neural network "signature" in the image, altering or eliminating the generator signature. More than that, the autoencoders were trained to recreate the image with different levels of noise, starting from 3% difference to 20%. This way, they would also introduce new artefacts and the image would not be 100% identical to the original. The image goes through one of the autoencoders chosen randomly.
- A deepfake detector is trained using the images coming from the autoencoders, as well as the original images.

4.1.3.1 Experiments For our setup we use FaceForensics++ [139] as a training dataset for the deepfake generator. We use the slightly compressed HQ version. To evaluate the generalization ability of the model, we test it on CelebDF [140] and DFDC Preview [141] datasets.

To train the autoencoders, we use the real videos from DFDC [142]. We use L1 loss to evaluate their performance, but we penalize the loss function if the error is less than 3%. This way, the autoencoders would introduce some kind of noise and perturbations. We also use some models that lose a lot of information such as AE1, AE2 and Unet1. Figure 23 presents some frames produced by different autoencoder models. Some of the frames are identical to the original, some have small errors in shades or skin color and the last row has a lot more inconsistencies, but still presents somewhat realistic images.

We trained 3 different neural network models for comparison: XceptionNet, Capsule Networks and EfficientNetB4, using both basic augmentation (blurring, random noise, random sharpening, cropping, affine transforms etc) as well as the Autoencoder Augmentation presented in Figure 22. Table 9 presents a comparison for the 3 models, each trained using both training paradigms. The results are for models trained on FaceForensics++ and evaluated on CelebDF, to show the generalization capability of the proposed method. As we can see, the Autoencoder-based augmentation yields better results for generalization than a basic training, for all 3 models. CapsNet models

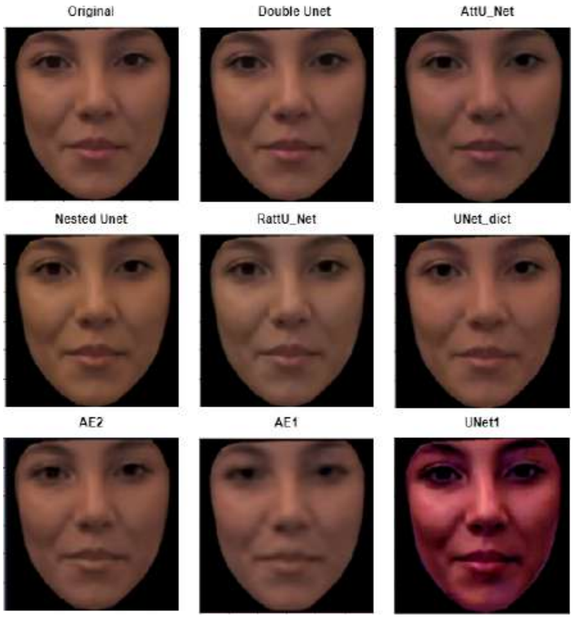


Figure 23. Comparative results of a frame recreated using multiple autoencoders. The quality is gradually lost, from top to bottom, but the majority of the images are still somewhat realistic.

have the lowest generalization capability, as they tend to overfit the most. On the other hand, EfficientNetB4 with basic augmentation tends to overfit the data from FaceForensics++, losing its ability to generalize on other datasets. However, the AE augmentation increases generalization by almost 9%. The XceptionNet model is somehow balanced, but still sees improvements of 5% for generalization when using AE augmentation.

Table 10 presents a comparison of the generalization ability of the XceptionNet model using different levels of augmentation:

1. No data augmentation
2. Only basic data augmentation (blurring, random noise, random sharpening, cropping, affine transforms etc)
3. Data augmentation with the 3 worst basic autoencoder models (2x AE1, 2xAE2, 4x UNet1)

Method	CelebDF AUC [%]
Xception - basic augmentation	68.96
Xception - AE augmentation	73.67
CapsNet - basic augmentation	64.32
CapsNet - AE augmentation	70.4
EfficientNetB4 - basic augmentation	66.93
EfficientNetB4 - AE augmentation	**75.85**

Table 9. Comparison of generalization for models trained FaceForensics++ and evaluated on CelebDF, with basic and Autoencoder augmentation

Method	CelebDF AUC [%]
Xception - no aug (1)	64.55
Xception - basic aug (2)	68.96
Xception - AE aug (3)	73.67
Xception - AE aug (4)	80.73
Xception - Full AE aug (5)	**82.62**
Xception - AE aug (6)	80.61

Table 10. Comparison between the generalization ability of a XceptionNet model, with different levels of augmentation, trained on FaceFGorensics++ and evaluated on CelebDF

4. Data augmentation with all the trained autoencoder and U-Net architectures, one model per architecture, for a total of 10 models
5. Data augmentation with with all the models mentioned above, multiple models at different epochs and multiple loss functions per architecture, for a total of 80 models
6. Data augmentation with all the models mentioned above, multiple models at different epochs and multiple loss functions per architecture, but without also using basic data augmentation like blurring, noise, color jitter etc.

Table 10 shows that each incremental step was necessary in increasing the performance of our detector.

Lastly, Table 11 presents a comparison between our proposed method and the state of the art. While our methods do not outperform the state of the art, there are a few positive key takeaways regarding their performance:

- Our Autoencoder Augmentation approach is aimed to improve the robustness of any existing deepfake detector. The approach can be used with any model in the training phase.
- While we did not obtain state-of-the-art best results, our results clearly show that this method brings a significant improvement. For example, Our proposed XceptionNet model outperforms a simple XceptionNet model without augmentation by almost 10% on CelebDF and by 1% on DFDC Preview.
- Our method outperforms some state-of-the-art models while using a very basic CNN model, in comparison to other methods using video-level approaches.

4.1.3.2 Conclusions In this study, we investigate a method to increase generalization for deepfake detectors using an Autoencoder Augmentation algorithm. The algorithm, uses a multitude of pretrained autoencoders to recreate images. This way, the "model signature" obtained from the deepfake generator could be modified or erased and a new signature would be added. The models are varied to obtain more diverse results. More than that, the autoencoders introduce some intentional error of their own, as they are pretrained to do so.

Our experiments showed that this method improves generalization by a significant amount on multiple models. More than that, while our approach does not achieve state-of-the-art level results, it is an incremental step that can be applied to any deepfake detection algorithm to improve robustness and generalization.

However, this method also has certain limitations, such as (1) being a frame-level approach or (2) having to use a multitude or pretrained models to obtain the necessary augmentations. We

Method	CelebDF AUC [%]	DFDC Preview AUC [%]
LipForensics [136]	82.4	73.5
Face X-Ray [128]	79.5	65.5
3D R50-FTCN [137]	86.9	74.0
Xception [139]	73.7	70.9
CNN-aug [143]	75.6	72.1
PCL + I2G [144]	90.03	74.37
EFNB4 + SBIs [138]	**93.18**	**86.15**
Ours - Xception Full AE aug (5)	82.62	71.52
Ours - EfficientNetB4 Full AE aug (5)	82.87	72.6

Table 11. Comparison of generalization for models trained FaceForensics++ and evaluated on CelebDF and DFDC Preview

plan to overcome one of those limitations by extending the model to video-level, by augmenting a multitude of consecutive frames with 3D CNN based autoencoders.

4.1.3.3 Relevant Publications

- Stanciu, Dan-Cristian, and Bogdan Ionescu. "Autoencoder-based Data Augmentation for Deepfake Detection." In Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation, pp. 19-27. 2023. https://dl.acm.org/doi/abs/10.1145/3592572.3592840

4.1.3.4 Relevant software, datasets and other resources

- The Pytorch implementation can be found in https://github.com/StanciuC12/deepfake-detection-ae-augmentation.

4.1.3.5 Relevance to AI4media use cases and media industry applications The methods developed in this work contribute to UC1 (AI for Social Media and Against Disinformation), and Feature 1A (Detection/Verification of Synthetic Media). Deepfake detection systems could be used by either social media platforms to automatically verify the authenticity of uploaded content, as well as journalists to verify real-world content. Methods such as this can be used to help the models generalize, making sure that all the cases are covered. More than that, our proposed method shows a good performance against attacks or sources of noise which can impact the ability for a model to detect a fake.

4.1.4 Reliability of deepfake detection

Contributing partner: CERTH

Since the landmark publication of [139], deep neural networks have dominated the field of deepfake detection as they were shown to be more accurate than their shallow counterparts. Deep networks however tend to overfit their training data and perform poorly on unseen manipulations [145]. This has sparked intense research interest in the issue of generalization, which involves extracting features that are sufficiently discriminative for general types of deepfakes. In this work, we focus on the largely unexplored issue of reliability of deepfake detection, focusing on face manipulations.

The motivation of our approach is that it is presently unclear if such discriminative features or a general class of deepfakes exist. Deepfake detection differs from standard object detection in that models try to detect artefacts representative of manipulation algorithms instead of semantic objects. This implies that deepfake detection is fundamentally an *attribution* problem, and framing it as such has been shown to increase detection accuracy [146]. A further consequence is that, considering the evolution of deepfake generation algorithms, deepfakes could conceivably be unrecognizable from real media in the future. Due to the above, we argue that it is critical to enhance the reliability of deepfake detectors so that they detect known artefacts with high confidence. In this way, even if false negatives are unavoidable (unless the manipulation method is close to a seen one), false positives should be better addressed.

A key consideration for reliability is *uncertainty quantification*, i.e., estimating the confidence of a model's decisions, which has also received considerable attention in recent years [147]. Ideally, AI models should estimate the confidence of their decisions and refrain from producing an output if their confidence is low. This is possible by acknowledging that the accuracy of AI models drops for inputs that stray away from the training data. While there have been many approaches to uncertainty quantification, e.g., softmax calibration, predict with reject options, and Bayesian neural networks, in our experiments, we have focused on conformal prediction [148], which can yield statistically meaningful estimates from *any* uncertainty measure.

4.1.4.1 Experiments Addressing deepfake detection at the frame level, we consider a labelled collection of real and fake face images (X, Y) where the fake faces come from K manipulation classes ($Y \in [0, K]$ with 0 corresponding to real faces). We also consider a convolutional model trained to discriminate over the manipulation classes, whose output can be easily become binary by converting the fake labels to 1.

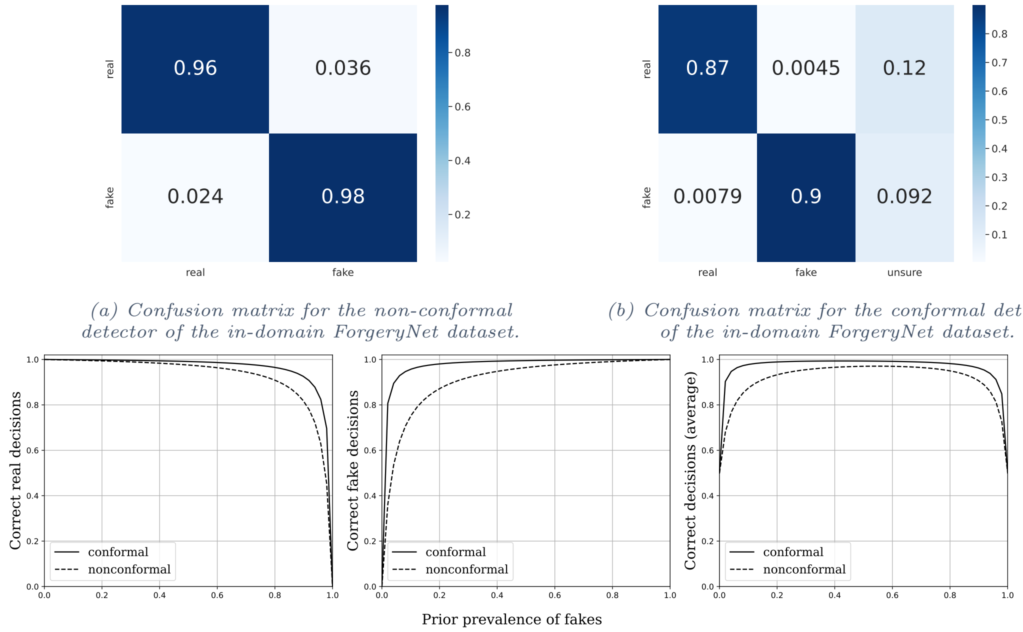
The conformal prediction framework requires i) a non-conformity score function $s(x, y)$ that quantifies the uncertainty of a prediction, and ii) a percentage α that expresses the target probability of correct classification after calibration. For the non-conformity score, we choose the softmax output of the true class based on its interpretation as an estimate of the posterior distribution $p(x|y)$:

$$s_i = 1 - \hat{p}(x|y_i). \tag{1}$$

Based on the *split conformal prediction* method, we further reserve a portion of the training data as calibration data to calculate a threshold $\hat{q}$ for the score, which is the $\lceil(|D|+1)(1-\alpha)\rceil$ $|D|$-th percentile of the calibration scores. This threshold is used at test time to output prediction sets $C(x_{test})$ for test samples x_{test} as:

$$C(x_{test}) = \{y : s(x, y) \leq \hat{q}\}. \tag{2}$$

The power of conformal prediction is that, regardless of the underlying distribution, the prediction sets are guaranteed to hold the correct label with probability $1 - \alpha$, provided that the test data come from the same distribution as the training data [148]. While this condition does not hold for out-of-distribution data, the presence of such data can be signaled through an empty prediction set. A conformal version of our detector thus outputs three values: real and fake based on the class of the maximum probability and uncertain if the prediction sets are empty.



(a) Confusion matrix for the non-conformal detector of the in-domain ForgeryNet dataset.

(b) Confusion matrix for the conformal detector of the in-domain ForgeryNet dataset.

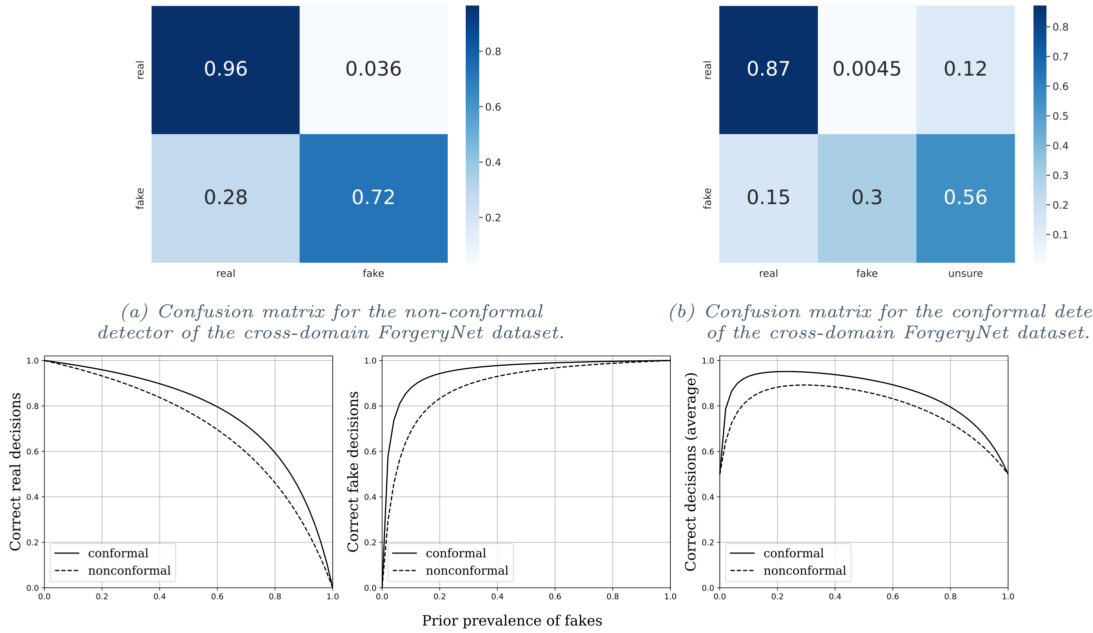
(c) Posterior detection probabilities for the in-domain ForgeryNet dataset.

Figure 24. Evaluation results for the in-domain ForgeryNet dataset (manipulation classes 0, 1, 4, 6, 7, 11).

For our experiments, we trained an EfficientNet-B0 model on a subset of the ForgeryNet image dataset [103]. This architecture was chosen for being simple yet capable, and this dataset for its large number of samples (>2M) and diverse manipulation methods (15 methods covering face transfer, swap, reenactment, editing, and combinations). In detail, we sampled 14K images from each manipulation category [5] and grouped one category per manipulation type (0, 1, 4, 6, 7, & 11) for an in-domain detection scenario and the remaining types for a cross-domain detection scenario. The in-domain samples were split with a (7,1,1,1) ratio for the training, validation, calibration, and test splits, while the cross-domain samples were enriched with the FF++, CelebDF, DFDC preview, and DFDC datasets. Since the latter are video datasets, they were sampled at 1 fps, and the faces were extracted with the MTCNN face detector cropped with 1.3 margin and resized to the target size of EfficientNet-B0. For the training, we used pretrained weights from ImageNet, the Adam optimizer with learning rate 0.0005, 0.00005 weight decay, and fine-tuned for 30 epochs with early stopping.

Figs. 24-29 show the confusion matrix of our detector with and without conformal decisions, as well as the posterior probability of a correct decision. Regarding the confusion matrices, we see that the true positive and negative rates of the non-conformal detector are >95% in the in-domain scenario, relatively high in the cross-domain scenario of ForgeryNet, and drop considerably in the cross-domain scenarios. In the cross-domain scenarios, it is noteworthy that the detector's outcome is not equivalent to pure chance but actively mistakes unseen manipulations as real faces, as described before. Using the conformal prediction framework, however, we see that a significant portion of fake decisions are moved to the uncertain category. This implies that although (unseen) fake faces register as real, they do not reach the confidence threshold to be categorized as such; hence, uncertainty remains over their classification. There is also uncertainty over the classification

[5] 14K was the number of samples from the smallest category.

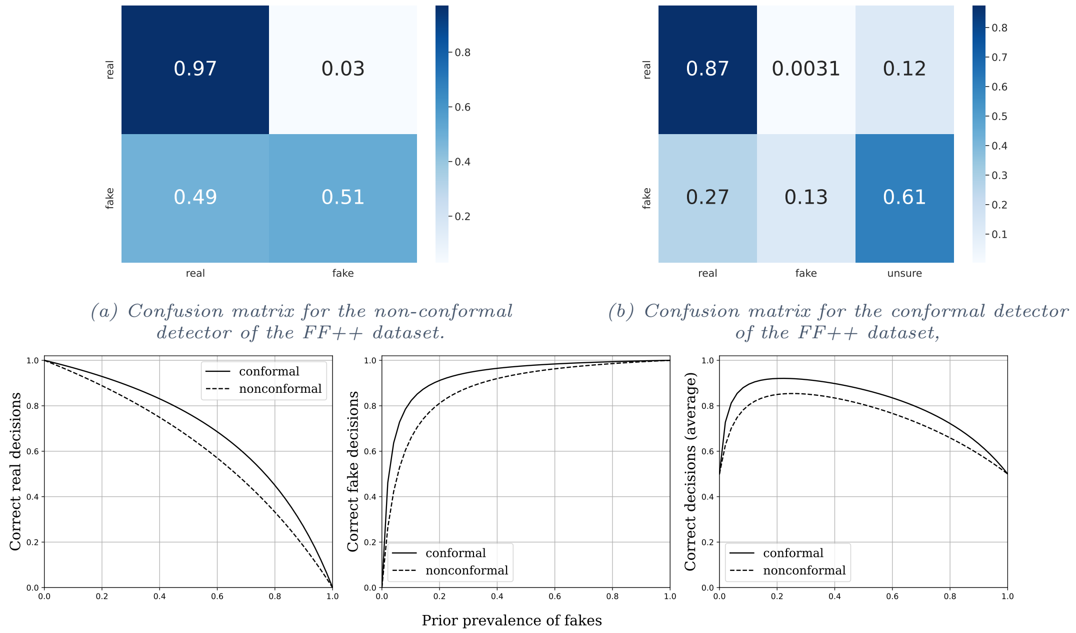


(a) Confusion matrix for the non-conformal detector of the cross-domain ForgeryNet dataset.

(b) Confusion matrix for the conformal detector of the cross-domain ForgeryNet dataset.

(c) Posterior detection probabilities for the in-domain ForgeryNet dataset.

Figure 25. Evaluation results for the cross-domain ForgeryNet dataset (manipulation classes 0, 2, 3, 5, 8, 9, 12, 13, 14, 15).

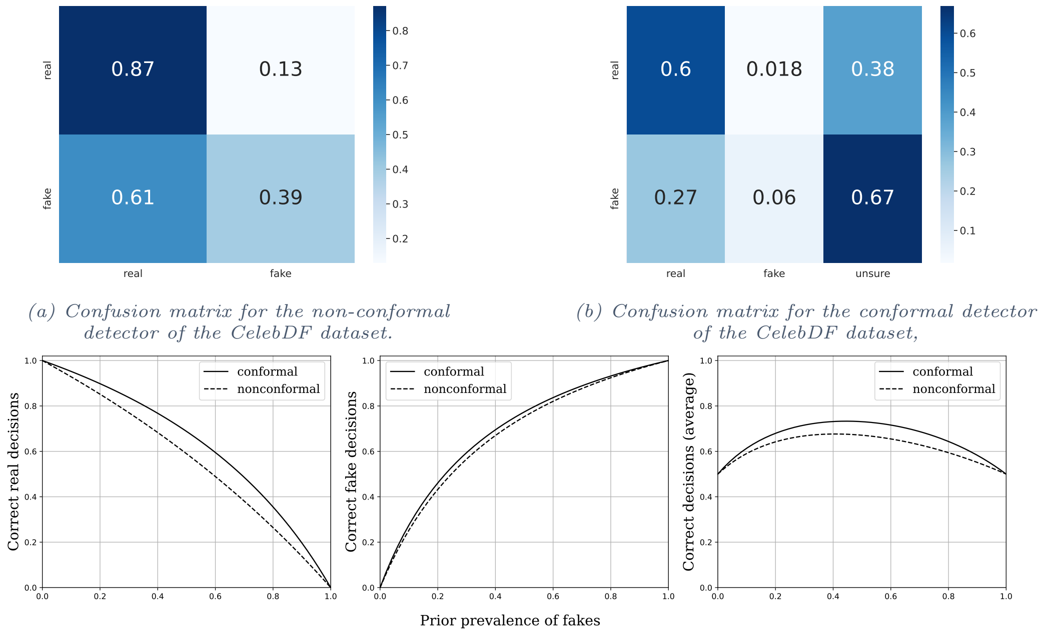


(a) Confusion matrix for the non-conformal detector of the FF++ dataset.

(b) Confusion matrix for the conformal detector of the FF++ dataset,

(c) Posterior detection probabilities for the FF++ dataset.

Figure 26. Evaluation results for the FF++ dataset (raw quality).

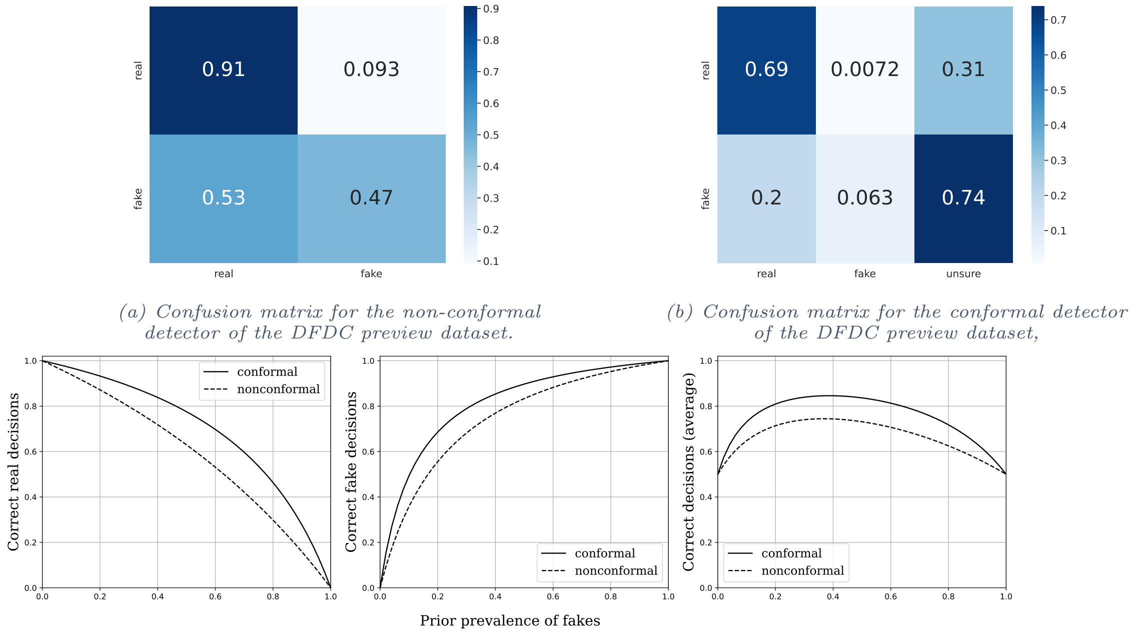


(a) Confusion matrix for the non-conformal detector of the CelebDF dataset.

(b) Confusion matrix for the conformal detector of the CelebDF dataset,

(c) Posterior detection probabilities for the CelebDF dataset.

Figure 27. Evaluation results for the CelebDF dataset.

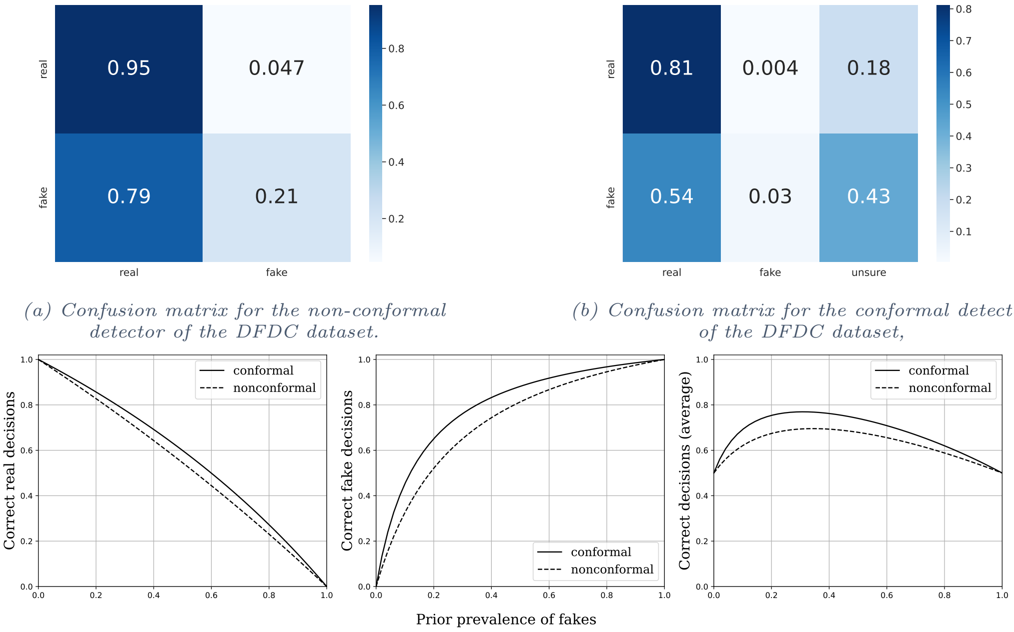


(a) Confusion matrix for the non-conformal detector of the DFDC preview dataset.

(b) Confusion matrix for the conformal detector of the DFDC preview dataset,

(c) Posterior detection probabilities for the DFDC preview dataset.

Figure 28. Evaluation results for the DFDC preview dataset.



(a) Confusion matrix for the non-conformal detector of the DFDC dataset.

(b) Confusion matrix for the conformal detector of the DFDC dataset,

(c) Posterior detection probabilities for the DFDC dataset.

Figure 29. Evaluation results for the DFDC dataset.

of real faces, especially in the CelebDF and the DFDC preview datasets, but it is significantly lower than the uncertainty of fake faces, which attests to the fact that our detector has learned real faces better.

To evaluate the change in the reliability of the detection, we resorted to the posterior probability correct decisions, i.e., the probability of a given decision being correct, which is different from the probability of a given image being classified correctly. From Bayes' theorem, this probability inconveniently depends on the prior prevalence of fake and real images which is generally unknown. To address this issue, based on the forward probabilities of the confusion matrices, we have plotted the posterior probabilities of correct decisions as a function of the prevalence of fake images. The graphs show that the reliability of correct decisions indeed increases with conformal models with an enhancement that reaches even 10%. Interestingly, the reliability increases for both real and fake decisions which shows that our approach can help even with unseen manipulations, if their artefacts resemble a known manipulation.

4.1.4.2 Conclusions In this work, looking beyond generalization, we have focused on the reliability of deepfake detection, which is an overlooked area in the literature. In particular, we have framed detection as an attribution instead of a binary classification task, and used the conformal prediction framework to capture uncertainty over ambiguous inputs. This approach increases the reliability of observed decisions by converting both false positives and false negatives to uncertain outputs. It is further agnostic to the underlying model architecture, which aids to its practical usefulness and applicability.

The main limitation of our work is that the theoretical guarantees of conformal prediction do not strictly apply to out-of-distribution data, despite our promising results. Hence, we expect the performance of our method to hinge on the existence of sufficiently diverse training samples in terms of manipulation techniques, which is limited by the existing datasets. In addition, the conformal prediction approach can be further optimized, e.g., via the choice of different conformal score functions. More generally, conformal prediction is just one approach of uncertainty quantification, which is currently a rapidly evolving area. Different techniques, e.g., based on the Bayesian framework, could be also effective for deepfake detection, which remains to be investigated.

Based on the above, our work could readily be extended to include different design parameters of conformal prediction like the score function, as well as different frameworks of uncertainty quantification. Another promising line of work is investigating the impact of diversity of the training data to our models' confidence. Ultimately, fuelled by the probabilistic nature of AI algorithms and the demand for trustworthiness, we believe that the importance of reliability will only grow and our work aspires to be a first step towards this direction.

4.1.4.3 Relevant publications

- No relevant publications published at this moment.

4.1.4.4 Relevant software, datasets and other resources

- The code is publicly available in https://github.com/ngiatsog/deepfakes-reliability.

4.1.4.5 Relevance to AI4media use cases and media industry applications Our approach is related to UC1 (AI for Social Media and Against Disinformation), specifically, to Features 1A (Detection/Verification of Synthetic Media) and 1E (Capability for Trustworthy AI by Design). According to 1A, journalists should use AI tools to evaluate the authenticity of multimedia content and aid in the verification process. The opaque inner workings, however, of most AI models can easily predispose the journalists to accept their outputs as truth. In reality, AI models are statistical and unreliable when they extrapolate beyond their training data, hence, their outputs are much less useful when they are not paired with confidence estimates. Therefore, by integrating our methods to UC1, we expect to increase journalists' and fact-checkers' trust in the detector's results, and facilitate adoption of deepfake detection tools by media professionals. Beyond UC1, reliability estimates are important for other media-related applications such as disinformation detection, moderation, and authentication in social networks.

4.1.5 Addressing real-world constraints of synthetic image detection

Contributing partner: CERTH

Distinguishing between generated and authentic images has attracted the interest of numerous researchers in the field of media forensics [149], [150]. The most widely used approach for detecting generated images involves training a neural network on a binary classification task (real vs fake) using a large corpus of labelled images. A key component in this development process is the use of a set of carefully selected image augmentations during training as demonstrated in prior work [149], [150]. Our proposed approach relies on this supervised learning setup.

Additionally, the generalization of the detectors is mostly studied with respect to the different generative architectures [143], [150].Herein, we take a different avenue and define generalization as the ability to detect synthesized images that depict different concept classes. We refer to this capability as generalization in *cross-concept settings*, which is often referred to as *Domain Generalization* in literature [151]. We study the behavior of the detector in such a way that it could be trained in one concept class of real and fake images, e.g., *human faces*, and can then be used to distinguish between real and fake images also in other concepts, e.g., *animal faces*.

We empirically find that using standard practices from the literature to train our detectors is not effective enough to generalize on cross-concept scenarios.

In this research, we tackle the cross-concept generalization challenge of synthetic image detectors by proposing a sampling strategy for the selection of generated images used for training. Prior work [143], [150] relies on randomly generating a large number of images that are used for training. Instead, we assess image quality based on a probabilistic Quality Calculation (QC) score, and then select the top-k images in terms of quality to train our detectors. This is under the assumption that high quality images will lead the network to focus less on the artifacts of the generative process and more on the characteristics that are invariant to the image content. We evaluate our method using fake images generated by StyleGAN2 [152] and the unconditional module of the recently introduced Latent Diffusion [131]. When training with the proposed sampling strategy, the performance is considerably improved compared to random selection.

4.1.5.1 Experiments. This section provides a summary of the conducted experiments.
Datasets: We obtain real images from publicly available datasets, FFHQ [153], AFHQ [30], and LSUN [154], containing images of human faces (FFHQ), dogs, cats, and a general class of wildlife (AFHQ), and nearly one million images of 10 scene categories and 20 object classes (LSUN). Then, we employ pretrained diffusion and GAN models to generate artificially generated images for different classes to evaluate the cross-concept scenario. More specifically, we use StyleGAN2 [40] to generate images from pretrained networks in FFHQ, AFHQ and LSUN-churches, while we also use Latent Diffusion [131] to generate images from pretrained networks in FFHQ, LSUN-bedrooms, and LSUN-churches. We denote each dataset as $\mathcal{X} + \mathcal{X}^A$, where $\mathcal{X}$ is the set of real images of a specific concept class and $\mathcal{X}^A$ is the corresponding set of fake images generated with architecture A, which takes values either G for StyleGAN2 or D for Latent Diffusion. Regarding the different concepts, we use $\mathcal{H}$ for human, $\mathcal{A}$ for animal, $\mathcal{C}$ for church, and $\mathcal{B}$ for bedroom datasets.
Cross-concept evaluation: Table 12 presents the AUC scores of detectors trained with random sampling versus the proposed method for images based on StyleGAN2 and Latent Diffusion. We run three training sessions for each concept and report the mean and standard deviation of AUC.

We first analyse the results of our intra-class experiments, where we train and evaluate a detector in the same concept class. The corresponding runs are coloured in gray in Table 12. Our findings indicate that there is no significant difference in the performance of the detector when using images of higher quality, due to the perfect performance in almost all cases (with the exception of the Bedroom concept in Latent Diffusion). This is consistent with previous studies that have shown that a robust augmentation scheme is sufficient to train very accurate detectors within the same domain [143], [150]. In our study, we used the same augmentation scheme for both randomly selected and quality-based selected subsets, which explains the lack of notable difference in the detector's performance between the two subsets.

Next, we discuss the results of our proposed methodology in the case where a detector is trained on generated and real images from a concept class and is evaluated on images of a different class. Our proposed method clearly improves the robustness of the StyleGAN2 detector in almost all cases (with improvements between 5 and 8.5%) except in the case when training on Animals ($\mathcal{A}$+$\mathcal{A}^G$) and testing on Churches ($\mathcal{C}$+$\mathcal{C}^G$). Furthermore, the standard deviation of the runs with the proposed QC sampling strategy is significantly lower compared with the runs where random sampling was used. This implies that using our strategy leads to more consistent and reliable models. Similar conclusions can be drawn from the experiments on Latent Diffusion sets. The AUC score is improved by almost 10% in some cases, and the standard deviation is generally low, except for one case. Overall, our experiments provide evidence of the effectiveness of the proposed method in improving the generalization performance across different concepts.

		test		
training	sampling	$\mathcal{H}$+$\mathcal{H}^G$	$\mathcal{A}$+$\mathcal{A}^G$	$\mathcal{C}$+$\mathcal{C}^G$
$\mathcal{H}$+$\mathcal{H}^G$	random	100.0±0.0	89.1±1.00	77.5±3.42
	QC (ours)	100.0±0.0	**94.8±1.22**	**83.5±0.94**
$\mathcal{A}$+$\mathcal{A}^G$	random	61.8±1.36	100.0±0.0	**89.1±0.71**
	QC (ours)	**66.4±1.08**	100.0±0.0	86.5±0.15
$\mathcal{C}$+$\mathcal{C}^G$	random	53.7±0.22	58.6±3.51	100.0±0.0
	QC (ours)	**59.4±0.19**	**67.1±0.16**	100.0±0.0

(a) StyleGAN2

		test		
training	sampling	$\mathcal{H}$+$\mathcal{H}^D$	$\mathcal{B}$+$\mathcal{B}^D$	$\mathcal{C}$+$\mathcal{C}^D$
$\mathcal{H}$+$\mathcal{H}^D$	random	100.0±0.0	64.4±2.31	66.7±0.88
	QC (ours)	100.0±0.0	**74.1±0.68**	**77.5±0.03**
$\mathcal{B}$+$\mathcal{B}^D$	random	52.1±1.47	96.3±0.79	**99.7±0.03**
	QC (ours)	**56.2±5.13**	**99.6±0.02**	99.4±0.02
$\mathcal{C}$+$\mathcal{C}^D$	random	54.2±4.08	96.3±1.19	100.0±0.0
	QC (ours)	**58.5±1.52**	**98.9±0.58**	100.0±0.0

(b) Latent Diffusion

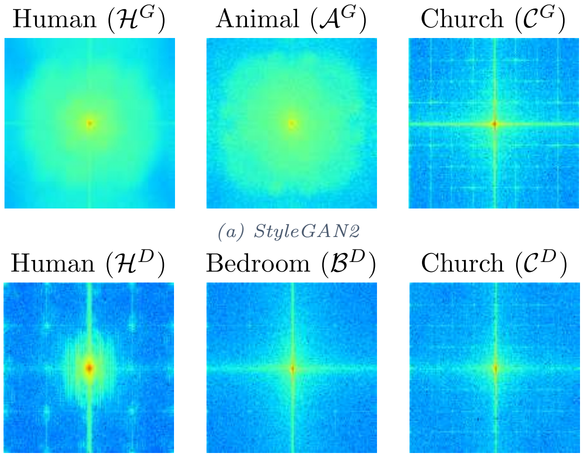
Table 12. AUC of detection model trained on randomly selected samples or based on the proposed QC score for each concept and generative model. Mean and standard deviation of three training sessions with different seeds are reported. ***Bold*** *indicates the best performance between the proposed and the random baseline. Gray colour indicates intra-concept evaluation.*

Results when test image quality varies: We also evaluate our model in the case of selecting different test set composition in terms of image quality. Specifically, the aim is to evaluate whether images of better quality are more difficult to detect. Hence, we split the test sets into four quartiles denoted as 0-25% (lowest quality), 25-50%, 50-75%, and 75-100% (highest quality) based on their QC score. We observe that in several cases of StyleGAN2 images, the detector achieves better AUC scores for low- or medium-quality images. This means that a test set would benefit from an automatic selection step based on quality, since it would be more challenging. Instead, in the case of images generated using the Latent Diffusion model, there is no apparent association between the test set composition in terms of image quality and the detection performance.

Results when the detector is trained with many classes and architectures: We also evaluate model performance when training based on QC sampling in three different datasets:

- $\mathcal{G}$: images from all concepts generated by StyleGAN2 and real images.
- $\mathcal{D}$: images from all concepts generated by Latent Diffusion and real images.
- $\mathcal{G} + \mathcal{D}$: Combination of the other two datasets.

The results in Table 13 indicate that the detector performance significantly improves when it is trained on a diverse set of generated images, including those produced by StyleGAN2 and Latent Diffusion models, and evaluated on a range of test sets. Conversely, we observe that detectors



(a) StyleGAN2

(b) Latent Diffusion

Figure 30. Visualization of the magnitude spectrograms for the different classes of the two generative architectures using the denoising network proposed in [155].

	test		
training	$\mathcal{G}$	$\mathcal{D}$	$\mathcal{G}+\mathcal{D}$
$\mathcal{G}$	99.9±0.00	60.5±1.53	77.3±0.88
$\mathcal{D}$	59.2±0.29	99.9±0.00	74.6±1.07
$\mathcal{G}+\mathcal{D}$	99.9±0.00	99.9±0.01	99.9±0.00

Table 13. AUC of detection model when trained on all the different classes and architectures. Mean and standard deviation of three sessions with different seeds are reported.

trained solely on images generated by a single architecture demonstrate limited generalization ability [156], [157]. Furthermore, the results show that the semantic content of each dataset is a factor affecting the detection accuracy of generated images.

Qualitative artifact analysis: In this section, we visualize the artifacts that are calculated using the Fast Fourier Transform for different generated classes of the same architecture. Following [157], we first randomly select 1,000 generated images per concept and model architecture. Then using the denoising function proposed in [155], we transform each image X to its denoised version $f(X)$ and then compute the residual $R(X) = X - f(X)$. Next, we average the residuals and apply a 2D Fast Fourier Transform in order to obtain the magnitude and phase spectrograms. The magnitude spectrograms are a good indicator for the analysis of the artifacts introduced by a generative model.

As seen in Figure 30, differences exist between the generated images of each concept: Generated images from models pretrained on datasets with similar characteristics present more similarities. For instance, FFHQ and AFHQ, which are used for Humans $\mathcal{H}^G$ and Animals $\mathcal{A}^G$, respectively, are high-quality and high-resolution datasets consisting of images with centered and aligned faces. This produces similar cloudy magnitude spectrograms for these two cases. Similar spectrograms also appear in the case of the Bedrooms $\mathcal{B}^D$ and Churches $\mathcal{C}^D$ generated from the models that were trained on the LSUN-Bedroom and LSUN-Church datasets, which contain images of significantly lower resolution and quality compared to FFHQ and AFHQ. Nevertheless, all images are preprocessed in the same way in our experiments; hence, we expect such differences to be mitigated during our evaluations. This observation reinforces the main assumption of this work, which is that images produced by generative architectures are class-dependent.

Model compression for synthetic image detection in real-world scenarios: While the effectiveness of the proposed methodology is significant, it is also crucial to take into account its efficiency, especially in scenarios necessitating deployment at the edge. To this end, our goal was to compress the initial detector into a lightweight architecture, while preserving its high performance. Pruning [158] and Knowledge Distillation (KD) [159] constitute two widely applied approaches for compressing a neural network. The objective of KD is to transfer knowledge from a powerful teacher network to a smaller and faster one, in order to extend its performance capabilities, while pruning aims to discard the redundant parameters of a neural network in order to reduce its storage requirements or/and inference time. Taking advantage of the capabilities of these to model compression techniques, we developed a method, namely InDistill[6], that combines knowledge distillation and channel pruning in a unified framework for the transfer of the critical information flow paths from a heavyweight teacher to a lightweight student. Such information is typically collapsed in previous methods due to an encoding stage prior to distillation. By contrast, InDistill leverages a pruning operation applied to the teacher's intermediate layers reducing their width to the corresponding student layers' width. In that way, we force architectural alignment enabling the intermediate layers to be directly distilled without the need of an encoding stage. Additionally, a curriculum learning-based training scheme is adopted considering the distillation difficulty of each layer and the critical learning periods in which the information flow paths are created.

model	#parameters	*accuracy*	*AUC*
Resnet-50 (teacher)	23.9 m	99.5	99.9
Resnet-18 (student)	11.4 m	99.2	99.9

Table 14. AUC of teacher and student detection models on $\mathcal{G}$ set.

InDistill was applied on the detector model for transitioning from the ResNet-50 teacher model consisting of 23.9 million parameters to a ResNet-18 student model with 11.4 million parameters. The target was to enable the student model to achieve competitive performance compared to the teacher model that demonstrates 99.5% accuracy and 99.9% AUC. Despite the student model possessing 52.3% fewer trainable parameters than the teacher, InDistill led to nearly identical performance levels, achieving an accuracy of 99.2% and an AUC score of 99.9%, as presented in Table 14. The excellent performance of the compressed model enabled us to develop an Android application for detecting GAN-generated images at the edge. To this end, we utilized the Pytorch Lite Interpreter which allows for deploying pre-trained models for inference on Android devices. Figure 31 presents the User Interface of the developed demo application[7]. Users have the option to either use an image from their local storage or fetch a synthetic image directly from `https://thispersondoesnotexist.com/`. Upon selection, the chosen image is processed by the lightweight model and the result accompanied by a confidence score is displayed to the user.

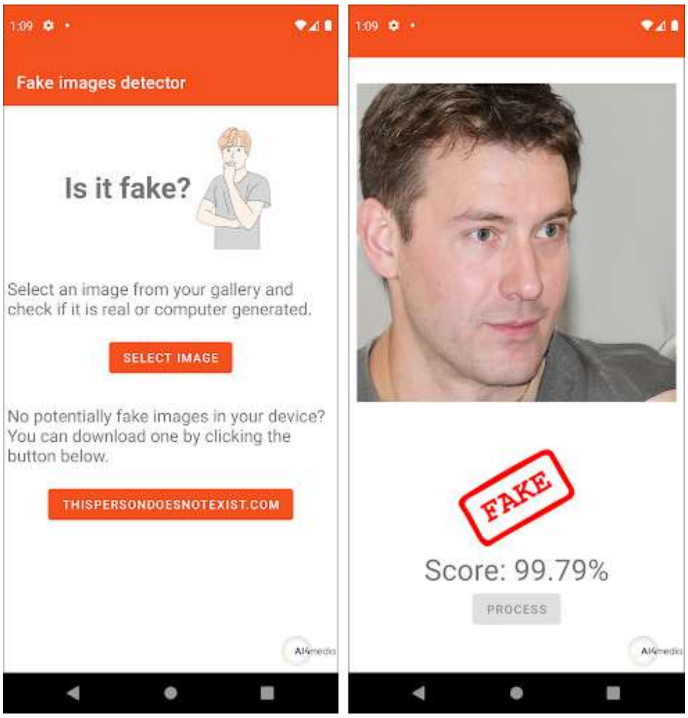
4.1.5.2 Conclusions

The main contributions of this work are:

1) We show the lack of generalization of state-of-the-art detectors in the cross-concept scenario.
2) We propose a sampling strategy for training data that considers image quality scoring.

[6] The InDistill method was developed in the context of T3.5 and will be presented in detail in the upcoming WP3 deliverable.

[7] Note that this is a prototype app developed by CERTH to help with the integration of the model. The fully fledged UC1 app that is discussed in D8.4, Sec. 2.5.3, will be delivered in the coming months.



Select an image from your gallery and check if it is real or computer generated.

No potentially fake images in your device? You can download one by clicking the button below.

Figure 31. The User Interface of the developed Android application for detecting GAN-generated images.

3) We demonstrate improved performance using the proposed approach in the cross-concept settings of three concept classes for two generative architectures.
4) We apply a knowledge distillation and curriculum learning-based method to enable synthetic image detection on off-the-shelf smartphones.
6) We provide our codes publicly available to facilitate future research on the field: https://github.com/dogoulis/qc-sgid and https://github.com/gsarridis/InDistill.

4.1.5.3 Relevant publications

- P. Dogoulis, G. Kordopatis-Zilos, I. Kompatsiaris and S. Papadopoulos, Improving Synthetically Generated Image Detection in Cross-Concept Settings, Proc. of the 2nd ACM Intern. Workshop on Multimedia AI against Disinformation, pp. 28-35, June 2023. [160].
  Zenodo record: https://zenodo.org/record/7984285.
- I. Sarridis, C. Koutlis, S. Papadopoulos and I. Kompatsiaris, InDistill: Transferring Knowledge From Pruned Intermediate Layers, arXiv preprint arXiv:2205.10003, 2022. [161].

4.1.5.4 Relevant software, datasets and other resources

- The Pytorch implementations can be found in
  https://github.com/dogoulis/qc-sgid and
  https://github.com/gsarridis/InDistill.

4.1.5.5 Relevance to AI4media use cases and media industry applications The presented method supports UC1 (AI for Social Media and Against Disinformation), specifically, the

Feature 1A (Detection/Verification of Synthetic Media). According to 1A, journalists should use AI tools to evaluate the authenticity of multimedia content and aid in the verification process. The proposed method has been demonstrated to be able to detect images that have been generated using StyleGAN2 and Latent Diffusion models, two of the most popular synthetic image models to date. Additionally, this capability is possible to port to smartphones, which extends the applicability of these tools to more user scenarios, e.g. in cases where the images to be checked are confidential or contain sensitive information and cannot be shared with third party services.

4.2 Audio-based DeepFake detection

This section focuses on current AI4Media activities related to the detection of audio DeepFakes. The methods presented hereafter thus apply to both audio files and audio streams of video files.

In Section 4.2.1, we describe how to generate a coherent set of audio DeepFakes to be used for training and testing of current and future DeepFake detection algorithms. In Section 4.2.2, we present a first approach to distinguish real speech recorded with a device from synthetic speech generated by text-to-speech algorithms. In Section 4.2.3 we report a novel method to distinguish which device was used to record an audio file in the presence of non-negligible background noise. Lastly, in Section 4.2.3, we introduce a method able to determine whether an audio track under examination is unaltered or if it has been manipulated by splicing one or multiple segments recorded by different device models, and to localize the splicing point.

4.2.1 Requirements for Synthetic speech generation

Contributing partners: CERTH, FHG-IDMT

In the course of the research project, a collaborative effort was undertaken by CERTH and FhG-IDMT to generate and collect instances of synthetic and natural speech. This initiative was driven by the need to address the prevailing challenges in synthetic speech detection, with the objective of aligning the data used for training and validation of detection methods with the requirements of the intended end-users [162].

During the data collection phase, it was observed that the usability of publicly available datasets was significantly compromised due to a variety of issues. These issues encompassed undisclosed synthesis algorithms, unpaired synthetic and real data, and an overly specific distribution of voice characteristics, often skewed towards a single speaker. Furthermore, it was noted that these datasets did not adequately cover contemporary synthesis methods [162].

A thorough analysis of public datasets was conducted, as shown in Table 15, revealing several challenges. The ASVspoof dataset, while extensive, was limited by undisclosed synthesis algorithms. The Half-Truth Dataset, despite its unique focus on detecting manipulated real audio, had limited usability for systematic research. The FoR dataset, although balanced in terms of gender and class, had undisclosed synthesis algorithms due to its challenge-driven nature. The WaveFake dataset, while diverse in its synthesis methods, was skewed by a single female voice. The ADD dataset had distribution restrictions, and the TIMIT-TTS dataset, although multi-speaker, was limited by a single female voice representation. These challenges underlined the need for a more comprehensive dataset for synthetic speech detection [162].

In response to these challenges, a position paper [162] was developed outlining the requirements for a common, public dataset to facilitate research and development in synthetic speech detection. The aim of these requirements was to bridge the existing gap between synthesis and detection methodologies.

The prerequisites for the proposed dataset were comprehensive and diverse. The dataset was required to be phonetically diverse, balanced in terms of gender, and inclusive of a broad spectrum of dialects. In addition, the process of data collection was structured to comply with legal regulations pertaining to data protection, such as the European General Data Protection Regulation (GDPR), the forthcoming California Privacy Rights Act (CPRA) in the United States, and other similar future initiatives. This compliance was deemed essential to ensure the dataset's usability by researchers worldwide.

Moreover, the conditions for recording were ideally established in acoustically controlled environments utilizing high-quality recording equipment. The dataset was also intended to comprise paired bona-fide and synthetic data. The synthetic data was expected to be created using long periods, of duration higher than a few seconds, to avoid wrong prosodies and to minimize their lack of expressiveness. The synthetic data was also expected to be paired with the real data it should be compared with, ensuring that all synthetic voices have a natural equivalent. This balance was expected to be maintained in the real data as well.

The research underlined the importance of the interpretability of the developed detection system, which was considered a key requirement in many court cases for the analysis to be admissible. The research proposed to focus on collecting requirements for new research datasets compliant with regulations and better representing real-case scenarios, on defining characteristics of future trustworthy detection methods. The importance of federated learning as a collaborative technique to quickly counter the constant flow of new synthesis methods released was also highlighted. Federated Learning (FL) enables the collaborative training of a neural network from siloed data centers and remote devices, while preserving data privacy. However, it also introduces challenges due to the heterogeneity of the data distribution among the network nodes. The research also emphasized the importance of Explainable AI (XAI) in the context of synthetic speech detection. XAI algorithms can provide insights and assist in the improvement of the existing AI solutions, particularly in understanding the distinct voice characteristics that distinguish synthetic speech samples from real audio samples.

4.2.1.1 Conclusions

Within this activity we collected several requirements of training and evaluation datasets suitable for research on synthetic speech detection.

This analysis was necessary, since the open datasets available at the time of writing still do not present the characteristics required for the production of a robust, interpretable detection system that can be used in the media sector industry.

The requirements and related challenges were reported in a public conference paper [162], which was presented to the research community at the main IEEE conference focused on forensics and security (WIFS), receiving large acceptance by the audience and therefore achieving the potential to steer future state-of-the-art research on the topic, beyond the activities carried out within the AI4Media project.

4.2.1.2 Relevant publications

The activities related to the analysis of the requirements of research datasets for synthetic speech detection led to the following conference paper:

- L. Cuccovillo, C. Papastergiopoulos, A. Vafeiadis, A. Yaroshchuk, P. Aichroth, K. Votis, and D. Tzovaras, D., “Open Challenges in Synthetic Speech Detection,” in IEEE International Workshop on Information Forensics and Security (WIFS), pp- 1–6, 2022 [162].

Dataset	Real Utterances	Fake Utterances	Notes
FoR	117,000	87,000	Gender balanced, class balanced and truncated versions of the original dataset are provided.
ASVSpoof	5,128	25,096	Different datasets are provided to tackle three major forms of spoofing attacks, namely replay, voice conversion and speech synthesis.
WaveFake	-	117,985	Fake utterances include a single female voice, resulting in data distribution bias.
ADD	5,319	45,367	Low usability for systematic research due to restrictions in distribution.
Half-Truth	26,554	26,554	Partially fake utterances are provided for the purpose of detecting manipulated real audio.
TIMIT-TTS	-	79120	Includes several multi-speaker synthesis methods while others are represented by a single female voice.

Table 15. Available datasets for the development of spoofing attacks detection models.

4.2.1.3 Relevant software, datasets and other resources

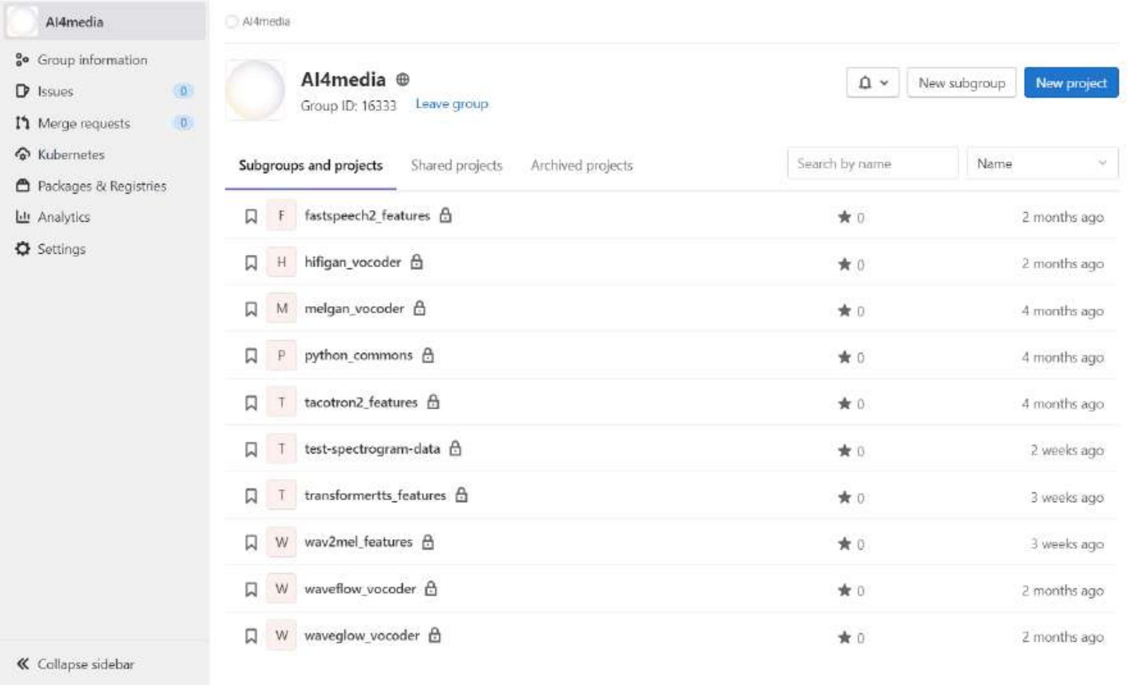
Software related to synthetic speech generation is collected collaboratively in the aforementioned shared GitLab repository, a screenshot of which is depicted in Figure 32.

The software is being used to create a open dataset for research on synthetic speech detection, according to the principles outlined in [162]. Whereas the analysis of the requirements and the selection of an initial set of synthesis algorithm has been carried out as part of AI4Media in the paper mentioned above, the dataset generation is being carried out as part of the vera.ai project, partially funded by the Horizon Europe program (grant agreement No. 101070093). After completing the generation the dataset will be made publicly available and the related characteristics will be reported in details in the upcoming D6.4 AI4Media deliverable, due in M48.

4.2.1.4 Relevance to AI4media use cases and media industry applications

This activity is a necessary prerequisite for the detection of synthetic speech, and thus is primarily related to UC1 (AI for Social Media Against Disinformation) and UC2 (Ai for News: The Smart News Assistant), that deal with disinformation detection. Furthermore, it relates to the "just in time content verification" requirement of UC3 (AI for Vision: Video Production and Content Automation), aiming at real-time detection of fabricated or manipulated content.

More specifically, our work helps identifying the requirement of training dataset required to deliver robust and unbiased synthetic audio detection. At the time of writing, such a dataset does not yet exist. Therefore, the information collected on the topic might be useful both to UC partners to verify the correctness of the training data involved in the detectors to come, and to the whole media industry as blueprint for contributing to the generation of such training and evaluation dataset, fostering the research activities required to leverage the potential of AI.



AI4media
Group ID: 16333 Leave group

F	fastspeech2_features	★ 0	2 months ago
H	hifigan_vocoder	★ 0	2 months ago
M	melgan_vocoder	★ 0	4 months ago
P	python_commons	★ 0	4 months ago
T	tacotron2_features	★ 0	4 months ago
T	test-spectrogram-data	★ 0	2 weeks ago
T	transformertts_features	★ 0	3 weeks ago
W	wav2mel_features	★ 0	3 weeks ago
W	waveflow_vocoder	★ 0	2 months ago
W	waveglow_vocoder	★ 0	2 months ago

Figure 32. T6.2: Common repositories for synthetic speech generation.

4.2.2 Audio DeepFake detector

Contributing partners: CERTH

The rise of synthetic speech presents novel challenges in the field of auditory security, necessitating rigorous countermeasures to ensure the integrity and authenticity of speaker verification systems. This study aims to contribute to this growing field by evaluating the effectiveness of the Conformer model [163], a hybrid architecture fusing Transformer and Convolutional Neural Network (CNN) elements, for the task of synthetic speech detection. Utilizing the Logical Access (LA) partition of the Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) 2019 dataset [164], we examine the impact of data augmentation and transfer learning techniques on the model's performance. Specifically, we introduce a comparative analysis of four training conditions to elucidate their influence on two key evaluation metrics: Equal Error Rate (EER) and t- Detection Cost Function (t-DCF). This analytical discourse serves to further our understanding of best practices in deploying advanced machine learning techniques to bolster the security framework of audio verification systems.

4.2.2.1 Experiments **Dataset:** In the sphere of audio security, the Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) 2019 challenge dataset [164] is an invaluable asset. This all-encompassing dataset, encompassing both genuine and artificially generated speech utterances, is the product of a broad array of Text-To-Speech (TTS) and Voice Conversion (VC) systems.

The ASVspoof dataset is structured into two fundamental sections: the 'defender' and 'attacker' subsets. The 'defender' subset includes genuine utterances alongside a substantial volume of spoofed utterances, the latter being a product of various TTS algorithms. Conversely, the 'attacker' subset furnishes unseen training data, facilitating the generation of synthetic speech that mirrors the characteristics of the target speakers.

The primary emphasis of this study is the Logical Access (LA) partition of the ASVspoof 2019 dataset. The LA subset has been specifically crafted for the discernment of genuine and spoofed utterances within communication networks, a pivotal component in strengthening the security fabric of modern digital communication systems.

For the purpose of this study, a balanced training dataset was created from the LA subset, ensuring equal representation of each class with a total of 11,000 samples, of which 7,000 were allocated for training, 2,000 for validation, and 2,000 for testing. This strategy was designed to mitigate any potential class bias during the training process, thus enhancing the reliability and robustness of the synthetic speech detection model.

Despite the ASVspoof 2019 dataset having certain constraints, notably the undisclosed nature of some synthesis algorithms, it continues to be an exceedingly valuable open-source resource for training and evaluation in the synthetic speech detection field. The dataset indirectly encapsulates possible interactions among feature extraction networks, voice conversion techniques, and vocoders in its data distribution, thereby enriching its utility for such applications.

Conformer Model: Transformers, first introduced by Vaswani et al. [163], revolutionized the field of natural language processing (NLP) by establishing a new paradigm for sequence learning tasks. Despite its initial application in machine translation, the success of the Transformer model soon permeated other NLP tasks, spurring the development of various Transformer-based models.

The Conformer model, proposed by Gulati et al. [165], innovatively merges the Transformer and Convolutional Neural Network (CNN) architectures. This hybrid approach was designed to address limitations of Transformer models in processing sequential data, particularly in the area of speech recognition.

The Conformer model utilizes a novel building block, the Conformer block, which seamlessly integrates the elements of both Transformer and CNN architectures. The block consists of four main components, depicted in Figure 33:

- *Feed-forward module:* A pair of feed-forward networks (FFN) positioned at the beginning and end of the block. Each FFN follows the traditional setup with two linear layers, separated by a Swish activation function.
- *Multi-head self-attention module:* Borrowed from the Transformer architecture, this module computes the attention scores for each pair of elements in the sequence, enabling the model to capture long-range dependencies.
- *Convolution module:* This is a 1D depthwise separable convolution module that performs local feature extraction. Depthwise separable convolutions, introduced by Chollet in MobileNet [166], are a variant of convolutions that reduce the model size and computation while maintaining performance.
- *Macaron-style feed-forward module:* This module, named after the MacaronNet [167], adds another feed-forward module in the middle of the block, further enhancing the model's capacity.

By combining local feature extraction capabilities of CNNs with global context understanding of Transformers, the Conformer model outperforms conventional Transformer models in various tasks, most notably in automatic speech recognition (ASR). This novel architecture marks an advancement in the lineage of Transformer-based models, underlining the potential of hybrid models for tackling complex sequence learning tasks.

Data Augmentation: The efficient and intelligent utilization of data plays a pivotal role in the development of robust and generalizable machine learning models. One such strategy is data augmentation, which seeks to enhance the diversity and size of the available dataset by applying a variety of transformations, hence mitigating the risk of overfitting and improving model

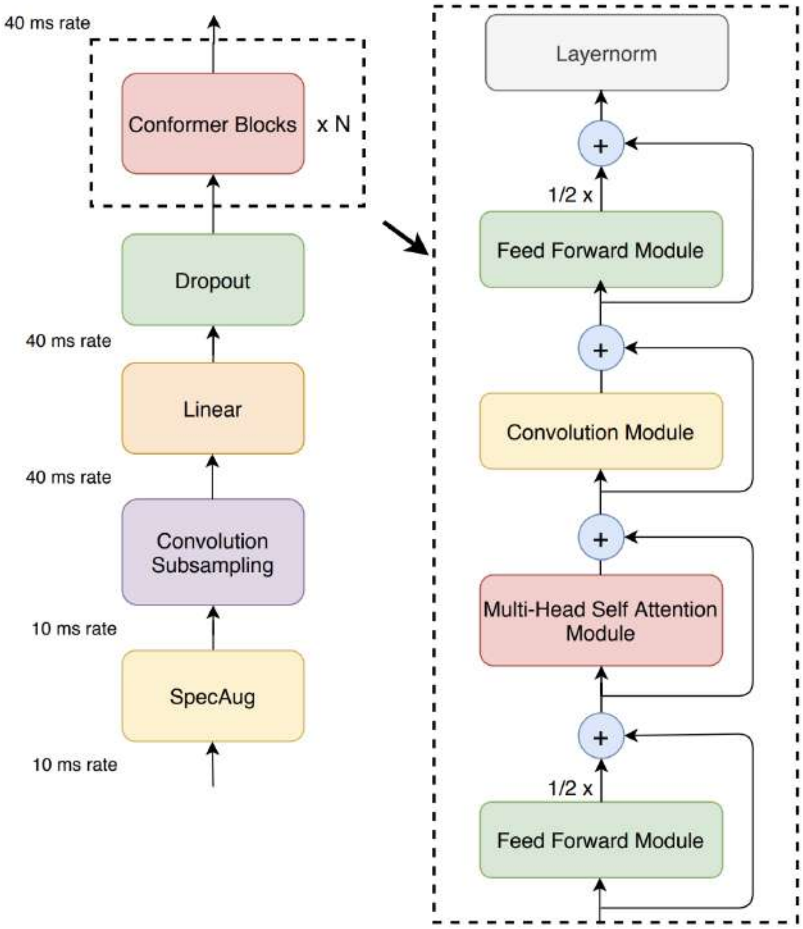


Figure 33. A representation of the Conformer block, showing its various components: the feed-forward module, the multi-head self-attention module, the convolution module, and the Macaron-style feed-forward module. The Conformer block combines the local feature extraction of CNNs with the global context understanding of Transformers. Figure adapted from [165].

generalizability [168].

In the realm of audio processing, there are several commonly applied data augmentation techniques, which include time masking, frequency masking, random trimming, and random noise injection. These techniques were applied to an audio sample and the resulting Mel-Spectrograms were visualized and compared to the original, non-augmented spectrogram that can be seen in Figure 34.

Time masking is a procedure that involves altering a random segment of the spectrogram along the time axis, effectively simulating the effect of temporarily muting the audio [169]. In contrast, frequency masking modifies a random section along the frequency dimension, mimicking the effect of a particular frequency band being obscured during the recording process.

Random trimming represents another augmentation approach, although its effect on the spectrogram may not be visually significant. It is a technique that removes a random section from the start or end of the audio sample, simulating the scenario where the audio recording starts late or ends early. It contributes to the model's ability to handle variable-length inputs and increases robustness against partial data [170].

The injection of random noise into the audio is a widely adopted data augmentation technique, serving to increase the model's robustness against background noise [171]. This technique involves adding a small quantity of random noise to the audio signal, mirroring the real-world scenarios where recordings are often contaminated with a variety of noises.

By employing these data augmentation techniques, the model's exposure to a diverse set of data is increased, thereby enhancing its ability to generalize and perform effectively on unseen data.

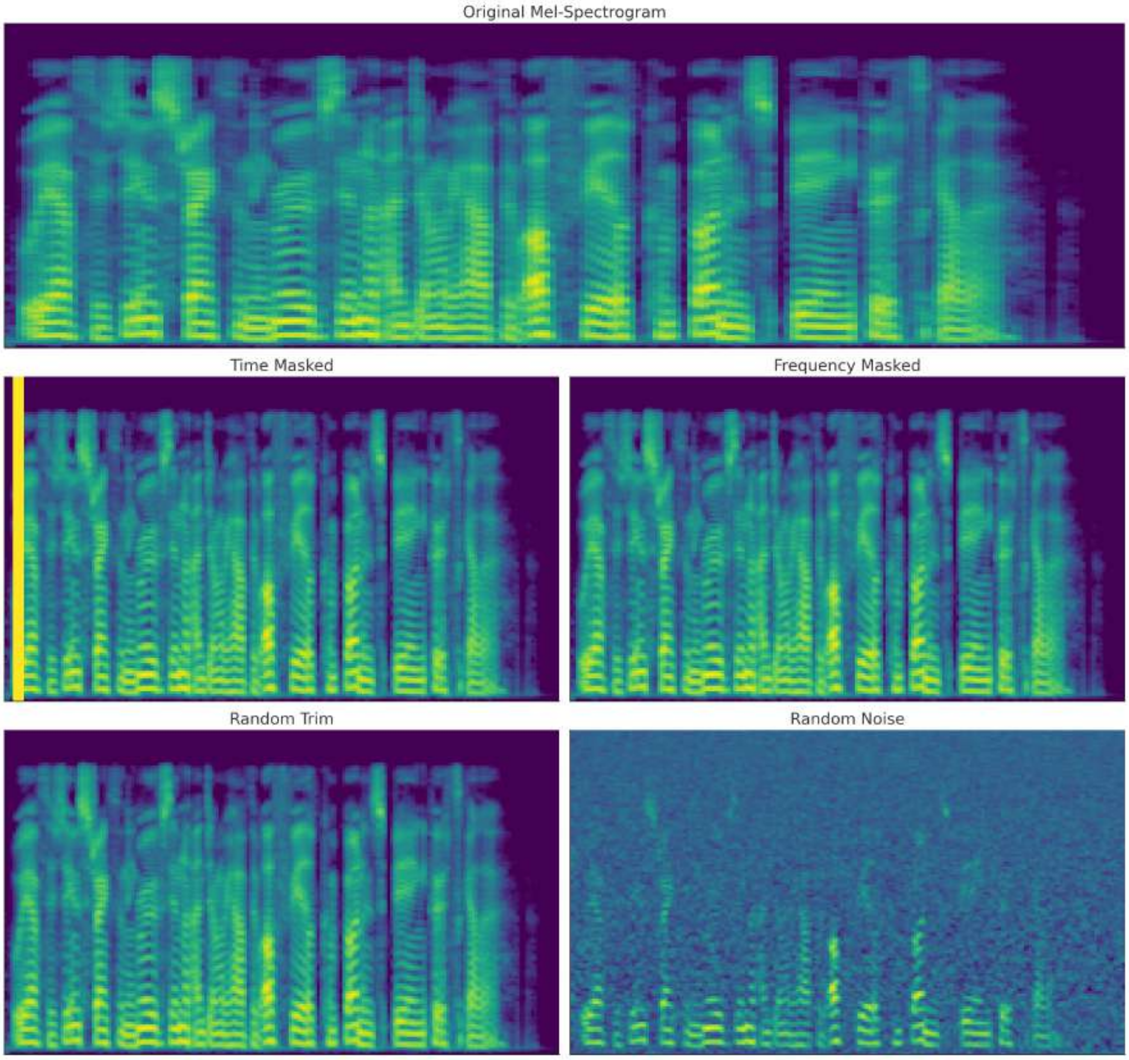


Figure 34. This figure illustrates the Mel-Spectrogram of an audio sample before and after various audio data augmentations. The original Mel-Spectrogram is shown at the top, and the spectrograms after Time Masking, Frequency Masking, Random Trimming, and Random Noise augmentations are displayed in a 2x2 grid below.

Training Process and Experimental results: Our study focused on the training of a Conformer model for the task of synthetic speech detection. The model was subjected to different training conditions to evaluate the impact of data augmentation and transfer learning on its performance. The training process was executed using the Adam optimizer with a learning rate of 1×10^{-4}, a batch size of 32, and for a maximum of 50 epochs. The loss function employed was binary cross-entropy.

We trained four different versions of the model:

1. A baseline model without data augmentation or transfer learning.
2. A model trained with data augmentation but without transfer learning.
3. A model trained with transfer learning but without data augmentation.
4. A model trained with both data augmentation and transfer learning.

Data augmentation was performed on the mel-spectrograms of the audio samples, which included time masking, frequency masking, random noise and audio trimming. Time masking was applied to a maximum of 20 time steps and frequency masking was applied to a maximum of 10 frequency bins. Each of these augmentations was applied with a probability of 0.1. These techniques are known to help the model generalize better to unseen data by providing more diverse training examples.

For the models trained with transfer learning, pre-trained weights from a similar Automatic Speech Recognition (ASR) task were used as the initial weights. This approach was chosen as an audio model based on spectrograms would have more relevant information for an audio task compared to using weights from a task like ImageNet, which is based on visual data. Transfer learning is known to help the model learn faster and achieve better performance by leveraging knowledge from a related task.

For the single-channel representation of the audio samples, Mel spectrograms were used. The Mel spectrogram is a visual representation of the normalized, square magnitude (power spectrum) of the Short-Time Fourier Transform (STFT) coefficients produced via the computation of Fourier Transform for successive frames in an audio signal[172]. However, unlike the STFT spectrogram, the frequency axis in a Mel spectrogram is scaled to the Mel scale, an approximation of the nonlinear scaling of the frequencies as perceived by humans. The Mel spectrograms were computed with a frequency axis of 128, FFT points of 2048, and a time axis of 256.

The performance of the models was evaluated using two metrics: Equal Error Rate (EER) and t-DCF. The Equal Error Rate (EER) [173] is a commonly used metric in binary classification tasks, especially in the field of biometrics and speaker verification. It represents the point at which both false acceptance rate (FAR) and false rejection rate (FRR) are equal. In the context of synthetic speech detection, a false acceptance occurs when a synthetic (fake) speech is incorrectly classified as genuine, while a false rejection occurs when genuine speech is incorrectly classified as synthetic. A lower EER indicates better performance of the model, with 0 being the ideal EER, indicating no false acceptances or rejections.

The t-DCF (t- Detection Cost Function) [174] is a more recent evaluation metric specifically designed for the ASVspoof challenge to measure the performance of countermeasures against spoofed speech attacks. The t-DCF takes into account both the performance of the countermeasure (CM) system and that of an automatic speaker verification (ASV) system. It provides a principled way of comparing different spoofing countermeasures under a range of different operating conditions, reflecting the trade-off between false alarms and missed detections. A lower t-DCF indicates better performance, with 0 being the ideal score.

The results are summarized in Table 16.

These results suggest that both data augmentation and transfer learning can improve the performance of the Conformer model on the task of fake speech detection. However, the best performance was achieved when both techniques were used together.

Augmentation	Transfer Learning	EER	t-DCF
No	No	10.8%	0.20
Yes	No	9.7%	0.19
No	Yes	9.3%	0.18
Yes	Yes	8.01%	0.17

Table 16. Performance of the Conformer model for fake speech detection under different training conditions.

4.2.2.2 Conclusions This study presented a comprehensive approach to the detection of synthetic speech, leveraging the synergistic combination of state-of-the-art feature extraction, data augmentation, and deep learning techniques. The integration of advanced audio data augmentation methods enriched the model's exposure to various audio characteristics, strengthening its generalization capabilities. The utilization of the robust Conformer model, with its unique fusion of Transformer and Convolutional Neural Network architectures, proved to be highly effective for this complex classification task. Importantly, the incorporation of transfer learning strategies amplified the model's performance, substantiating the value of pre-existing knowledge in augmenting predictive accuracy. Experimental results on the ASVSpoof 2019 dataset were promising, with the model achieving superior performance metrics in terms of both EER and t-DCF, particularly when data augmentation and transfer learning were applied in tandem.

4.2.2.3 Relevant publications

- No relevant publications published yet.

4.2.2.4 Relevant software, datasets and other resources

- No additional resources published yet.

4.2.2.5 Relevance to AI4media use cases and media industry applications This research is focused on the development of a robust deep-learning-based framework for the detection of synthetic audio, specifically engineered to be incorporated into UC1's existing synthetic audio detection application. Targeting a user base of media professionals, including those in journalism, film, and gaming, the enhanced application aspires to be more than a mere binary classifier. It aims to deliver an end-to-end solution that not only labels audio samples as 'real' or 'synthetic' but also enhances the decision-making process through transparent, explainable AI methodologies.

4.2.3 Microphone classification in noisy environments

Contributing partners: `FHG-IDMT`

Microphone classification is a classic problem related to the authenticity analysis of audio recordings. Given an unlabeled audio signal $x(t)$ and a set $\mathcal{X} = \{x_i\}$ of *known* recording devices, the goal of microphone classification is to determine which device x_i was used to acquire the audio signal under investigation, given a device fingerprint extracted from the device. In terms of machine learning, the operation consists a closed-set classification task, in which a pre-trained model classifier predicts a label x_i, given a feature vector extracted from the input signal $x(t)$, as shown in Figure 35.

The problem has been thoroughly addressed in the literature, e.g. in [175]–[179], by creating device fingerprints which model the microphone frequency response of the device, denoted by $F_{\text{mic}}(f)$ in

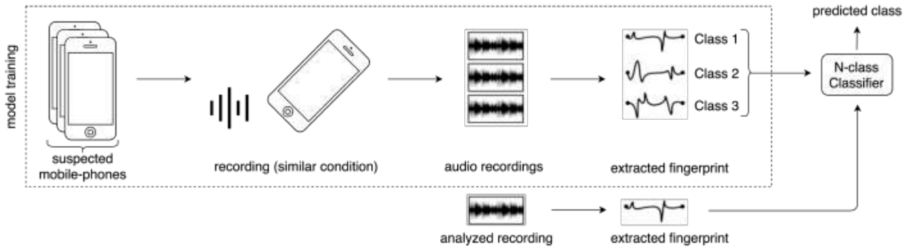


Figure 35. High level schema for closed-set microphone classification.

eq. (3)

$$x(t) = \int F_{\text{mic}}(f) \cdot [S(f) + N_{\text{env}}(f)] \, df. \tag{3}$$

In the equation above, $S(f)$ denotes the input (speech) signal in the frequency domain, $N_{\text{env}}(f)$ denotes any environmental additive noise in the frequency domain, and $x(t)$ the resulting audio signal under analysis. State-of-the-art methods work remarkably well in nearly *noiseless* conditions – i.e., by assuming $N_{\text{env}}(f) = 0$ in eq. (3) – but suffer from a great performance decrease whenever applied to *noisy* conditions, when the assumption does not hold; e.g., the accuracy of [176], [177] drops dramatically from about 99% to about 36% if the Signal to Noise Ratio (SNR) drops from $+\infty$ to 20dB.

In our method, we addressed the issue by introducing a denoising block, highlighted in green in Figure 36, which transforms the log-magnitude spectrogram of $x(t)$ by removing the influence of the additive noise $N_{\text{env}}(f)$. The rationale behind this choice was to develop a universal approach, to be applied independently from the specific feature extraction mechanism or the selected classification model: the log-magnitude spectrogram is a basis foundation not only of [176], [177], i.e., the baseline that we included for the first experiments on the topic, but of the near totality of publications addressing microphone classification, such as the aforementioned [178].

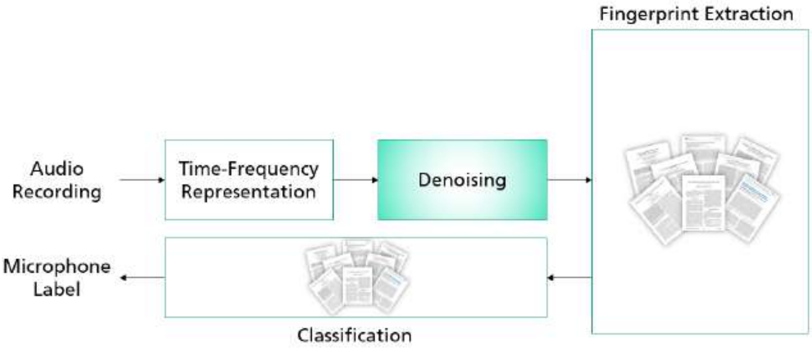


Figure 36. Schema of the proposed approach for microphone classification in noisy conditions.

4.2.3.1 Experiments

The first experiments, which were the focus of our first publication of microphone classification in noisy environments [180], aimed at determining the denoising algorithm best fitting the purpose among several candidates:

- *Total Variation [181]:* Based on Digital Signal Processing (DSP), performs image denoising based on the minimization of a constrained minimization problem.

- *Non-local means [182]:* Based on DSP, performs image denoising by replacing each pixel by a weighted average of all *similar* pixels in the rest of the image.
- *Bilateral filtering [183]:* Based on DSP, performs image denoising by replacing each pixel by a weighted average of all *similar, nearby* pixels.
- *Wavelet BayesShrink [184]:* Based on DSP, leverages wavelet decomposition to filter-out high frequency components associated to the noise.
- *DnCNN [185]:* AI-based approach for image denoising aiming at solving the problem by correctly predicting not the original clean image u, but rather the residual noise n associated to it.
- *Audio Denoising Autoencoder (Audio DAE) [186]:* AI-based approach for audio denoising in the time domain, aiming at reconstructing the clean audio signal $u(t)$ from its noise corrupted version $v(t) = u(t) + n(t)$

The modified pipeline including denoising was then tested on the MOBIPHONE dataset [187]. This dataset contains audio recordings from 21 mobile-phones produced by 7 manufacturers, and was specifically devised for evaluating microphone classification. It contains utterances lasting 30 seconds each, spoken by 12 female and 12 male speakers and captured in a silent laboratory environment using a sampling rate of 16 kHz and the GSM-AMR encoding. The final recordings were then distributed as uncompressed WAV files with PCM encoding. Speakers, encoding and devices are thus completely uncorrelated with the ones present in the LibriSpeech corpus.

To obtain our final *noiseless* dataset for training and testing, we split the MOBIPHONE recordings in non-overlapping segments of 4.112 seconds, obtaining 168 *noiseless* examples per class. The same segments were then corrupted using additive white Gaussian noise with 25dB SNR, obtaining 168 *noisy* examples per class. For brevity, we will hereon use $MOBI_{+\infty}$ to denote the *noiseless* examples, and $MOBI_{25}$ to denote the *noisy* ones – i.e., the index denotes the SNR.

The dataset preparation for the denoiser comparisons was performed as follows:

1. We corrupted all audio files in the MOBIPHONE dataset with additive white Gaussian noise, using a SNR of 25 dB.
2. We extracted log-power spectrograms from the clean MOBIPHONE dataset, obtaining a reference set $X_{ideal}(f)$
3. We extracted denoised log-power spectrograms from the noisy audio files, obtaining a benchmark set $\widehat{X}_{25}(f)$
4. We split both reference set and benchmark set into training and testing portions, obtaining the four distinct sets $X_{ideal}^{train}(f)$, $\widehat{X}_{25}^{train}(f)$,$X_{ideal}^{test}(f)$, $\widehat{X}_{25}^{test}(f)$

After obtaining the four aforementioned sets, we compared the outcome of the denoising using 3 different metrics:

1. PSNR: Average Peak Signal-to-Noise Ratio (PSNR) between corresponding pairs of $X_{ideal}^{test}(f)$ and $\widehat{X}_{25}^{test}(f)$
2. SSIM: Average Structural Similarity Index Measure (SSIM) between corresponding pairs of $X_{ideal}^{test}(f)$ and $\widehat{X}_{25}^{test}(f)$
3. MCA: Microphone Classification Accuracy (MCA) of the baseline trained on $X_{ideal}^{train}(f)$ and tested on $\widehat{X}_{25}^{test}(f)$

The first two metrics relate directly to the visual quality of the denoising: The PSNR quantifies the closeness in terms of pixel energy between the original spectrogram and the denoised one, while the SSIM their similarity in terms of luminance, contrast and structure. The MCA metric is meant to capture to which extent the classification is possible after the denoising operation: aggressive denoisers may remove too much content from the log-power spectrogram, while ineffective ones may remove too little disturbance for being of any help.

The outcome of this evaluation is reported in Table 17, according to which we selected the DnCNN [185] as most appropriate denoising algorithm. Its PSNR is superior to the Audio DAE by more than 6dB, and the SSIM is beyond the one achieved by the DSP-based denoisers. The most promising score, however, was the MCA itself: without retraining the classifier, which would probably improve the performances but could be a costly operation, we achieved an accuracy of about 69%, which is significantly beyond the accuracy of all other alternatives.

Given a target original image X and an input image $X_{\text{noise}} = X + N$ corrupted by Gaussian noise N, the DnCNN network is able to compute an estimate $\hat{N}$ of the input noise, and thus an estimate $\hat{X}$ of the original image. In our proposal, the input image X and the Gaussian noise N coincide respectively with the input spectrogram $S(f)$ and environmental noise $N_{\text{env}}(f)$ from eq. (3), as depicted in Figure 37.

Denoiser	Performances (#)		
	PSNR	SSIM	MCA
Total Variation [181]	20.56	0.81	34.40
Non-Local Means [182]	20.59	0.83	36.69
Bilater Filtering [183]	20.33	0.81	33.67
Wavelet BayesShrink [184]	20.62	0.82	39.06
DnCNN Architecture [185]	**27.80**	**0.86**	**69.09**
Audio DAE Architecture [186]	21.02	0.76	26.47

Table 17. Denoising Baseline Benchmarks

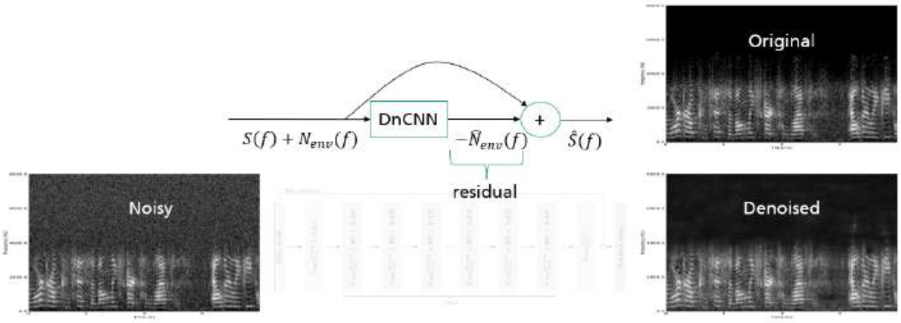


Figure 37. DnCNN application to audio spectrograms.

In a second study, which eventually led to a second publication on the topic of microphone classification in noisy conditions [188], we focused on determining the applicability of the method on various algorithms, using different basic features for microphone classification:

- f_1: A full blind estimate of the microphone frequency response [189].
- f_2: A hand-crafted multidimensional feature for microphone classification [176]
- f_3: A descriptor of average energy differences between frequency bands [178]

Table 18 shows the different accuracies obtained downstream the identification process with and without the denoiser preprocessing, as the SNR of the input audio recording varies. A different Support Vector Machine (SVM) classifier was trained for each feature, while the *Clean* (noiseless, i.e., $MOBI_{+\infty}$) training set was kept constant across SNRs. Analyzing the results, we can see how the noise removal process improves the performance of all three features.

Feature	DnCNN	Test Dataset				
		20dB	25dB	30dB	35dB	Original
f_1	✓	57.5	69.1	83.7	84.2	95.1
	✗	36.0	41.8	50.4	60.6	99.2
f_2	✓	47.2	49.3	53.9	57.5	96.5
	✗	26.7	33.3	39.7	50.8	99.9
f_3	✓	64.8	74.4	78.0	85.6	99.3
	✗	42.6	52.8	58.2	77.7	99.2

Table 18. Microphone Classification in Noisy Conditions: Identification Accuracy [%] between the three features for different test SNRs, with and without the DnCNN denoiser contribution, using the Clean *training for the SVM classifier.*

Up to this point, we tested the system with different SNRs while using the sole clean files for training. To overcome this limitation, we trained three different SVMs per feature, on different variations of the MOBIPHONE datasets:

- *Clean*: classical training, with features obtained from noiseless recordings;
- *Mixed*: multi-scene data augmentation training, with features obtained from recordings having a variable SNR between 20dB and 35dB and the noiseless one, i.e., the original;
- *Denoised*: multi-scene data augmentation training, with features obtained from denoised recordings that had variable SNR between 20 and 35dB and the noiseless one.

The corresponding results are shown in Table 19. We can observe that M(*ixed*) training turns out to be somehow effective for f_2 and f_3 concerning all five SNRs, if compared to case C(*lean*). The feature f_1, behaves differently, to the point that from 30dB SNR we notice a convenience in training on noiseless recordings. Most importantly, training on the D(*enoised*) set seems to be consistently the best option, leading to a strong boost in the identification performance. Indeed, the results obtained while training on the augmented D(*enoised*) set approach the ones obtained in the noiseless ideal case.

4.2.3.2 Conclusions

Within this activity, we proposed a robust method capable of identifying a recording device in noisy conditions, i.e., we advanced the existing state of the art to cope with real-life recordings which otherwise could not have been processed.

The limitation of the proposed method lies in the accuracy itself (which might be included) and in the lack of a large scale evaluation: Even though the results suggest that a classification based on the model of the device might be feasible, the generality of this statement has yet to be proved.

Furthermore, the paper did not address open-set classification, in which the model of the device used in the recording at hand might fall outside of the training set: Using the method as-is, without considering this limitation, might lead to a strong bias in the conclusions drawn by the final user.

4.2.3.3 Relevant publications

The activities related to microphone classification in noisy conditions led to two conference papers on the topic:

- L. Cuccovillo, A. Giganti, P. Bestagini, P. Aichroth, and S. Tubaro, "Spectral denoising for microphone classification," in ACM International Workshop on Multimedia AI against Disinformation (MAD), pp. 10–17, 2022 [180].
- A. Giganti, L. Cuccovillo, P. Bestagini, P. Aichroth, and S. Tubaro, "Speaker-independent microphone identification in noisy conditions," in European Signal Processing Conference (EUSIPCO), pp. 1047–1051, 2022 [188].

4.2.3.4 Relevant software, datasets and other resources

- No additional resources published yet.

4.2.3.5 Relevance to AI4media use cases and media industry applications

This activity contributes to the detection of manipulated speech, and thus is primarily related to UC1 (AI for Social Media Against Disinformation) and UC2 (AI for News: The Smart News Assistant), that deal with disinformation detection. Furthermore, it relates to the "just in time content verification" requirement of UC3 (AI for Vision: Video Production and Content Automation), aiming at real-time detection of fabricated or manipulated content.

More specifically, our work might help identifying the model of the recording device used for creating a specific audio signal. The resulting information could be used to verify the correctness of the statements provided upon the content, thus supporting the detection of decontextualization of pristine material (e.g., a recording with a video camera reported as being captured by a mobile phone).

Furthermore, the same information could be used in court-cases for verification of evidence allegations, or could be leveraged by the media industry to annotate large media archives (e.g., bounding the microphone model to the expected audio quality).

Feature	DnCNN	Training Dataset	Test Dataset				
			20dB	25dB	30dB	35dB	Original
f_1	✗	M	67.3	68.9	67.3	68.0	83.4
	✓	C	57.5	69.1	83.7	84.2	95.1
	✓	D	83.6	92.6	96.7	95.6	95.3
f_2	✗	M	86.0	94.2	96.3	96.2	94.3
	✓	C	47.2	49.3	53.9	57.5	96.5
	✓	D	92.7	96.2	98.1	98.5	93.5
f_3	✗	M	75.3	85.6	91.9	96.5	98.5
	✓	C	64.8	74.4	78.0	85.6	99.3
	✓	D	93.6	97.8	98.4	98.0	96.2

Table 19. Microphone Classification in Noisy Conditions: Identification Accuracy [%] of our system, for all the three features and tested using different SNRs and SVM training sets, i.e., M(ixed), C(lean), D(enoised).

4.2.4 Microphone traces analysis for audio splicing detection

Contributing partner: FHG-IDMT

In recent years, the multimedia forensic community has dedicated significant efforts to developing solutions for evaluating the integrity and authenticity of multimedia objects. Most of the focus at present addresses manipulations performed using advanced deep learning techniques, among which the creation of fully synthetic audio files ("deepfakes"). However, it is crucial to recognize that alongside these complex forgeries, there are simple yet highly effective manipulation techniques that do not require the use of state-of-the-art editing tools. These techniques remain a significant threat.

One such example is audio splicing for speech signals, where multiple speech segments from various recordings of an individual are concatenated and combined to create a new fake speech. This simply yet effective method has the power to completely alter the meaning of an existing speech by merely adding a few words. With this research we aimed to address the overlooked problem of detecting and locating audio splicing across different models of acquisition devices. Our objective was to determine whether an audio track under examination is unaltered or if it has been manipulated by splicing one or multiple segments obtained from various device models. Additionally, if a recording is identified as spliced, we strived to identify the specific point in the temporal dimension where the modification has been introduced.

Our research, published in a corresponding journal paper [190], relied on the premise that each distinct audio segment, when combined with others, possesses unique characteristics that set it apart. By analyzing whether these audio recording characteristics undergo changes over time, we can effectively detect and pinpoint the presence of splicing. We therefore proposed to employ a Convolutional Neural Network (CNN)-based technique for extracting audio embeddings specific to various models of acquisition devices, considering different time windows. Subsequently, we performed a clustering process to identify potential inconsistencies among the embeddings, which could indicate the presence of multiple devices being used. If splicing is detected, we proceed to determine the exact point of splicing by measuring the distances between embeddings and further enhance the accuracy of the detection results. The high level schema is depicted in Figure 38.

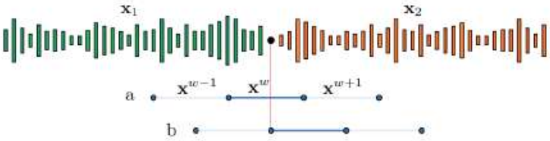


Figure 38. High level schema for splicing detection. In case (a), the window x^w contains samples from multiple models: embeddings of x^{w-1} and x^{w+1} will be far from each other; In case (b), the splicing point falls exactly between the windows x^{w-1} and x^w.

Let us denote with x an input recording, and with x_q the same input recording from which the first q samples were removed. Futhermore, let us denote with $X_q^{\text{emb}} = \{x_q^{\text{emb}}[w]\}, w \in [1, W]$ a set of embeddings obtained for all W non overlapping analysis frames of each x_q. Finally, let us denote with y_q a detection function obtained by computing the distance of consecutive embeddings, i.e.:

$$y_q[w] = \mathrm{d}\left(x_q^{\text{emb}}[w], x_q^{\text{emb}}[w-1]\right), w \in [1, W-1], \tag{4}$$

with $\mathrm{d}(\cdot, \cdot)$ a suitable distance for the embeddings. The splicing point can be uniquely determined by evaluating

$$\hat{w} = \arg\max_{w \in W} \left(\max_{q \in Q} y_q[w] \right), \tag{5}$$

as long as both the embeddings and the distance function are well defined.

To extract the embeddings, we propose to rely on a the next-to-the-last layer of a CNN network for microphone classification [191], which is trained using Binary Cross Entropy (BCE) on one-hot encoded labels of the microphone devices. A suitable metric for such embeddings is the cosine distance, i.e.:

$$y_q[w] = 1 - \frac{x_q^{\text{emb}}[w] \cdot x_q^{\text{emb}}[w-1]}{\|x_q^{\text{emb}}[w]\| \cdot \|x_q^{\text{emb}}[w-1]\|}. \tag{6}$$

4.2.4.1 Experiments

Several experiments were conducted to measure the quality of the proposed approach, and can be read in full details in [190]. Within this pages, we would like to focus on the comparison in terms of splicing detection of our methods against two baselines that we implemented specifically as comparison.

As first baseline, we considered [192], i.e., a method for splicing detection based on *reverberation* clues. As second baseline, we considered the [177], i.e., a open-set microphone classification method with an explicit discrimination function comparable to the one we proposed, but based on a blind estimate of the transmission channel.

Figure 39 shows the Receiver Operating Characteristic curves (ROC curves) obtained for the two baselines (Baseline 1 and Baseline 2) and our method (Proposed) tested on a clean dataset (i.e., tracks not affected by noise or compression), produced by thresholding the maximum value of the detection function $y_q[w]$, i.e. the output of the most prominent splicing point. In the detection test, we consider as true positives all the spliced tracks that were correctly detected as such, and as true negatives all the pristine tracks that were correctly detected as pristine.

These results show that the proposed method achieves the same Area Under the Curve (AUC) of the best baseline. However, this is not the only metric of interest. To better capture the behaviour of the method, Table 20 presents the performance in terms of both AUC and balanced accuracy when the best working point is selected. These results show that the proposed preliminary method outperforms both baselines, reaching a balanced accuracy of 96%, whith a maximum localization error of about 0.012 seconds.

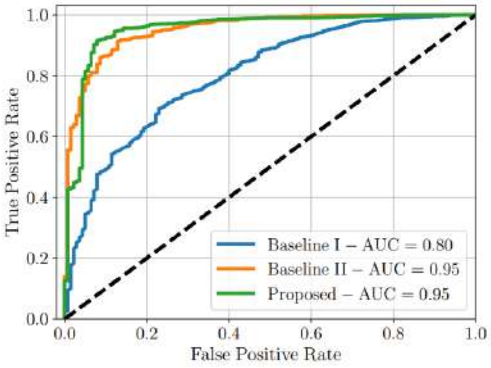


Figure 39. Audio splicing detection performance

Method	AUC	Accuracy
Baseline 1 [192]	0.80	0.77
Baseline 2 [177]	0.95	0.86
Proposed Method	0.95	0.96

Table 20. Audio splicing detection performance

4.2.4.2 Conclusions Within this activity, we proposed a novel method capable of detecting splicing points based on the traces left by the microphone devices. This kind of tampering detection was never attempted before, and remarkably makes use of AI up to the estimation of the microphone embeddings. Therefore, the detection process is completely transparent to the user, and can be monitored improving trustworthiness and fault tolerance.

Beyond the initial application of tampetind detection, the proposed technique proved to be potentially applicable to synthetic speech detection: Embeddings trained for microphone analysis proved to be usable for the attribution of the synthesis algorithm applied to generate the test recordings. In the future we would like to investigate this matter further, and to verify if the embeddings might also be used for blind synthetic speech detection.

4.2.4.3 Relevant publications

The activities related to microphone traces analysis for audio splicing detection led to the following journal paper on the topic:

- D. U. Leonzio, L. Cuccovillo, P. Bestagini, M. Marcon, P. Aichroth, and S. Tubaro, "Audio splicing detection and localization based on acquisition device traces," IEEE Transactions on Information Forensics and Security (TIFS), vol. 18, pp. 4157–4172, 2023 [190].

4.2.4.4 Relevant software, datasets and other resources

- No additional resources published yet.

4.2.4.5 Relevance to AI4media use cases and media industry applications

This activity contributes to the detection of manipulated speech, and thus is primarily related to UC1 (AI for Social Media Against Disinformation) and UC2 (AI for News: The Smart News Assistant), that deal with disinformation detection. Furthermore, it relates to the "just in time content verification" requirement of UC3 (AI for Vision: Video Production and Content Automation), aiming at real-time detection of fabricated or manipulated content.

More specifically, our work might help identifying recordings which were manipulated by splicing together segments captured by different devices. Whenever this splicing occur within a sentence which supposedly should not contain any manipulation, the alarm might be used to flag the content as potentially questionable, and therefore to be double-checked.

The same information could be used in court-cases for tampering detection of questioned evidence, or could be leveraged by the media industry to annotate large media archives (e.g., by automatically segmenting interview, or by supporting the automatic generation of cuesheets for archive content produced using legacy editing software).

4.3 Content-based Detection of Bot-generated Tweets

Contributing partners: CEA

Recent advances in automatic text generation enable the generation of short coherent text that imitates the style of the human-elicited text on which the models have been trained [193]–[196]. Potential misuse of these models includes the fast spreading of disinformation. In the context of an increasing political polarization, this could be detrimental to democracy and carry on actions which may cause public troubles. Methods which automatically discriminate short texts which are generated by humans from those generated by bots should be investigated to prevent such actions. However, this task is very challenging when very little background information is available for an account. Early detection is therefore crucial in order to stop disinformation campaigns before they spread significantly on the network [197].

Previous research focused on the identification of bots based on the analysis and the identification of anomalous behavior [197], [198] or on the mining of profile information [199]. Closer to our work is the approach proposed by the authors of [200]. They collected original tweets from human accounts (accounts the content of which is written by a person) and their fake bot counterparts maintained by people on the Twitter platform (accounts the content of which is written by text generation algorithms). They analyzed the performance of several classification models according to the technology used for tweet generation. They show that RoBERTa [201], a pretrained language model based on the transformer architecture [163] obtains the best performance across all configurations. The authors retrieved 23 bot accounts and 17 human accounts through the platform API. Although individual tweets are different in both train and test datasets, accounts are not unique in either part. This prevents from evaluating the generalization capabilities of their approach. Moreover, the authors highlight that tweets generated by GPT-2 [194] are more difficult to detect than tweets generated by older methods (e.g. AWD-LSTM). These approaches assume a significant amount of background information or of automatically- and human-generated text is available for each account. They do not enable the early detection of bot accounts, while detection is most useful before accounts start spreading misinformation.

We generalize the approach introduced in [200] by proposing a method which detects bot-generated tweets after the occurrence of a single tweet. Equally important, no network or profile information is required. Put simply, our method aims to detect bots after only one tweet and has the potential to counter disinformation campaign in an effective manner. Our main contributions are the following:

- We create a new dataset for deep fake tweet detection which contains 47 political and public personalities Twitter accounts. We generate their fake counterparts using GPT-2.
- We investigate the use of the newly generated dataset to improve the performance of a RoBERTa classifier on the [200] dataset.
- We investigate the generalization capabilities of a classification algorithm across Twitter accounts by creating a new dataset based on TweepFake.

4.3.1 Experiments

4.3.1.1 Datasets

Experiments and reported results are based on the TweepFake dataset and on a newly collected dataset. TweepFake contains 25,572 tweets (half humans, half bots). Tweets are generated with several algorithms. Further information about the human accounts and their bot counterparts is available in the original paper [200].

The newly collected dataset contains 47 accounts and 725,227 original tweets. As we mentioned earlier, we keep half of them to finetune one GPT-2 model per account and the other half is split into train (290,081), dev (36,260) and test sets (36,282). We generate the same amount of fake tweets for each part of the corpus resulting in a balanced dataset.

4.3.1.2 Tested Models

Our classifiers are based on the RoBERTa model (small version) to which we add a classification head. The head is composed of a 2-layer feedforward neural network whose hidden layer has the same size as the input layer. We use tanh as activation function an a dropout rate of 0.1 on both input and hiddent layers. The classification model is trained to minimize a Binary Cross Entropy loss. We train for 5 epochs with a learning rate of $1e^{-5}$ for the transformer model and $1e^{-3}$ for the classification head and a linear warmup of a 1/2 epoch (10%). We use AdamW as optimizer and a mini-batch size of 16. We implement our models using the library *transformers*[8]. We keep the best performing model on the development set (tested at the end of each epoch) and apply it on

[8] https://github.com/huggingface/transformers

the test set at the end of training. To assess the robustness of our model to the random seed, we run 5 experiments per configuration and report mean scores along with their standard deviations.

Concerning generation, we fine-tune the large version of GPT-2 (774M parameters) for one epoch with a learning rate of $1e^{-5}$ and a mini-batch size of 16. We use a temperature of 0.7 during tweets generation. Our implementation is based on the library *aitextgen*[9].

Dataset	Human			Bot			Global
#	P	R	F1	P	R	F1	Acc.
TweepFake	0.922 ±0.004	0.902 ±0.008	0.912 ±0.004	0.904 ±0.007	0.924 ±0.004	0.914 ±0.003	0.913 ±0.004
1	0.705 ±0.018	0.854 ±0.035	0.771 ±0.010	0.816 ±0.027	0.641 ±0.043	0.716 ±0.021	0.747 ±0.011
2	0.667 ±0.020	0.788 ±0.054	0.721 ±0.016	0.743 ±0.028	0.603 ±0.059	0.663 ±0.026	0.695 ±0.010
3	0.945 ±0.018	0.809 ±0.028	0.871 ±0.010	0.834 ±0.017	0.952 ±0.018	0.889 ±0.004	0.880 ±0.007
4	0.932 ±0.009	0.822 ±0.010	0.873 ±0.003	0.841 ±0.006	0.940 ±0.009	0.888 ±0.003	0.881 ±0.003
5	0.582 ±0.033	0.837 ±0.036	0.685 ±0.017	0.705 ±0.034	0.392 ±0.096	0.498 ±0.082	0.615 ±0.036
6	0.950 ±0.018	0.928 ±0.024	0.939 ±0.009	0.930 ±0.020	0.951 ±0.019	0.940 ±0.008	0.939 ±0.008
7	0.640 ±0.021	0.764 ±0.041	0.696 ±0.008	0.708 ±0.016	0.568 ±0.059	0.628 ±0.030	0.666 ±0.011
8	0.598 ±0.018	0.897 ±0.044	0.716 ±0.011	0.799 ±0.041	0.393 ±0.068	0.522 ±0.053	0.645 ±0.019
9	0.961 ±0.013	0.850 ±0.019	0.902 ±0.010	0.866 ±0.014	0.965 ±0.013	0.913 ±0.008	0.908 ±0.009
10	0.947 ±0.018	0.848 ±0.024	0.894 ±0.007	0.863 ±0.017	0.951 ±0.018	0.905 ±0.004	0.900 ±0.000

Table 21. Results for cross-account experiments with RoBERTa on TweepFake. We split randomly the TweepFake dataset into train, dev and test sets along the human(s)-bot(s) pair axis following the 80/10/10 ratio. We repeat this step 10 times and we run 5 experiments with each dataset. We report mean Precision (P), Recall (R) and F1-score (F1) for each account type (Human vs. Bot) and the mean global accuracy (Acc.). We present also the standard deviation for each score. For comparison, we report a simple RoBERTa baseline on TweepFake original split.

4.3.1.3 Results

Generalization capabilities. Performance obtained for the different corpus splits is presented in Table 21 along with our baseline. Depending on the random split, the global accuracy is varying from 0.615 to 0.939 with several standard deviation going beyond four points, suggesting that in some cases, the model encounters difficulties to generalize across accounts. These results show that generalization is a key issue for detecting generated tweets at an early stage. Models and algorithms that are proposed in the literature must ensure that they are able to generalize beyond the accounts used in their datasets.

Improving GPT-2 tweet detection. To assess the quality of our generated dataset, we sample 10 random datasets, split across the human(s)-bot(s) pairs, and learn one[10] classification model (RoBERTa) per split. Results are presented in Table 22. The performance obtained on these datasets is higher than the one obtained on GPT-2 tweets from TweepFake, suggesting that our generated tweets are easier to discriminate. We select the best performing model on bot accounts (number 3 in our case) and keep model weights for the next experiments.

4.3.2 Conclusions

Our study highlights the need for building models that generalize better beyond the accounts and the technologies used in the training datasets. Our experiments on random splits of TweepFake show that performance can vary from 0.615 to 0.939 according to the split. By generating fake

[9] https://github.com/minimaxir/aitextgen

[10] We do not run 5 experiments per split due to low computing budget.

Split	Human			Bot			Global
#	P	R	F1	P	R	F1	Acc.
1	0.927	0.946	0.936	0.945	0.925	0.935	0.936
2	0.935	0.946	0.941	0.946	0.935	0.940	0.940
3	0.941	0.951	0.946	0.951	0.941	0.946	0.946
4	0.953	0.899	0.925	0.904	0.955	0.929	0.927
5	0.916	0.904	0.910	0.905	0.917	0.911	0.911
6	0.945	0.912	0.929	0.915	0.947	0.931	0.930
7	0.922	0.932	0.927	0.931	0.921	0.926	0.926
8	0.970	0.914	0.941	0.919	0.972	0.944	0.943
9	0.927	0.945	0.936	0.944	0.926	0.935	0.935
10	0.928	0.907	0.917	0.909	0.930	0.919	0.918

Table 22. Results for experiments on the generated dataset. We split randomly the TweepFake dataset into train, dev and test sets along the human(s)-bot(s) pair axis following the 80/10/10 ratio. We repeat this step 10 times and we run 1 experiment with each dataset. We report mean Precision (P), Recall (R) and F1-score (F1) for each account type (Human vs. Bot) and the mean global accuracy (Acc.).

tweets with GPT-2 and incorporating them within two simple approaches, we are able to boost the classification performance for splits with low scores.

Our experiments open new research avenues in the domain. First, the quality of generated tweets needs to be improved. Finding a way of selecting hard examples could allow us to generate tweets that will help the classifier to make decisions. Second, results obtained for GPT-2 tweets suggest that the use of text generation algorithms will become more and more difficult to detect. Texts generated with newer models will likely be harder to detect and the detection models, as well as the training datasets, should be continuously updated to keep pace with the fast evolution of the works in the field of generative AI.

4.3.3 Relevant publications

- Tourille, J., Sow, B., and Popescu, A. (2022, June). Automatic Detection of Bot-generated Tweets. In Proceedings of the 1st International Workshop on Multimedia AI against Disinformation (pp. 44-51), 2022. [202].
  Zenodo record: https://zenodo.org/record/8004475.

4.3.4 Relevant software, datasets and other resources

- The Pytorch implementations can be found in
  https://github.com/zipengxuc/PPE

4.3.5 Relevance to AI4media use cases and media industry applications

Our approach for bot-generated tweet detection was already packaged and integrated in the demonstrator of UC1 that focuses on disinformation detection. Our method can detect bots after only one tweet and has the potential to counter disinformation campaigns in an effective manner.

5 Hybrid, privacy-enhanced recommendation (T6.3)

Contributing partners: FhG-IDMT, UPB

T6.3 focuses on the development of a novel privacy-aware hybrid recommendation framework. When drafting the proposal, the goal was to connect T6.3 and T4.5 (Methods for detection and mitigation of bias affecting fairness in recommender systems) tightly to one of the AI4Media use cases, to ensure that the research can be backed up by real data and evaluated in a real system. This turned out to be difficult, not least due to the lack of available user data and confirmed business cases, and caused delays for the task, which was postponed by FhG-IDMT several times, as we wanted to avoid "decoupled" research that is not connect to a use case.

At some point, however, we decided to start research on knowledge graph based approaches (see section 5.1) that we considered to be broadly applicable, and to create the use case connection by proposing (together with VRT) a recommender system challenge in the second open call of AI4Media, which led to the acceptance and funding of the MAGNET (*Automatic Recommendation of In-Context Media Content to Support Exploratory Research in Journalism*)[11] project.

5.1 Explainable Knowledge Graph based recommendation for News

Contributing partner: FhG-IDMT

For the research, we started work on explainable recommender systems in the news domain to tackle the problem of detecting bias, which will hopefully provide means to mitigate it as well. To this end, a paper titled "An Explainable Knowledge Graph-Based News Recommendation System" has been submitted and accepted at the KDIR 2023 conference, with the following synopsis:

"The paper outlines an explainable knowledge graph-based recommendation system that aims to provide personalized news recommendations and tries to explain why an item is recommended to a particular user. The system leverages a knowledge graph (KG) that models the relationships between items and users' preferences, as well as external knowledge sources such as item features and user profiles. The main objectives of this study are to train a recommendation model that can predict whether a user will click on a news article or not, and then obtain the explainable recommendations for the same purpose. This is achieved with three steps: Firstly, KG of the MIND dataset are generated based on the history and, the clicked information of the users, the category and subcategory of the news. Then, the path reasoning approaches are utilized to reach explainable paths of recommended news/items. Thirdly, the proposed KG-based model is evaluated using MIND News data sets. Experiments have been conducted using the MIND-demo and MIND-small datasets, which are the open-source English news datasets for public research scope. Experimental results indicate that the proposed approach performs better in terms of recommendation explainability, making it a promising basis for developing transparent and interpretable recommendation systems."

5.1.1 Method

The following presents a new explainable KG-based algorithm for personalized recommendations. With its ability to incorporate a large amount of knowledge from various sources and represent it in a comprehensible and transparent manner, it can provide users with recommendations that are not only accurate but also easy to understand. It can be especially useful in domains where trust and transparency are crucial. We use the publicly available MIND News datasets which vary in domain, extensiveness, and sparsity. For scalability purposes, the reduced version of this dataset is utilized and obtained with the following steps:

[11] https://www.ai4media.eu/winners2/magnet/

In the first step, the raw news data is cleaned and pre-processed to generate explainable recommendations. The raw dataset is composed by the following files and illustrated in the left part of figure 40.

- behaviours.tsv: List of users and some demographical data.
- news.tsv: the catalog of news and the entities.

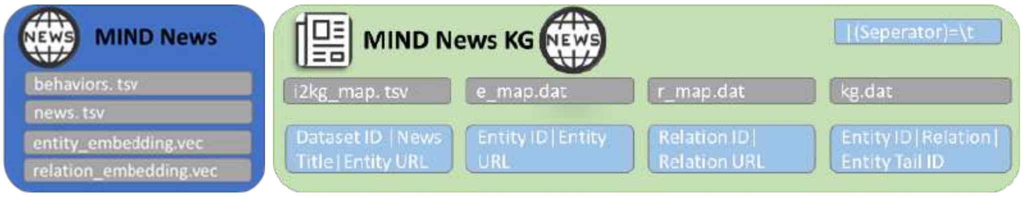


Figure 40. The raw data (left) and (right) the standardized KG model of the Mind News Dataset [203]

A dataset reduced to its K-cores (i.e., dense subsets) is a subset with removed items and users, such that each of the remaining users and items have k reviews each. Hence, these datasets are reduced to its 5-cores and transformed to the standardized format. Then, simple time-based data and knowledge graph embeddings are obtained for training stage and finally, the proposed Explainable Knowledge Graph-Based Recommendation Model (EKG-RM) is evaluated on this metadata. The standardized KG model is composed by 4 different main files and generated as in [203] and illustrated in the right part of Figure 40.

The relations are extracted from behaviors and the news files of the MIND dataset. The history, the non-clicked and clicked information from behaviors files are used to define history and clicked relations. The category, and subcategory information from the news file are used to define same category relations. Hence three different relations are defined in the KG model of MIND dataset, which are history, clicked and same category.

In the second step, the path reasoning models for KG-based recommendation systems are applied on the proposed model. A path-reasoning algorithm starts from a specific user and proceeds through the graph to discover the preferable items in the graph for the target user. The objective is that if the system bases its results on an explicit reasoning path, it is easy to interpret the reasoning process leading to each recommendation, i.e., providing the relevant reasoning paths in the graph as interpretable evidence for why a particular recommendation is made. For instance, considering user A, the proposed model is trying to find candidate News B and News C, along with their explainable paths in the graph.

The PGPR (Policy-Guided Path Reasoning, [204]) model relies on a Reinforcement Learning (RL) agent that is conditioned to the user and trained to navigate to potentially relevant items. The RL agent can be performed with an explicit multi-step path reasoning over the graph starting from a given user node to find out appropriate items in the graph for the target user. Then, the path from the user u to the item i can be used to explain the recommendation.

The CAFÉ (CoArse-to-FinE neural symbolic reasoning approach, [205]) model creates a personalized user profile based on transactions of the user in the KG and utilizes neural symbolic reasoning modules in the path reasoning stage. A layout tree is generated with the modules based on the user profile, then this tree is used by the path reasoning algorithm to generate a set of recommendation paths. The path inference is structured using a layout tree, which offers more efficiency compared to PGPR.

		Precision (%)	Recall (%)	Hit Ratio (%)	NDCG^1
MIND-Demo	PGPR	1.17	3.11	10.45	4.40
	CAFE	**2.85**	**8.02**	**5.17**	**6.28**
Mind-Small	PGPR	0.55	2.66	5.22	2.41
	CAFE	**1.28**	**5.88**	**2.33**	**3.97**

Table 23. Experimental results of applying PGPR and CAFE on the MIND-Demo and MIND-Small datasets.
[1] Normalized discounted cumulative gain

5.1.2 Experiments

The proposed algorithm is applied on the open-source English MIND News dataset[12] that was collected from Microsoft News website in 2019. Users were randomly selected who had at least 5 news clicks, and each user is hashed into an anonymized ID to protect user privacy. The news click behaviours of the users were formatted into impression logs and valued as 1 for click and 0 for non-click. In addition, small versions of MIND-Small and MIND-Demo dataset were released by randomly sampling 50,000 and 5,000 users and their behaviour logs.

The experimental results based on the evaluation metrics of the MIND-Demo and MIND-Small datasets are given in Table 23. The experimental results demonstrate that the CAFÉ model performs better than the PGPR model on the MIND News datasets and should be taken as a baseline for future work.

5.1.3 Conclusions

Explainable recommendation systems based on KGs have emerged as a promising approach to address the challenges of traditional recommendation systems. By leveraging KGs, such systems can model complex relationships between entities and provide personalized recommendations that are more accurate and diverse. Moreover, the explainability aspect of these systems enables users to understand the underlying reasoning behind the recommendations, thereby improving their trustworthiness.

The proposed approach applies KG graph generation combining extracted news metadata and five different relationships, and state-of-art PGPR and CAFE algorithms to find explanation paths between a user and the recommended items in the KG to explain the recommendations. Experimental results on real-world MIND-Small and MIND-Demo datasets indicate flexibility of the model to incorporate multiple relation types and show that the proposed approach offers a promising solution for explainable news recommendations to users.

As for directions for future work, the explainable recommendation quality needs further improvement by extending the KG and the selecting appropriate parameters for the MIND-News dataset. Another objective as a future work is to advance the development and implementation of privacy-preserved news explainable recommendation systems that prioritize user privacy and provide enhanced user control over their personal data.

5.1.4 Relevant publications

- Z. Kurt, T. Köllmer, and P. Aichroth; "An Explainable Knowledge Graph-Based News Recommendation System", Submitted and accepted at KDIR2023 conference.

[12] https://msnews.github.io/

5.1.5 Relevant software, datasets and other resources

- No additional resources published yet.

5.1.6 Relevance to AI4media use cases and media industry applications

This work is potentially relevant for UC2 “AI for News - The Smart News Assistant”, but explainable recommender systems can be applied widely in many media industry applications. The societal damages by filter bubbles are evident, and one important way to tackle these problems are steps to make recommender systems more explainable.

6 AI for Healthier Political Debate (T6.4)

Contributing partners: BSC, AUTH, UvA

The intersection of political debate and technology has given rise to unprecedented challenges, including the spread of misinformation, echo chambers, and polarized narratives that hinder constructive dialogue. This Section covers some important aspects related to the study of AI models applied to political debates. The contributions to this Section propose novel methods for sentiment analysis and opinion polling, that promise more accurate, frequent, and inexpensive pubic opinion polling, while also looking into new sentiment definition, based on a 4-dimensional scale. Furthermore the partners propose several methods for studying other interesting dimensions of political debate, including fallacies, propagandist and argumentative speech, debate leadership, and ephemerality. This Section also proposes an analysis of the objective and subjective perspectives.

6.1 Political Tweet Sentiment Analysis for Public Opinion Polling

Contributing partner: AUTH

Public opinion measurement is a classical political analysis tool, e.g. for predicting election results. Recently, opinion polls by population sampling and questioning have been used to gratify this need rather efficiently. However, they are expensive to run and their results may be biased primarily due to improper population sampling. As social networks are used as political dialogue fora, there is an opportunity to engineer automated processes for measuring public opinion using social network texts, e.g. political tweets. In this work, an innovative way for employing tweet sentiment analysis results is proposed and also, a novel hybrid way to regress poll results from tweets is introduced. This method enables more accurate, frequent, and inexpensive public opinion estimation.

Public opinion consists of concepts, ideas, and statements that seem too abstract to quantify. However, the popularity of certain public *entities* whether political parties, politicians, or even products and services is much more easily quantifiable. Let us consider the cause of n (political) entities, each having an unknown popularity score $p_i, i = 1, ..., n$. As political voting is a competitive procedure, the political score (essentially the voting intention) represents the percentage of people that would prefer entity (political party or candidate) i from all other entities. The popularity score distribution $p_i, i = 1, ..., n$ is initially unknown. It can be estimated in various ways, e.g. through conventional population sampling and polling (through questioning) or by social media data analysis. Any polling method leads to a probability distribution estimate $\hat{\mathbf{p}} = [\hat{p_1}, ..., \hat{p_n}]^T$ that should be as close as possible to the probability distribution $\mathbf{p} = [p_1, ..., p_n]^T$ that can become known only infrequently in special occasions, e.g. through an election/voting procedure. The estimation error $|\mathbf{p} - \hat{\mathbf{p}}|$ is small if certain conditions are met, e.g. that the population sampling set is representative and relatively large. Of course, different population samples and polling methods, lead to different estimates $\hat{\mathbf{p}}$.

Social media users that are politically active, e.g. by posting political tweets, form a special population subject that can be monitored very easily. Let $\mathcal{P}$ be the total population set and $\mathcal{P}_m, \mathcal{P}_o$ be the population subsets of a) people that are politically active in social media and b) people participating in a public opinion poll. Each member of the set $\mathcal{P}_m$ produces political texts that, in many cases, refer to the political entities $1, ..., n$ in question. Social media hashtags can be used for such an association of text to a political entity. Text sentiment analysis can be used to classify such texts into sentiment classes $\{positive, neutral, negative\}$ and quantify their respective text (e.g. tweet) numbers a_i, b_i, c_i respectively for each political entity $i = 1, ..., n$. The data analysis problem at hand is to regress $\hat{\mathbf{p}}$ from the sentiment data set $\mathcal{S} = (\hat{a_i}, \hat{b_i}, \hat{c_i}), i = 1, ..., n$. As sets $\mathcal{P}, \mathcal{P}_m, \mathcal{P}_o$ differ, the respective estimates $\hat{\mathbf{p}}, \hat{\mathbf{p_m}}, \hat{\mathbf{p_o}}$ will be different. Since voting results are too

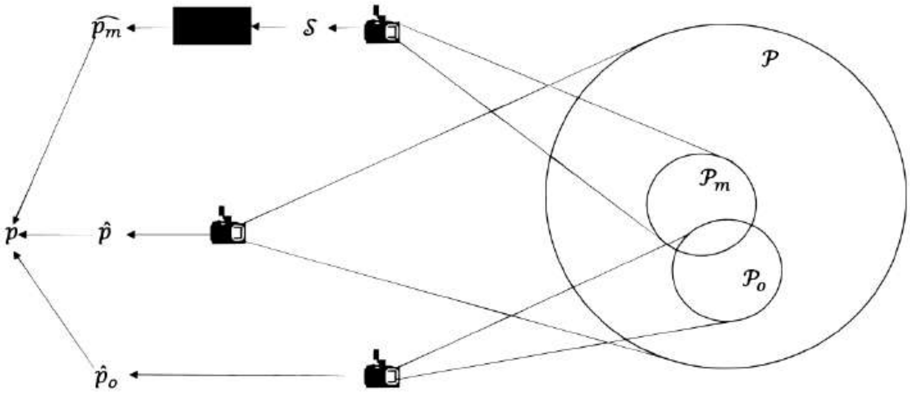


Figure 41. The population set $\mathcal{P}$, the twitter users subset $\mathcal{P}_m$ and the political opinion poll participators subset $\mathcal{P}_o$.

infrequent and transitional polling results are more frequent, we can use $\mathbf{\hat{p_o}}$ to try and regress opinion poll results from $\mathcal{S}$, without performing new traditional opinion polls (Figure 41).

6.1.1 Experiments

The popularity score distribution $p_i, i = 1, ..., n$ can be heuristically estimated from sentiment-labeled political tweets as follows. Firstly, we perform tweet sentiment analysis and automatically tag each political tweet corresponding to a political party (as identified by the tweet hashtags) as positive, neutral or negative.

Let, $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ be n - dimension vectors where a_i, b_i, c_i represent the total number of positive, neutral, negative tweets for party $i, ..., n$. Let vector $\mathbf{d} = \mathbf{a} + \mathbf{b}$ represent the sum of positive and neutral tweets numbers. The heuristic popularity score:

$$\hat{p_i}(\mathbf{c}, \mathbf{d}) = \frac{d_i}{d_t} \cdot (c_t - c_i), \tag{7}$$

$$d_t = \sum_{i=1}^{n} d_i, c_t = \sum_{i=1}^{n} c_i$$

distributes the total negative tweet count (without the ones of candidate i) according to candidate's own positive and neutral comment numbers. As the popularity score distribution should satisfy $\sum_{i=1}^{n} \hat{p_i}(\mathbf{c}, \mathbf{d}) = 1$, we modify this heuristic estimator accordingly:

$$\hat{p_i}(\mathbf{c}, \mathbf{d}) = \frac{n \cdot d_i \cdot (\mathrm{c}_t - c_i) + \mathbf{d}^T\mathbf{c}}{n \cdot c_t \cdot d_t} \tag{8}$$

The proposed popularity score estimator was compared, in our experiments, with popularity estimators proposed in [206], [207], as well as the obvious estimators of assigning popularity scores according to the percentage of positive tweets in [208], or according to the percentage of references for candidate i [209]:

$$\hat{p_i}(\mathbf{a}, \mathbf{c}) = \frac{a_i}{c_i} \tag{9}$$

$$\hat{p_i}(\mathbf{a},\mathbf{c}) = log\frac{a_i+1}{c_i+1} \tag{10}$$

$$\hat{p_i}(\mathbf{a}) = \frac{a_i}{\sum_{i=1}^{n} a_i} \tag{11}$$

$$\hat{p_i}(\mathbf{a},\mathbf{b},\mathbf{c}) = \frac{a_i+b_i+c_i}{\sum_{i=1}^{n}(a_i+b_i+c_i)} \tag{12}$$

We have also extended the bi-party political score estimator [210]:

$$\hat{p_i}(\mathbf{a},\mathbf{c}) = \frac{a_i}{a_i+c_i}\frac{N_i}{N_i+N_j} \tag{13}$$

where, N_i is the total number of tweets for entity i. To the multi-party case:

$$\hat{p_i}(\mathbf{a},\mathbf{b},\mathbf{c}) = \frac{a_i}{a_i+c_i}\frac{a_i+b_i+c_i}{\sum_{i=1}^{n}(a_i+b_i+c_i)} \tag{14}$$

and used it in our experiments.

Heuristic popularity score predictions over the years have shown promising results, but prediction accuracy is sub-optimal as neither ground truth nor optimization criteria have been used in their derivation. Given this fact, we can handle this accuracy loss by resorting to past public opinion poll results and using them as ground truth data. To this end, we implemented a regression model mapping the aforementioned positive, neutral, and negative counts $a_i, b_i, c_i, i = 1, ..., t$ for a certain time window before an opinion poll, on public opinion poll data. A simple regression model:

$$\hat{\mathbf{p}} = \mathbf{w}^T\mathbf{x} + \mathbf{b} \tag{15}$$

is sufficient. $\mathbf{x} = [\mathbf{a_1}^T||\mathbf{b_1}^T||\mathbf{c_1}^T||...||\mathbf{a_t}^T||\mathbf{b_t}^T||\mathbf{c_t}^T]$ is the input vector of size $3nt$, $\hat{p}$ is the poll data estimator $\hat{\mathbf{p}} = [\hat{p_1}, ..., \hat{p_n}]^T$ and $\mathbf{w}, \mathbf{b}$ are trainable regression parameters. Training of $\mathbf{w}, \mathbf{b}$ is performed using the Mean Squared Error (MSE) loss $(\hat{\mathbf{p_o}} - \hat{\mathbf{p}})^2$. Once the regression model (9) is trained on sample data $\mathcal{D} = (\hat{p_{oi}}, \mathbf{x_i}), i = 1, ..., L$, with L being the total number of training opinion polls, it can estimate using $a_i, b_i, c_i, i = 1, ..., t$ within the immediate past to provide a current political popularity score estimate $\hat{\mathbf{p}}$.

The chosen approach is to identify outlying polls and avoid using them in regression model training. Supposing that $u_{ij}(t_k)$ is the estimation of entity j popularity in a poll conducted by company $i, ..., m$ at date t_k. As the poll dates differ across polling companies, we perform linear interpolation for each political entity between two consecutive polls conducted by the same company for a given date t, by using the formula:

$$u_{ij}(t) = u_{ij}(t_k) + (t - t_k)\frac{u_{ij}(t_{k+1}) - u_{ij}(t_k)}{t_{k+1} - t_k}, t_k \le t \le t_{k+1} \tag{16}$$

Then we can compute the Mean Absolute Error (MAE) $e_i(t)$ for company i on date t between the polls of company i and the rest of the m:

$$e_i(t) = \sum_{k=1,k\neq i}^{m} \frac{\sum_{j=1}^{n}|u_{ij}(t) - u_{kj}(t)|}{n}, i = 1, ..., m \tag{17}$$

If this error is above a threshold T, an outlying poll is deleted and filtered out from regression. To this end, the error threshold T can be estimated, e.g. by the average of all errors $e_i(t)$ by day, throughout the entire polling period.

Heuristic estimators	200 days	180 days	150 days	100 days
8	**5.07%**	**5.23%**	**5.3%**	**5.3%**
9	20.73%	20.51%	20.38%	20.5%
10	18.7%	18.87%	18.96%	19.14%
11	7.01%	6.91%	7%	7.29%
12	9,39%	9.5%	9.47%	8.95%
14	6,44%	6.91%	7.22%	7.9%

Table 24. Heuristic estimators' MAE compared to opinion polls

More than 300,000 tweets have been gathered from about six Greek political parliamentary parties, using the Twitter API from the 14th June 2022 until the 31st December 2022. All tweets have been labeled as neutral, positive, or negative using the Transformer proposed in [211], which scores 79% accuracy, tested on ground truth Greek political tweets [212]. During the data-gathering period, 19 public opinion polls were collected and utilized to train the proposed regression model and validate the proposed techniques.

In order to evaluate and compare five different heuristic estimators 8-12,14 on political Greek Twitter data, we calculated the popularity score, for every party inside the Greek parliament during the data collection period. Then, in order to compare the different heuristic estimator outputs we calculated the Mean Absolute Error (MAE), defined as the average error of each predictor between the opinion poll results (used as ground truth) and the estimator output:

$$e = \frac{\sum_{i=1}^{n} |\hat{p_{oi}} - \hat{p_i}|}{n} \tag{18}$$

where s_i and p_i are the opinion poll results and their estimation, respectively. As estimators 9 and 10 do not sum to 1 for all n entities, we normalised them first:

$$\acute{p_i} = \frac{\hat{p_i}}{\sum_{i=1}^{n} \hat{p_i}} \tag{19}$$

Table 24 presents testing results during 4 periods, starting from 31 Dec 2022 until 100 to 200 days backward. It is clear that the proposed heuristic estimator 8 outperforms other formulas during all windows tested.

6.1.2 Conclusions

In this paper, we proposed two new methods for estimating political popularity scores through sentiment analysis of Twitter data: both a heuristic and a regression method are proposed. They both provide rather good estimation of political popularity scores. Although, the difference between formula popularity estimators and hybrid Twitter-poll ones is still considerable, as NLP tools get more advanced the results we get from political forecasting through Twitter should become more and more accurate, but for the time being poll analysis outperforms them.

6.1.3 Relevant publications

- I. Pitas and A.Kaimakamidis, "Political Tweet Sentiment Analysis for Public Opinion Polling", Technical Report submitted as conference paper [213].

6.1.4 Relevant software, datasets and other resources

- Political Barometer is a software estimating political public opinion, by using political tweets, which are collected and analysed daily. The relevant web app of the software can be found in http://icarus.csd.auth.gr/political-barometer/
- AUTH has created the "GreekPolitics Dataset" in the context of the "AI4Media" and is described in [212]. To access the "GreekPolitics Dataset" refer to https://aiia.csd.auth.gr/auth-greekpolitics-dataset/

6.1.5 Relevance to AI4media use cases and media industry applications

This work finds relevance in two distinct AI4Media use cases, UC2 "AI for News - The Smart News Assistant" and UC4 "AI for Social Sciences and Humanities", where AI-driven tools are integral to achieving specific goals in the fields of news creation and social sciences and humanities (SSH) research. Specifically for UC2, our method offers journalists an opportunity to enhance public opinion measurement, providing more accurate, frequent, and cost-effective results compared to traditional polling methods. Moreover, since one of the big challenges of SSH researchers is the inefficiency of manual media investigation methods, our research aligns with UC4 by automating and scaling up linguistic analysis procedures.

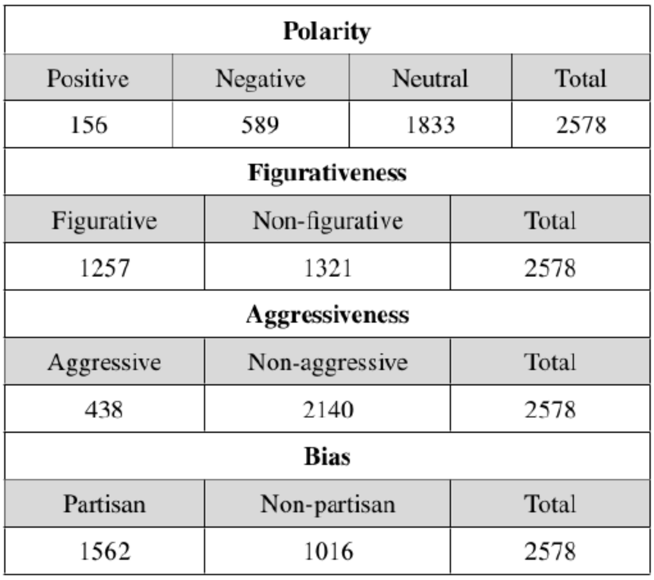
6.2 GreekPolitics: Sentiment Analysis on Greek Politically Charged Tweets

Contributing partner: AUTH

The rapid growth of on-line social media platforms has rendered opinion mining/sentiment analysis a critical area of Natural Language Processing (NLP) research. This work focuses on analyzing Twitter posts (tweets), written in the Greek language and politically charged in content. This is a rather underexplored topic, due to the scarcity of publicly available annotated datasets. Thus, we present and release "GreekPolitics", i.e., a dataset of Greek tweets with politically charged content, independently annotated across four different sentiment dimensions: polarity, figurativeness, aggressiveness, and bias. GreekPolitics has been evaluated comprehensively in a classification setting, separately for each sentiment, using state-of-the-art Deep Neural Networks (DNNs) and data augmentation methods.

GreekPolitics was designed as a large-scale dataset of politically charged Greek tweets, accompanied by full ground-truth labels along the proposed 4 sentiment dimensions. This multidimensional sentiment definition is expected to allow fuller semantic characterization of a tweet. GreekPolitics Twitter posts were collected based on specific query hashtags related to the Greek political scene, using the official Twitter API. These hashtags are mainly related to the names of the various political parties and popular politicians represented in the Greek parliament over the past decade, while variants of them (in both the Greek and the Latin alphabet) were also exploited. Collected tweets span a large time scale, from January 2014 up to March 2021. After an initial data-cleaning stage, we ended up with over 8,000 tweets. After removing retweets, duplicates, or poorly written posts, the final dataset contained 2,578 unique tweets.

Based on the proposed 4-dimensional sentiment definition, each individual tweet was independently annotated for classification with respect to polarity, figurativeness, aggressiveness, and bias.



Polarity			
Positive	Negative	Neutral	Total
156	589	1833	2578

Figurativeness		
Figurative	Non-figurative	Total
1257	1321	2578

Aggressiveness		
Aggressive	Non-aggressive	Total
438	2140	2578

Bias		
Partisan	Non-partisan	Total
1562	1016	2578

Figure 42. Number of annotated tweets per sentiment class.

As far as *polarity* is concerned, each tweet was assigned one out of three possible class labels: "positive", "negative" or "neutral". For *figurativeness*, each tweet was assigned the ground-truth label of either "figurative" (ironic, sarcastic or figurative in general) or "normal" (i.e., non-figurative, literal). Regarding *aggressiveness*, each tweet was annotated with either an "aggressive" (offensive, abusive, racist or aggressive in general) or "normal" (i.e., non-aggressive) label. Finally, for *bias*, each tweet was annotated as either a "partisan" (if it expressed a strong, supportive and adamant opinion) or a "neutral" (i.e., non-partisan) one. Therefore, each tweet was manually and independently annotated with: i) 1 label for a 3-class classification task, and ii) 3 different labels for 3 binary classification tasks (one per task).

Manually annotating a tweet with the employed four labels (one per dimension) may potentially lead to different results for each individual human annotator. In order to provide as objective ground-truth annotations as possible, a team of three volunteers was asked to classify each tweet in the dataset with respect to the classes of each different sentiment dimension. Inter-annotator agreement was subsequently calculated and annotations with majority agreement were selected as the actual annotations of the tweets in question. The resulting class split for each sentiment is presented in Figure 42.

Figure 43 shows a huge inter-class imbalance with regard to the ground-truth class size: in the case of polarity, positive tweets are significantly less than negative or neutral ones. This is because most Twitter users tend to express rather negative or neutral opinions regarding politics. This effect may cause great issues when training and evaluating machine learning models since a classifier could undesirably learn to favor at the test stage those classes that were the most large-sized during training.

After the collection and annotation of tweets, they were preprocessed in order to produce the final GreekPolitics dataset. While reading and comprehending a tweet's content is done effortlessly by a human, providing raw input to machine learning models is not ideal. Tweets may include undesirable content, such as hashtags, URLs, or emojis, which could impede successful sentiment analysis of their actual text. Thus, the following preprocessing steps were followed:

- Remove all mentions (i.e., text starting with @), URLs, emojis, or any other special character.

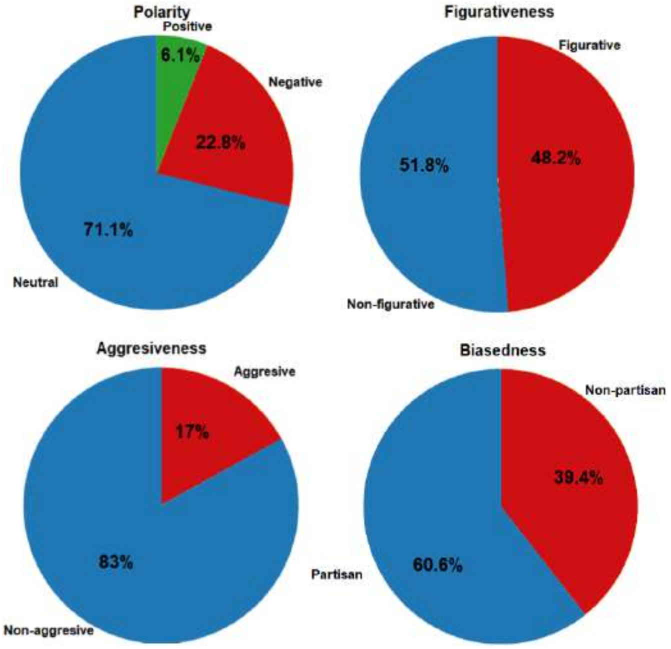


Figure 43. Data distribution for every classification task (i.e., sentiment dimension).

Since hashtags may be entirely meaningful, they were retained.

- Strip accents. Greek words usually contain accents that are striped. Remove all punctuation. Punctuation marks do not help us discriminate between different text sentiments.
- Remove multiple spaces and line breakers so that each tweet is expressed in one line and each word is separated by a single space.
- Convert all words to lowercase in order to achieve data uniformity and avoid ambiguity.
- Tokenize the sentences. Tokenization is the process of splitting a piece of text into smaller units called tokens. For GreekPolitics, each text was tokenized by considering the words as splitting tokens.

6.2.1 Experiments

Four different classifiers were trained separately for the four classification tasks (i.e., sentiment dimensions). Polarity analysis was addressed as a 3-class classification task, while binary classifiers were employed for the remaining three sentiment dimensions. Two different DNNs were investigated as alternative options: i) a Convolutional Neural Network (CNN) adopted from [214], and ii) a Transformer adopted from [215]. The CNN has 5 1D convolutional layers, each one with an increasing amount of output filters. The Transformer has an encoder module composed of a multi-head self-attention layer, a normalization layer with a residual connection, two fully-connected layers, and a final normalization layer with a residual connection. The output of both the CNN and the Transformer is fed to a fully-connected classification layer with as many neurons as the number of classes.

A pretrained FastText [216] word embedding DNN was first utilized for transforming a given word into a unique 300-dimensional vector. Subsequently, the resulting vector representations of all words in a tweet were concatenated into a matrix $T \in \mathbb{R}^{M \times 300}$, where M is chosen as the maximum sentence length in terms of word count. If a sentence does not exceed the maximum length T, it is appropriately padded with zero-value vectors. Given the text of the specific tweets in GreekPolitics, M was set to 60.

80%/20% of the GreekPolitics tweets were used for the training/test set, respectively. DNNs were trained separately for each sentiment dimension, using random parameter initialization, the Adam optimizer, a categorical cross-entropy loss, a learning rate of 0.001, a mini-batch size of 64, and a total of 60 epochs.

6.2.1.1 Polarity Table 25 reports results on 3-class polarity classification, both on the overall test set and on each individual class. The huge class imbalance (i.e., the "positive" contains only 6.1% of the GreekPolitics tweets) leads unsurprisingly to poor class-specific performance. Thus, Table 25 reports evaluation results based on the best-measured performance on the minority classes, as well as precision and recall metrics. Initial comparisons in the original dataset split are showcased in the top section of Table 25 (i.e., the models with the suffix 1). The superiority of the Transformer against the CNN and the unacceptable accuracy in the minority classes are evident.

Model	Augment.	Pruning	Acc. positive	Acc. negative	Acc. neutral	Overall Acc.	Precision	Recall
CNN-1	-	-	0.24	0.42	0.90	0.76	0.75	0.73
Transformer-1	-	-	**0.39**	**0.67**	**0.89**	**0.81**	**0.805**	**0.806**
CNN-2	✓	-	0.57	0.61	0.81	0.78	0.77	0.73
Transformer-2	✓	-	**0.62**	**0.74**	**0.89**	**0.83**	**0.84**	**0.83**
CNN-3	✓	✓	0.74	0.67	0.79	0.74	0.75	0.74
Transformer-3	✓	✓	**0.83**	**0.78**	**0.82**	**0.81**	**0.83**	**0.82**

Table 25. Individual class accuracies, overall accuracy and overall precision/recall metrics for polarity classification. The ✓ symbol indicates that this method was applied for training the respective model.

In order to improve accuracy in the minority classes, data augmentation was applied. Ideally, new tweets should be generated based on the existing ones, which would maintain identical ground-truth sentiment but would be expressed in a different manner. Two text augmentation methods were applied:

1. Back-translation [217]. A sentence is translated into another language, e.g., from Greek to English, and then back to the source language. If the newly generated sentence is different but holds the exact same meaning, it is used as an augmented version of the original one.

2. Synonym substitution [218] [219]. In synonym substitution, given a certain probability, each word in a sentence is replaced with a synonym one acquired from an external vocabulary of synonyms, e.g., Thesaurus. The new sentence is again used as an augmented version of the original one.

Augmentation was applied only on 80% of the "positive" samples, i.e., the samples used for training, with the test set remaining untouched. Thus, out of the 156 "positive" training samples,

Model	Acc. figurative	Acc. non-figurative	Overall Acc.	Prec.	Recall
CNN	0.60	0.71	0.66	0.64	0.68
Transf.	**0.70**	**0.75**	**0.72**	**0.70**	**0.72**

Table 26. Individual class accuracies, overall accuracy and overall precision/recall metrics for figurativeness classification.

Model	Aug.	Acc. aggressive	Acc. non-aggressive	Overall Acc.	Pre.	Rec.
CNN-1	-	0.52	0.86	0.69	0.66	0.68
Transf.-1	-	**0.55**	**0.96**	**0.88**	**0.86**	**0.88**
CNN-2	✓	0.64	0.85	0.75	0.74	0.74
Transf.-2	✓	**0.69**	**0.91**	**0.86**	**0.85**	**0.86**

Table 27. Individual class accuracies, overall accuracy, and overall precision/recall metrics for aggressiveness classification.

Model	Aug.	Acc. partisan	Acc. non-partisan	Overall Acc.	Pre.	Rec.
CNN-1	-	0.83	0.44	0.68	0.65	0.66
Transf.-1	-	**0.85**	**0.67**	**0.77**	**0.70**	**0.72**
CNN-2	✓	0.84	0.52	0.70	0.69	0.69
Transf.-2	✓	**0.86**	**0.77**	**0.81**	**0.79**	**0.81**

Table 28. Individual class accuracies, overall accuracy, and overall precision/recall metrics for bias classification.

additional 111 tweets were generated. After retraining from scratch on the augmented dataset for polarity classification, the results reported in the middle section of Table 25 (i.e., the models with the suffix 2) were obtained. As it can be seen, augmentation resulted in a huge increase in test accuracy for the "positive" class. However, the inter-class accuracy gap remains considerable (0.62/0.74/0.89 for "positive"/"negative"/"neutral", in the Transformer model).

To further reduce the inter-class accuracy gap, the "neutral" training samples were pruned, while the augmented "positive" ones were retained. Thus, out of the 1833 "neutral" samples, 737 were removed. The results are reported in the bottom section of Table 25 (i.e., the models with the suffix 3). It can be seen that the test accuracy for the minority classes increased significantly, with a small drop in the accuracy of the "neutral" class. This inter-class balancing effect under the presence of ground-truth class imbalances is relevant in application domains where it is more important to estimate the comparative/relative volumes of "positive", "neutral" and "negative" tweets, rather than accurately characterize a specific tweet.

6.2.1.2 Figurativeness Figurativeness classification results are reported on Table 26. The ground-truth classes are rather balanced and the obtained accuracy is acceptable, so no training dataset processing was applied to boost test performance. The Transformer again surpasses the CNN.

6.2.1.3 Aggressiveness Aggressiveness classification results are reported on Table 27. A huge class imbalance was again observed, as the "offensive" tweets are few (i.e., 20% of the dataset). The augmentation strategies previously described for polarity classification were adapted and independently/separately applied: out of the 438 "offensive" training samples, an additional 256 augmented samples were generated. Initial comparison results for the original dataset are showcased in the top section of Table 27 (i.e., the models with the suffix 1). Results produced by retraining the employed models on the augmented scenario are reported in the bottom section of Table 27 (i.e., the models with the suffix 2). As before, augmentation leads to a massive increase in model accuracy for the minority class, while maintaining acceptable accuracy for the initially dominant class.

6.2.1.4 Bias Bias classification results are reported in Table 28. Moderate data imbalance was observed here for the two classes (39% and 61%). To counter it, the augmented tweets already generated for the polarity and aggressiveness tasks (annotated as "non-partisan") were employed for retraining from scratch. Thus, 202 augmented training tweets were added to the original 1016 "non-partisan" ones. Initial comparison results in the original dataset are showcased in the top section of Table 28 (i.e., the models with the suffix 1), while results for the augmented dataset are reported in the bottom section of Table 28 (i.e., the models with the suffix 2).

6.2.2 Conclusions

This paper introduced the recently captured/annotated "GreekPolitics" dataset for sentiment analysis of politically charged Twitter posts in the Greek language. The tweets have been manually labeled for classification across 4 independent sentiment dimensions (polarity, figurativeness, aggressiveness, and bias). A thorough experimental study was conducted by utilizing state-of-the-art Deep Neural Networks (DNNs), which yielded promising results. The domain-specific problem of data class imbalance (too few positive tweets) was tackled in the experimental evaluation by employing standard data augmentation tricks, based on generating novel tweet samples from the existing ones. Such methods can generate a limited amount of new training samples and should be validated by humans. Interesting future work could explore more sophisticated data augmentation for natural language text, which could ideally be applied without any human supervision.

6.2.3 Relevant publications

- Patsiouras E., Koroni I., Mademlis I. and Pitas I., "GreekPolitics: Sentiment Analysis on Greek Politically Charged Tweets", European Signal Processing Conference (EUSIPCO), 2023 [212].

6.2.4 Relevant software, datasets and other resources

- AUTH has created the "GreekPolitics Dataset" in the context of the "AI4Media" and is described in [212]. To access the "GreekPolitics Dataset" refer to https://aiia.csd.auth.gr/auth-greekpolitics-dataset/

6.2.5 Relevance to AI4media use cases and media industry applications

The "GreekPolitics" dataset presents an opportunity to integrate valuable data and AI-driven capabilities into both UC2 "AI for News - The Smart News Assistant" and UC4 "AI for Social Sciences and Humanities". Journalists in UC2 can use it to monitor and analyze political discussions, while SSH researchers in UC4 can leverage it for in-depth investigations into political discourse and sentiment analysis in the Greek language. This dataset aligns with the goals of both use cases, offering the potential to enhance news creation workflows and support research in the social sciences and humanities.

6.3 Fallacious Argument Classification in Political Debates

Contributing partner: 3IA-UCA

Fallacies have a long history in argumentation, contributing significantly to critical thinking education. In today's world, their importance has grown even further, as contemporary argumentation technologies face challenges in detecting misleading and manipulative information in news articles and political discourse, as well as generating counter-narratives.

However, classifying arguments as fallacious remains a difficult task. To address this challenge, we present a two-fold contribution:

- Firstly, we introduce a novel annotated resource comprising 31 political debates from U.S. Presidential Campaigns. We expanded the annotations of the ElecDeb60To16 dataset [220]. This dataset collects televised debates from U.S. presidential election campaigns between 1960 and 2016. To the best of our knowledge, it is the largest dataset of political debates that includes annotations for both argumentative components (Evidence, Claim) and relations (Support, Attack).
- Secondly, we propose a neural architecture based on transformers, which outperforms state of the art results and standard baselines in classifying fallacious arguments.

Despite the few existing approaches ([221], [222]), classifying fallacious arguments largely remains an unsolved task. Our contribution advances the state of the art with a novel resource and an effective method.

6.3.1 ElecDeb60to16

The exploratory study on the arguments of the debates was to determine which fallacy types, as outlined in the annotation scheme of [223] and the categorization of [224], were predominantly present in political discourse. Therefore, the resource contains 1,628 fallacious arguments categorized into six main categories of fallacies: *Ad Hominem, Appeal To Authority, Appeal To Emotion, False Cause, Slogan, Slippery Slope.*

Following the careful formulation of the annotation guidelines, three annotators with expertise in computational linguistics conducted the dataset annotation. After completing the initial round of annotation on a data sample, the guidelines were revised to address disagreements among the annotators, particularly regarding the boundaries of the spans to be annotated.

Next, nine sections from five distinct debates held in different years were annotated to calculate inter-annotator agreement, which is reported in Table 29 as showing moderate agreement. Subsequently, an expert annotator reconciled the annotations before incorporating them into the dataset that was made publicly available.

Fallacy Type	**Observed Agr.**	**Krippendorff's α**
Ad Hominem	0.9961	0.5315
Appeal to Authority	0.9945	0.5806
Appeal to Emotion	0.9759	0.4640
Slogans	0.9989	0.5995

Table 29. IAA, three annotators, 9 sections from 5 different debates (only 4 types of fallacies were present in the annotated data sample.)

6.3.2 Fallacy Classification Task

We cast the fallacious argument classification task as a sequence classification problem. First, we focus on a multi-class classification task to classify the fallacies observed in the debates. Then, we enhance our classifier with argumentation-based features (i.e., argumentative components and relations) within each fallacious argument. To this end, we test and adapt SOTA language models based on the transformer architecture, as they are challenging baselines to compare with. We considered BERT [225] and RoBERTa [226] as baselines.

Each debate in the dataset consists of two main components:

1. The portion of the debate containing the fallacious argument: This part refers to the specific segment within the presidential debate where the fallacious argument is presented.

2. The fallacious argument snippet itself: This refers to the actual excerpt or text snippet that constitutes the fallacious argument within the debate.

In this work, we use more advanced PLMs to tackle contents of considerable length, i.e., Longformer [227] and Transformer-XL [228] which have the ability to capture long-input texts to perform the classification.

6.3.3 Proposed Approach

Our approach is based on the Longformer model empowered with the argumentation features, and the context of the fallacious argument in the debate.

Longformer is a transformer-based model featuring an attention mechanism that scales linearly with the length of the input sequence. This scalability enables it to efficiently process documents with thousands of tokens or longer. The attention mechanism in Longformer combines windowed local-context self-attention with global attention of the context. The local attention in Longformer is mainly employed to create contextual representations, while the global attention enables the model to build complete sequence representations for making predictions. The model undergoes pre-training using the Masked Language Modeling (MLM) approach, similar to RoBERTa [226].

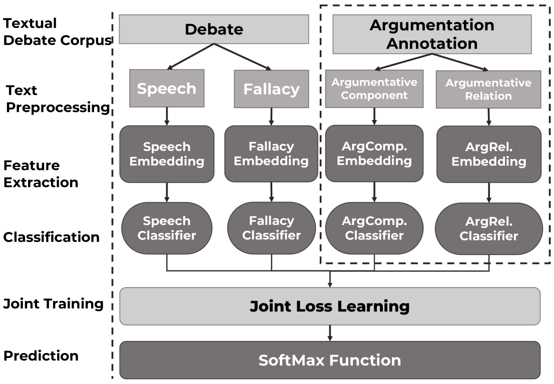


Figure 44. Approach for the task of fallacious argument classification.

Figure 44 visualizes our neural architecture approach for fallacious argument classification. Each debate is processed into four components: the dialogue context, the fallacious argument snippet, the argument component, and argument relation. Each component is then extracted in the embedded vectors using the PLM of interest. Each embedding has its own transformer-based classifier to finally obtain a logit (L).

We produced a different loss per classifier i.e., fallacy-snippet ($loss_{snippet}$), speech ($loss_{speech}$), argument component ($loss_{ArgComp}$), and argument relation ($loss_{ArgRel}$). We then join the $loss$ of each classifier to have a joint-loss learning with $\alpha = 0.5$ [229]. We arrange these alignments of L to calculate the average loss as a joint loss ($loss_{joint\_loss}$) from each $loss$ element. The function used before back-propagation is

$$loss_{joint_{l}oss} = \alpha * \frac{(loss_{speech} + loss_{snippet} + loss_{ArgComp} + loss_{ArgRel})}{N_{loss}} \tag{20}$$

where N_{loss} stands for the number of $loss$ elements taken into the model.

6.3.4 Experiments

Various experimental settings were applied to evaluate different transformer-based Pre-trained Language Models (PLMs) on the classification of main fallacy categories. The models tested include BERT, RoBERTa, Longformer, and Transformer-XL. The implementation was based on the huggingface transformer library, using PyTorch version 1.7.0. The learning rate was set to 5e-5, dropout to 0.1, and batch size to 1 for Transformer-XL, while the batch size was 8 for the other models. The maximum size length for the number of tokens varied for each model: 128 for BERT and RoBERTa, 128 for the fallacious argument snippet and 8192 for the context speech text in Transformer-XL, and 128 for the fallacious argument snippet and 4,096 for the speech context in Longformer.

Additional experiments were conducted using the best-performing model (Longformer + $loss_{joint}$) to *(i)* classify the 14 fallacious argument sub-categories, and *(ii)* classify the main categories while enriching the dataset with argumentative component and relation features in an ablation test setting as shown in Table 30.

In all experimental settings, 80% of the dataset was used for training, and the remaining 20% for testing. The split was performed by sklearn with a random seed for label distribution. To obtain more reliable results, the experiments were conducted three times, and the results were averaged.

Model	Dataset	$loss_{joint_{l}oss}$	Arg. Feat.	Prec.	Rec.	Macro avg F1
BERT	Main Category	No	None	0,62	0,55	0,55
RoBERTa	Main Category	No	None	0,58	0,56	0,53
Longformer	Main Category	No	None	0,64	0,6	0,57
Longformer	Main Category	Yes	None	0,66	0,61	**0,61**
Transformer-XL	Main Category	No	None	0,61	0,45	0,47
Transformer-XL	Main Category	Yes	None	0,61	0,51	0,53
Longformer	Sub-category	Yes	None	0,44	0,45	0,42
Longformer	Main Category	Yes	Comp.	0,88	0,81	**0,83**
Longformer	Main Category	Yes	Rel.	0,87	0,81	**0,83**
Longformer	Main Category	Yes	Comp + Rel	0,84	0,85	**0,84**

Table 30. Results of the multi-class sequence tagging task, on the average of three runs.

Furthermore, we conducted an ablation test in the multi-class classification setting to analyze the impact of argumentative features (components and relations) on classifying main fallacious argument categories. Adding argumentation features to the neural model resulted in significantly improved results, as shown in Table 31. The additional context from argumentative components, combined with speech context, enhanced the model's classification performance for fallacious arguments.

	Original dataset	**Arg. Comp.**	**Arg. Rel.**	**Arg. Comp. & Rel.**
	F1	F1	F1	F1
Ad Hominem	0,56	0,85	0,81	0,81
App. to Auth.	0,65	0,85	0,84	0,91
App. to Em.	0,85	0,93	0,93	0,94
False Cause	0,43	0,80	0,82	0,80
Slippery slope	0,50	0,78	0,79	0,84
Slogans	0,67	0,76	0,88	0,77
accuracy	0,75	0,88	0,89	0,89
macro_avg	**0,61**	**0,83**	**0,83**	**0,84**
weighted_avg	0,74	0,88	0,89	0,89

Table 31. Ablation test with argumentative features in details.

6.3.5 Conclusions

Fallacies continue to be a contentious issue in argumentation, particularly in real-world scenarios like political debates [230]. Existing argumentation schemes for identifying such flawed reasoning can be ineffective, as they often overlook the specific content and dialectical context of the fallacy. In our research, we aim to empirically explore the relationship between the argumentative content and the context of fallacies in our dataset.

Additionally, we have observed that almost every known type of fallacy is closely related to sound arguments in a debate. Thus, we will investigate how to generate sound arguments from the identified fallacies and their context. Furthermore, countering the formal invalidity of these fallacious arguments through newly generated counter-arguments poses a challenging follow-up task in our work.

6.3.6 Relevant publications

- Goffredo, P., Haddadan, S., Vorakitphan, V., Cabrio, E., & Villata, S. (2022, July). Fallacious argument classification in political debates. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI (pp. 4143-4149). [231]

6.3.7 Relevant software, datasets and other resources

- No additional resources published yet.

6.3.8 Relevance to AI4media use cases and media industry applications

This study holds significance in UC2 "AI for News" as it provides valuable findings on detecting and classifying different categories of fallacies in political debates, along with analyzing argumentative features defining such messages. The findings from this study can enhance the precision and credibility of politicians' statements during their debates.

Moreover, it aligns with UC4 “AI for Social Sciences and Humanities” and UC1 AI “Social Media and Against Disinformation” by tackling the necessity to identify misleading and uninformed content. The ability to recognize fallacy aids in combating the dissemination of misinformation and mitigating its negative impacts on society. Indeed, this research plays a crucial role as an initial step towards effectively detecting, explaining, and countering fallacies and disinformation.

6.4 Investigating Semantic and Argumentative Features for Supervised Propagandist Message Detection and Classification

Contributing partner: 3IA-UCA

Propaganda represents an effective, even though often misleading, communication strategy to promote a cause or a viewpoint ([232]–[235]). The ability to effectively identify and manifestly label such kind of misleading and potentially harmful content is of primary importance to restrain the spread of such information to avoid detrimental consequences for society. We tackle this challenging issue by proposing a textual propaganda detection model. More precisely, we address the following research questions: (i) how to automatically identify propaganda in textual documents and further classify them into fine-grained categories?, and (ii) what are the linguistic distinctive features of propaganda text snippets? The contribution of this work consists not only in proposing a new effective neural architecture to automatically identify and classify propaganda in text but in presenting a detailed linguistic analysis of the features characterizing propaganda messages.

Our work focuses on the propaganda detection and classification task, casting it as a binary and as a multi-class classification task, and we address it both at sentence-level and at fragment-level. We investigate different architectures of recent language models (*i.e.*, BERT, RoBERTa), combining them with a rich set of linguistic features ranging from sentiment and emotion to argumentation features, to rhetorical stylistic ones. The experiments we conducted on two standard benchmarks (the NLP4IF’ 19 [236] and SemEval’20-Task 11 [237] datasets) show that the proposed architectures achieve satisfying results, outperforming state-of-the-art systems on most of the propaganda detection and classification subtasks. Furthermore, we analyzed how the most relevant features for propaganda detection impact the fine-grained classification of the different techniques employed in propagandist text, revealing the importance of semantic and argumentation features.

6.4.1 Experiments

6.4.1.1 Feature Analysis We investigate a set of features that we assume to play a role in propaganda. We divide these features into four groups: (i) persuasion, (ii) sentiment, (iii) message simplicity, and (iv) argumentation. Although many of the following metrics are obtained at token-level, aggregation methods (*e.g.* averaging, summation) were used to characterize each sentence.

Persuasion In this group, we consider variables as *speech style*, *concreteness*, and *subjectivity*. We measure them using the lexicon-based tools proposed in [238], [239], and [240] respectively. Also, we rely on BERT (base-uncased) [241] to extract *lexical complexity* features.

Sentiment We employ various lexicons to gather variables related to sentiment. We use SentiWordNet 3.0 [242] for *sentiment labels* (positive, negative, or neutral), DepecheMood++ lexicon [243] for *emotion labels* (*i.e.*, afraid, amused, angry, annoyed, don’t care, happy, inspired, sad), the lexicon proposed in [244] for *VAD* (valence, arousal, and dominance) scores, a lexicon from [245] for *connotation*, and a lexicon from [246] for *politeness* measurement.

Message Simplicity We measure *exageration* employing an imageability lexicon [247]. To capture properties related to *text length* we count the average char-length, actual char-length, word length, punctuation frequency, capital-case frequency per sentence [248], and we apply length encoding at character-level, plus one additional dimension for non-alphabetical char count. Since some

propaganda structures usually contain *pronouns*, we create a lexicon of 123 pronouns in English and perform one-hot encoding of commonly used pronouns in English.
Argumentation To extract argumentative features representing our data, we train a supervised classifier for the task of argumentative sentence classification on the persuasive essays dataset [249]. We first binary classify sentences into argumentative text (claims and premises), or non-argumentative. Then, using only argumentative text sentences, we classify them into claims or premises. To address these tasks, we build and fine-tune a BERT classifier.

6.4.1.2 Models The authors in [236] define the Fine-grained Propaganda Detection task as two sub-tasks with different granularity: (i) Sentence-Level Classification task (SLC), which asks to predict whether a sentence contains at least one propaganda technique, and (ii) Fragment-Level Classification task (FLC), which asks to identify both the spans and the type of propaganda technique.

We used two available datasets for propaganda detection: NLP4IF'19 [236] and SemEval'20 T11 [237].

Sentence-level Classification We employ different models to tackle this task:

1. Pre-trained BERT [241] without fine-tuning and using default hyperparameters.
2. Fine-tuned BERT.
3. Fine-tuned T5 [250], a text-to-text transformer, where we fine-tuned it using the T5ForConditionalGeneration approach. In this setup, the input is a sentence (as a question), and the output is an answer (as a label).
4. Linear-Neuron Attention BERT [251]. This was the winning approach of the NLP4IF'19 shared task. This model builds on the BERT architecture with specific hyperparameter modifications.
5. Multi-granularity BERT [236], which utilizes the BERT transformer with a multi-granularity network on top. This network includes multi-classifiers for different granularity levels of text.
6. Multi-granularity + Featured BERT, an extension of the model proposed by [236]. It focuses on the last layer of sentence-level granularity and incorporates proposed features into a BERT classifier, yielding logits.
7. BERT + Featured BiLSTM, where we use a pre-trained BERT transformer architecture along with a Bidirectional Long Short-Term Memory (BiLSTM) architecture on top. The BERT model's output is combined with propaganda features and serves as input for the BiLSTM model.
8. BERT + Featured Logistic Regression. This approach follows the same idea as the previous model, but instead of a BiLSTM, a logistic regression model is used.

Table 32 presents the results for the SLC task (propaganda vs no propaganda). Each experiment was executed 5 times, and the reported metrics represent the macro-average across all runs. On the NLP4IF'19 dataset, our proposed models surpassed the state-of-the-art models, with an F1-score of 0.72 and a precision-score of 0.80 for BERT + Featured Logistic Regression model and Featured BiLSTM model respectively. For the SemEval'20-T11 dataset, the proposed architecture achieved the best F1-score when using BERT + Featured Logistic Regression. Utilizing semantic features alone demonstrated slightly better results than combining them with argumentation features.

Model	NLP4IF'19 Test Set			SemEval'20-T11 Dev. Set		
	F1	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
BERT ***Baseline***	0.52	0.53	0.50	0.48	0.48	0.48
SOTA						
Fine-tuned BERT	0.58	0.63	0.53	0.61	0.63	0.60
Fine-tuned T5	0.64	0.64	0.65	0.66	0.65	0.66
Linear-Neuron Attention BERT	0.63	0.60	0.67	0.66	0.69	0.63
Multi-granularity BERT	0.61	0.60	0.62	0.65	0.68	0.63
Proposed Architecture w/ Semantic Features						
Multi-granularity + Featured BERT	0.63	0.65	0.61	0.67	**0.71**	0.64
BERT + Featured BiLSTM	0.65	**0.80**	0.55	0.65	0.75	0.58
BERT + Featured Logistic Regression	**0.72**	0.74	**0.70**	**0.68**	**0.71**	0.66
Proposed Architecture w/ Semantic Features + Argumentation Features						
BERT + Featured Logistic Regression	0.71	0.72	0.69	**0.68**	0.70	**0.67**

Table 32. Results on the Sentence-level classification (SLC) task (binary task).

Fragment-level Classification on NLP4IF'19 Dataset The evaluation of the FLC task varied depending on the dataset being used. In the NLP4IF'19 Dataset, which comprises 18 annotated propaganda techniques, predictions were performed at the token-level. It is important to consider that multiple tokens can belong to the same span, and tokens without any propaganda bias are marked as "no propaganda."

For this evaluation, the following models are employed:

1. Fine-tuned BERT (baseline) [13].
2. Fine-tuned RoBERTa (baseline).
3. State-of-the-art Model: This corresponds to the winning team in the NLP4IF'19 shared task [252].
4. Transformer + CRF. We either used a pre-trained BERT or pre-trained RoBERTa model, followed by a CRF layer as the final layer.

Models	NLP4IF'19											
	Average	**Appeal-Fear**	**Black-White**	**Casual-Over.**	**Doubt**	**Exag.-Min.**	**Flag-Waving**	**Loaded-L.**	**Namecalling**	**Reductio-Hit.**	**Repetition**	**Slogans**
Baseline												
Fine-tuned BERT	0.03	0.09	0.04	0.03	0.07	0.07	0.17	0.08	0.07	0.02	0.01	0.04
Fine-tuned RoBERTa	0.02	0.06	0.02	0.01	0.05	0.07	0.10	0.09	0.07	0.01	0.01	0.02
SOTA (from NLP4IF'19) [253]	0.24	0.21	0.09	.0	0.17	0.16	0.44	0.33	0.40	.0	0.01	0.13
Proposed Architecture												
Fine-tuned BERT + CRF (5 epochs)	0.13	0.27	.0	0.04	0.08	0.20	0.59	0.26	0.28	**0.08**	0.01	0.10
Fine-tuned BERT + CRF (15 epochs)	0.11	0.25	0.02	0.04	0.07	0.28	**0.61**	0.25	0.22	0.04	**0.04**	0.13
Fine-tuned RoBERTa + CRF (5 epochs)	0.16	0.32	.0	**0.09**	0.11	0.35	0.37	**0.42**	0.37	.0	.0	0.06
Fine-tuned RoBERTa + CRF (7 epochs)	0.14	**0.40**	**0.23**	0.08	0.13	**0.37**	0.46	0.37	0.31	0.05	.0	0.17
Fine-tuned RoBERTa + CRF (10 epochs)	0.15	0.30	0.19	**0.09**	0.13	0.31	0.53	0.35	0.29	.0	0.01	0.25
Fine-tuned RoBERTa + CRF (12 epochs)	0.15	0.31	0.17	0.05	**0.19**	0.31	0.47	0.33	0.32	.0	0.03	0.14
Fine-tuned RoBERTa + CRF (15 epochs)	0.16	0.35	0.16	0.03	0.16	0.35	0.49	0.33	0.27	.0	0.01	0.24
Fine-tuned RoBERTa + CRF (5 epochs 3x-Oversampled)	0.15	0.34	0.14	0.07	0.13	0.30	0.52	0.34	0.27	0.05	0.02	**0.33**

Table 33. Experimental results on fragment-level classification on NLP4IF'19 test set. All scores are reported in micro-F1 (as in the original challenge). Scores in bold are the ones outperforming SOTA model.

[13] huggingface.co/transformers/

Models	SemEval'20 T11														
	Average	Appeal_to_Authority	Appeal_to_fear-prejudice	Bandwagon,Reductio_ad_hit.	Black-White-Fallacy	Casual-Oversimplification	Doubt	Exaggeration,Minimisation	Flag-Waving	Loaded_Language	Name_Calling,Labeling	Repetition	Slogans	Thought-terminating_Cliches	Whatab.,Straw_Men,Red_Her.
SOTA (from SemEval'20 T11) [254]	0.64	0.48	0.47	0.08	0.51	0.23	0.56	0.37	0.70	0.78	0.76	0.59	0.59	0.39	0.28
Proposed Architecture + Proposed argumentation features															
Fine-tuned RoBERTa (3 epochs)	0.53	0.08	0.34	0.14	0.17	0.06	0.52	0.32	0.61	0.72	0.68	0.22	0.12	**0.42**	.0
Fine-tuned RoBERTa (5 epochs)	0.53	0.14	0.34	0.17	0.26	0.09	0.46	0.35	0.60	0.73	0.72	0.17	0.36	0.30	0.18
Fine-tuned RoBERTa (10 epochs)	0.51	0.18	0.33	0.13	0.37	0.22	0.37	0.33	0.58	0.73	0.68	0.17	0.34	0.17	0.23
Fine-tuned RoBERTa (15 epochs)	0.51	0.14	0.29	0.12	0.31	0.14	0.42	0.35	0.55	0.73	0.69	0.13	0.35	0.25	0.21
Proposed Architecture + All proposed features															
Fine-tuned RoBERTa (3 epochs)	0.54	0.16	0.38	**0.20**	0.29	0.18	0.50	0.33	0.60	0.72	0.65	0.23	0.29	0.32	0.09
Fine-tuned RoBERTa (5 epochs)	0.52	0.09	0.35	0.13	0.31	0.21	0.43	0.34	0.61	0.74	0.70	0.21	0.23	0.33	0.12
Fine-tuned RoBERTa (10 epochs)	0.51	0.09	0.31	0.17	0.37	0.28	0.36	0.35	0.54	0.73	0.70	0.19	0.38	0.14	0.19
Fine-tuned RoBERTa (15 epochs)	0.51	0.15	0.32	0.07	0.40	**0.29**	0.37	0.31	0.54	0.75	0.66	0.18	0.43	0.14	0.12

Table 34. Results on span classification on SemEval'20 T11 test set (micro-F1). Scores in bold are the boost of F1 metric from proposed architectures compared to the SOTA model.

Table 33 reports on the obtained performances. Evaluation is reported as the average of micro-F1 scores of 5 run-times. Our proposed architectures, which utilize transformers with CRF output layers, outperform the state-of-the-art (SOTA) model for several propaganda techniques.

Fragment-level Classification on SemEval'20 T11 Dataset In the SemEval'20 T11 dataset, 14 propaganda techniques are annotated. Our focus lies in the Technique-Classification task (TC), which we approach as a sentence-span classification problem. To make predictions, we combine the logits of tokenized elements from both the sentence and the span. Moreover, we add semantic and argumentation features to enhance the performance. The following models are used:

1. Baseline. We utilize BERT or RoBERTa and implement a joint loss function ($loss_{\text{sentence}}$), combining individual loss functions per sentence ($loss_{\text{sentence}}$) and per span ($loss_{\text{span}}$). We define $loss_{\text{joint\_loss}}$ as:

$$loss_{\text{joint\_loss}} = \alpha \times \frac{(loss_{\text{sentence}} + loss_{\text{span}})}{N_{\text{loss}}}$$

   where N_{loss} stands for a number of $loss$ elements that are taken into the model (two, in this case).

2. State-of-the-art Model. This corresponds to the winning team [254] in the SemEval'20 T11 challenge.

3. Proposed Architecture. Our model combines a transformer architecture with a Bidirectional Long Short-Term Memory (BiLSTM). In addition to textual input, we feed the model with semantic and argumentation features. We apply the following joint loss function:

$$loss_{\text{joint\_loss}} = \alpha \times \frac{(loss_{\text{sentence}} + loss_{\text{span}} + loss_{\text{proposed\_features}})}{N_{\text{loss}}}$$

Table 34 presents the results obtained from 5 runs, measured as micro-F1 scores. Scores in bold indicate significant improvements compared to the state-of-the-art (SOTA) model. Notably, RoBERTa with argumentation features outperforms the SOTA model for the "Thought-terminating_Cliches" category. Moreover, by using all semantic and argumentation features together, we observe some improvements over "Bandwagon, Reductio_ad_hitlerum" and "Casual-Oversimplification". In general, we noticed that using different training epochs help to detect different propaganda techniques. We observe that some techniques tend to be learnt best at low training epochs (*i.e.*, "Bandwagon,Reductio_ad_hitlerum," "Thought-terminating_Cliches"), some at high training epochs (*i.e.*, "Casual-Oversimplification").

6.4.2 Conclusions

We proposed a new neural architecture combined with state-of-the-art language models and a rich set of linguistic features for the detection of propaganda messages in text, and their further classification along with standard propaganda techniques. Despite the boost in accuracy we achieved on two standard benchmarks for propaganda detection and classification, this task remains challenging, in particular regarding the fine-grained classification of the different propaganda classes. The state-of-the-art results on this subtask require further improvement to actually embed these solutions in real-world systems.

6.4.3 Relevant publications

- The contributing partners are currently working on a paper describing this contribution.

6.4.4 Relevant software, datasets and other resources

- No additional resources published yet.

6.4.5 Relevance to AI4media use cases and media industry applications

This research is relevant in UC2 "AI for News" since it offers valuable insights into the detection and classification of propaganda in news articles, alongside linguistic analysis of the features defining such propaganda messages. This can aid in enhancing the accuracy and credibility of news reporting.

Furthermore, it contributes to UC4 "AI for Social Sciences and Humanities" and UC1 "AI for Social Media and Against Disinformation" by addressing the need to identify misleading and potentially harmful content. Being able to discern such propaganda helps to prevent the spread of misinformation and mitigate its detrimental consequences for society.

6.5 Combining Objective and Subjective Perspectives for Political News Understanding

Contributing partner: CEA

Political news provides essential information, the interpretation of which shapes our understanding of political actions and events. Comprehensive analysis of the vast amount of political news available online is only possible by automating the process. Existing news analysis tools mostly characterize content through objective metrics. For instance, an analysis of the citation graph enables the clustering of news outlets based on their political affinity [255], an aggregation of TV and radio broadcast metadata to estimate political biases [256], an analysis of Facebook data associated with political parties enables an estimation of political polarization [257], and gender detection in spoken content leads to the quantification of speaking time for politicians of each gender [258]. These studies provide an objective characterization of news, but do not account for the subjective perspective of the author or of the news outlet which could be captured by an opinion-oriented analysis. This situation is in part a consequence of the limited availability of target-dependent sentiment classification (TSC) resources for the political domain [259], that can enable such analysis.

Our main contribution is a news analysis framework that integrates recent NLP components, information retrieval techniques, and external knowledge bases to obtain a rich text representation of news articles that combine objective and subjective perspectives. In a nutshell, the pipeline consists of: (1) named entity detection, (2) named entity linking, (3) named entity characterization via knowledge bases, (4) target-dependent sentiment classification, and (5) topic detection. It notably includes mentioned entities, sentiment associated with each entity, entity demographics, and political topic(s). Metadata such as the outlet name and the publication dates are also available. The combination of these dimensions enables the proposal of novel and rich insights about news outlets, entities and demographic segments. The analysis framework is instantiated for political news, using a corpus collected from French media. The main findings are the following:

- mainstream political orientations are presented in a rather balanced way in the major news outlets, but the radical left and right [260] are positively and negatively biased, respectively;
- the mentions and sentiments associated with political orientations vary across topics;
- sentiment scores are generally negative, with important variation between news outlets;
- the most positive and negative sentiment scores for individual politicians are well correlated with the public perception of their actions;
- mentions are biased toward male politicians, but sentiment scores of female politicians are higher;
- there is an age bias toward older politicians;
- the French semi-presidential system is reflected in the news, with a dominance of politicians who hold, held or were presidential candidates.

Some of these findings are illustrated below and the remaining ones are available in the associated publication, which is currently under review. These findings can be of great use for multiple stakeholders. News outlets can analyze their positioning and make it more transparent to the general public. Social scientists can gain new insights into the online representation of the political landscape. Political parties can monitor the online reporting of their political actions. The analysis can be run for other countries since the NLP components are multilingual and knowledge bases are international.

6.5.1 Findings

6.5.1.1 Mapping of Political Orientations We analyze the positioning of news outlets with respect to major political orientations by aggregating mentions and sentiment scores. The mention-oriented analysis is similar to existing ones [256], while the sentiment-oriented analysis adds a subjective perspective. We use the five-points scale of political orientations: far left, center left, center, center right, far right [260].

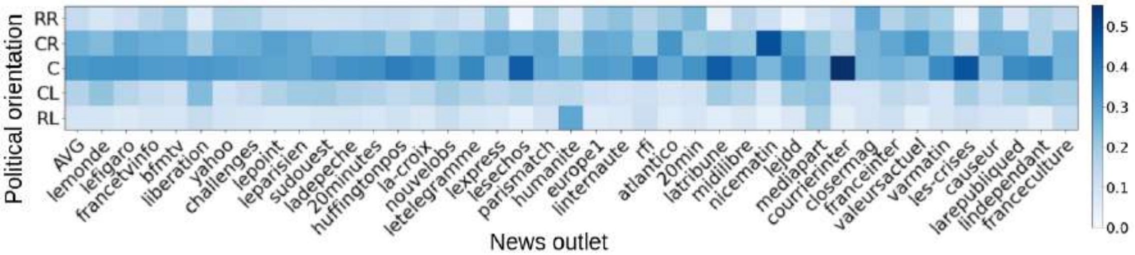


Figure 45. Distribution of mentions of political orientations in news outlets.

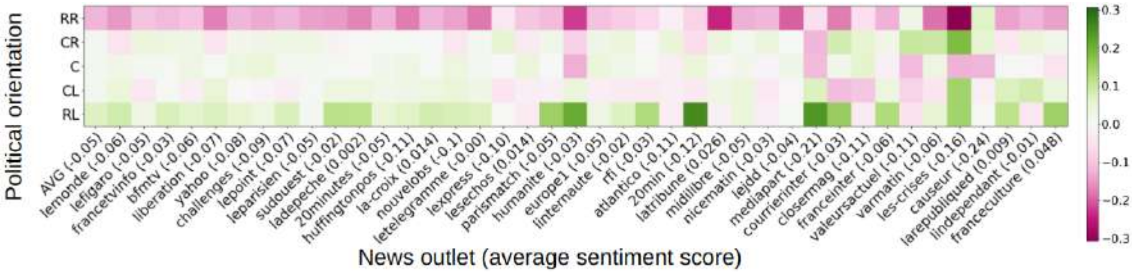


Figure 46. Deviation of the sentiment associated with major political orientations from the average sentiment of each source. Average sentiment of linked politicians is indicated in parentheses for each source.

Figures 45 and 46 summarize the mappings of mentions and of sentiment scores *versus* political orientations. The comparison of the two figures indicates that there is no correlation between mentions and sentiment since the Pearson correlation coefficient between the AVG columns in the two figures is 0.06. This result shows that the two types of analyses are complementary and provide additional insights into political news. Right-leaning orientations are more cited than their left-leaning counterparts but the sentiment associated with right-leaning orientations is more negative. This contrast is even clearer for RR and RL, the two radical orientations that are represented in the two figures.

Figure 45 shows that the center, including the French governing party, is the most mentioned orientation. The right-wing tendencies are more represented than their left-wing counterparts. Outlet-wise, the center is overrepresented in economic newspapers (*latribune* and *lesechos*), but also in *les-crises*, a right-wing anti-establishment outlet. RL is strongly present in *humanite*, the newspaper of the French Communist Party, while the extreme right is often cited by sources such as *closermag*, a tabloid, or *causeur*, a right-wing outlet.

Figure 46 illustrates the sentiment-oriented positioning of news outlets. It is quantified by the difference between the sentiment score per political orientation and the average sentiment score of the source. The positioning of sources toward political orientations varies, and this is a positive finding since diversified opinions are required for a healthy democratic debate. The representation of mainstream orientations (CL, C, CR) is rather balanced in the major sources (left of the figure) and more variable in other sources (right of the figure). We also note that, overall, there is a positive bias toward radical left (RL) politicians and a negative bias toward radical right (RR) ones. Notable exceptions for RL include *valeursactuelles*, *closermag* and *lindependant*. The only two sources which have a slightly positive positioning toward RR relative to their average positioning are *valeursactuelles* and *causeur*, but the average sentiment scores of RR remain overall negative even for these two sources. Intuitively, the center, which includes the current ruling party, is most criticized by news outlets which are left-wing (*humanite* and *mediapart*) and right-wing (*valeursactuelles* and *causeur*).

6.5.1.2 Topic Oriented Analysis The results for mentions and sentiments associated with political topics are presented in Figures 47 and 48. The mentions of centrist orientation, which includes the governing party, dominate most topics. This is particularly the case for the health effects of Covid-19 and for the war in Ukraine, two topics for which the government's communication is prevalent in the public space. Corruption is one notable exception, with the center-right being most cited because several of its politicians were involved in major corruption scandals. There are some differences even for closely related topics, such as the pairs related to Covid-19 and Ukraine. There, the mentions of the center are more dominant for Covid Health and for Ukraine War.

The deviation of sentiment associated with political orientations from the average of the topic, illustrated in Figure 48, can be interpreted as a proxy for the credibility of political orientations

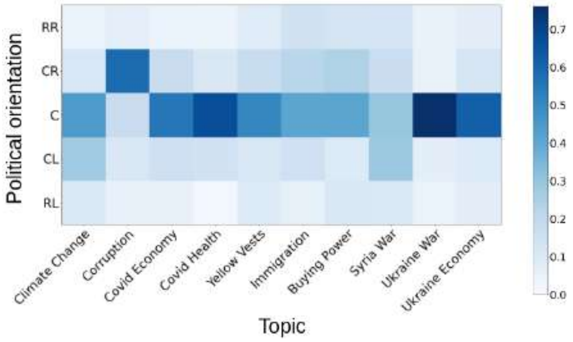


Figure 47. Distribution of mentions of political orientations for ten impactful political topics.

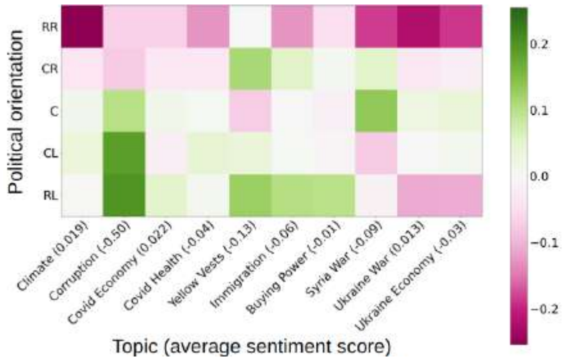


Figure 48. Deviation of the sentiment for major political orientations from the average sentiment of each topic. Average topic sentiment is given in parentheses.

for the respective topic. For instance, corruption has a particularly negative representation, with an average sentiment of -0.5. While still in the negative range, RL, CL and C orientations are perceived better than CR and RR on this crucial political topic. The radical right has the lowest scores on average, with particularly negative positioning for climate change and the war in Ukraine. This finding is probably explained by the relatively small importance given to environment in RR platforms and by the favorable positioning with respect to Russia in the past for the war in Ukraine. The radical left is also perceived negatively on the two Ukraine related topics for a similar reason. However, the sentiment toward RL is more positive than that of other orientations for most other topics.

6.5.1.3 Gender Representation Existing studies which quantify the representation of genders in politics [261] show that men are much more present than women. We deepen this analysis by adding a subjective component to understand how female and male politicians are represented in the news [14]. We present statistics about the mentions of female and male politicians (Figure 49), as well as the mean sentiment classification scores (Figure 50). The results from Figure 49 show that the news from the analyzed corpus contain significantly more mentions of male than of female politicians. The percentage of female mentions varies from 16.7% (*lemonde*) to 22.3% (*leparisien*). Considering that the overall proportion of women in French politics is higher (38% in Parliament[15] and 50% in the Government), the results from Figure 49 show a clear underrepresen-

[14] The corpus includes only 8 mentions of other gender.

[15] https://data.ipu.org/content/france

tation of women. This bias is higher than the one observed for French audio-visual media, where female politicians amount for 30% of the total mentions. The difference is probably explained by the stronger regulatory pressure on audio-visual media compared to written media. Gender bias is also stable through time. This finding indicates that the debate about the representation of gender in society has only a limited effect on the quantitative representation of women in politics.

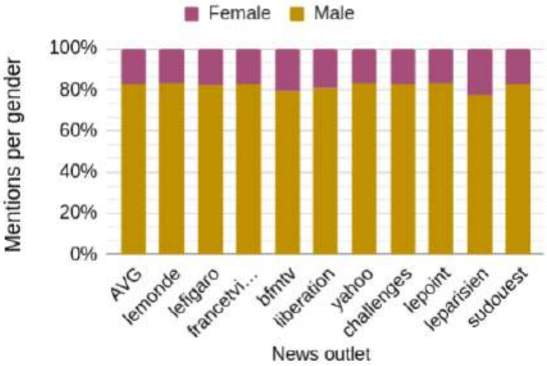


Figure 49. Percentage of gender mentions in the top-10 news outlets and on average.

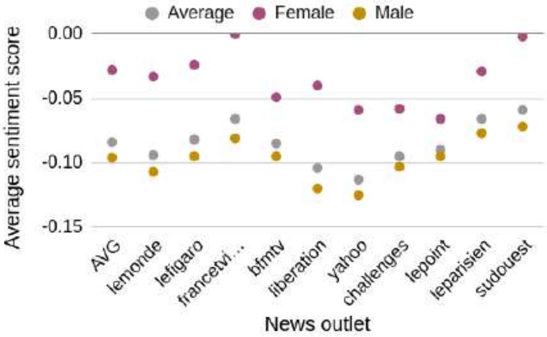


Figure 50. Mean sentiment classification scores by gender in the top-10 news outlets and on average.

The gender-focused analysis of sentiment from Figure 50 shows that the mentions of female politicians are more favorable than those of their male counterparts. This trend is respected for individual outlets, with some variation of the gap between the two genders. This favorable coverage is a positive signal for reducing gender bias in politics, but more efforts are needed to reduce the mention gap.

6.5.2 Conclusions

We introduced an application of NLP to the analysis of political news. The proposed framework combines recent NLP components, information retrieval and structured information to provide rich insights about the representation of politics in the news. The NLP components are multilingual and the resources ensure international coverage. The approach could thus be easily transferred to other countries and other languages in order to understand the similarities and differences of political representations in different democratic countries. We provide a comprehensive analysis focused on sources, politicians, and the representation of demographic segments.

We discuss below a series of limitations of the proposed analysis framework:

- The collected dataset includes a diversity of French media which are available online, but is focused on classical media. While the dataset content might induce biases in the analysis, we carefully designed experiments so as to have sufficiently large samples in each case.
- The sources are represented by a variable number of news and of mentions of entities. The outlet-related imbalance is mitigated by the fact that the analyses are performed on aggregated samples of data with sufficient data point for each sample.
- Sentiment classification leads to stable aggregated results, despite imperfections for individual cases. This was verified by ablating the analyzed dataset and studying the stability of results.
- Potential for misuse is associated to any AI technique which has societal impact, and it cannot be prevented in advance. Notably, analyses such as the one presented here should be presented as objectively as possible, and authors should restrain from conveying their own political opinions.

6.5.3 Relevant publications

- E. Dufraisse, A. Popescu, J. Tourille, A. Brun, O. Hamon, "Combining Objective and Subjective Perspectives for Political News Understanding", 2023, *under review* [229].

6.5.4 Relevant software, datasets and other resources

- No additional resources published yet.

6.5.5 Relevance to AI4media use cases and media industry applications

This research is currently being integrated in UC2 "AI for News", in order to give insights to journalists about the representation of political tendencies and of genders in political news. It can also be of use in UC4 "AI four Social Sciences and Humanities" since it offers support for an in-depth analysis of news by political scientists and communication experts. Initial ideas about integration in UC4 were discussed during a speculative design workshop held in June 2023 in Amsterdam. The analysis framework can be of use to media researchers and social scientists to understand the representation of politics in the media. After a domain-adaptation of resources, it can be used to analyze the representation of other fields in the media, including economics, health or sports.

6.6 Predicting Political Debate in a Complex International Organisation

Contributing partner: UvA

In this work, we devise a machine-learning protocol to tackle the complex task of investigating debate leaders in a multi-national organisation: the Intergovernmental Panel on Climate Change (IPCC).

The IPCC, a prominent international body, is entrusted with assessing scientific information related to climate change and its impacts. Comprising experts from various countries, the IPCC operates at the crossroads of scientific research and policy formulation. Its leadership structure includes the Bureau, a high-ranking committee responsible for overseeing the organization's work and steering its activities. This research employs anomaly-detection techniques to identify potential candidates for the Bureau, thereby constructing an intricate yet implicit model of leadership based on the behaviors and attributes of existing formal leaders.

The difficulty of this task lies in the impossibility to spell out the characteristics that define debate leadership in a complex and highly distributed organization, endowed with a hybrid mission at the interface between science and politics. To bypass this difficulty, we start from a sample of formal organisational leaders defined by the fact of having been officially nominated for the Bureau of the IPCC – among the highest positions in the organisation. We use a series of anomaly-detection techniques to identify IPCC contributors that are or might be Bureau candidates. We find that we can construct a precise albeit implicit model of IPCC leadership despite its social and political complexity. We then suggest various explainable AI methods to investigate why the model has selected members of the IPCC as Bureau candidates. Our analysis of the AI model and of its errors suggests interesting findings about asymmetries in the data and in the IPCC as well as shortcomings of the techniques we employed.

The insights gleaned from this analysis transcend the confines of the IPCC and can be applied to the broader context of "AI for Healthier Political Debate." This knowledge can be harnessed to design AI-driven solutions that promote constructive discussions, enhance decision-making processes, and mitigate potential biases or shortcomings within various organizations grappling with the intersection of science, policy, and politics. As societies tackle diverse challenges, the lessons drawn from this research can inform strategies for fostering more informed, inclusive, and fruitful dialogues within multifaceted international bodies and analogous institutions.

6.6.1 Background, Data and Methods

While leadership is a classic topic in political research, most work tends to focus on actual leaders, i.e. individuals who hold specific chairs in international organisations. While we cannot formulate an explicit definition of IPCC leadership, we know from our own previous fieldwork that the Bureau nominees are generally chosen among individuals who are recognised and respected within the organization, or by the national bureaucracies that nominate them. We can thus reasonably assume that Bureau candidates possess features connected to IPCC leadership and its mandate, even though these features cannot be easily made explicit. We use machine learning to capture these latent features and explainable AI to investigate them.

To create our model, we first hand-crafted several features that we expect to support its predictions. For featurization, we draw on a database of IPCC authors and delegates that we started to collect in two previous research projects and extended and updated since. The database contains the names of all individuals who have contributed to the first five assessment reports (ARs) of the IPCC. Great effort was invested to disambiguate homonyms and merge different names of the same person, but errors may remain. In total, we have counted 5,676 individual contributors to the IPCC. Our database separates the different roles held by the same individual, thus containing about 18,000 rows, each corresponding to the contribution by a specific individual in a given role (delegate, Bureau member, Coordinating Lead Author, Lead Author, Review Editor, Contributing Author). For each of the contributions, we also collected the national affiliation as declared by the contributor.

The IPCC proved to be a perfect example of how complicated international leadership can be and that a black-box model, contrary to intuition, could help us unravel this complexity. The situation in which we carried out our experiment, we believe, is not exceptional. It is common for social scientists to be interested in groups with unclear boundaries (e.g., the leaders of an international organisation) that they can identify by some prototypical examples (e.g., the official leaders of the organisation) rather than by an exact definition. Despite the complexity and blurriness of our target group, we succeeded in training an effective model, combining autoencoders and isolation forests, and in developing a non-trivial measure of leadership. Figure 51 showcases that our models can discriminate leaders, potential leaders, which are those IPPC members that we expect to become leaders, and others.

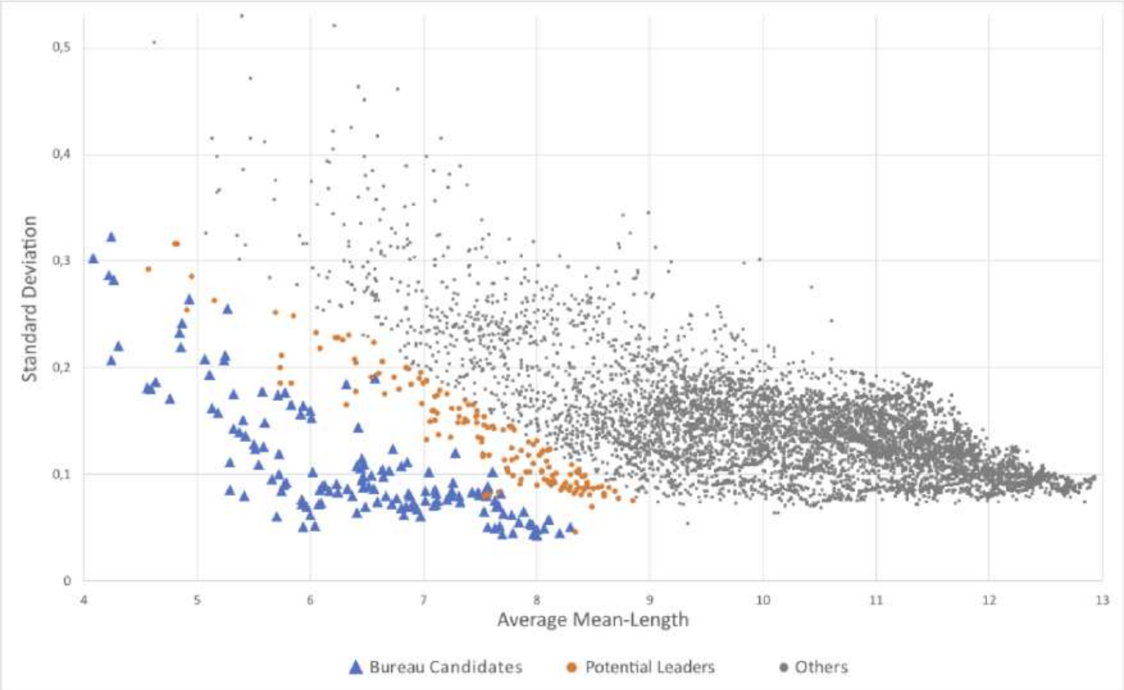


Figure 51. Average of and standard deviation for the MeanLength of the individuals of our corpus across the 10 implementations of our protocol.

6.6.2 Experiments

To make our protocol relevant for research into the development of political debates, we repurposed our model shifting its use from prediction to the description of political leadership relations in a complex international organisation. We developed a leadership-score as a continuous metric based on the outputs of our models and capable of capturing different degrees of exceptionality and to adapt to boundary cases. Repurposing predictive analytics techniques allowed us to describe the high-dimensional relationality of political leadership and to shed light on the different ways in which Global North and Global South countries, as well as from different regions of the World, contribute to the IPCC.

The paper also provides an extensive showcase of AI-explainability techniques for social and historical research. Using our leadership score and the classes of leaders and potential leaders, we could define surrogate models, for which we could derive the most important features. This analysis allowed us to separate the definition of leadership inherent to our model (which combines different forms of engagement with the organization) from the likelihood of being nominated to the IPCC Bureau (which is highly dependent on the recency of the latest contribution). We could also show how for particular individuals, decisions are made on the basis of the multifaceted interaction of several of our features, which a more straight-forward analysis would have missed. We could describe borderline cases that can be especially interesting to investigate. For instance, Anonymous1 looks from his website to be a good candidate for leadership as an author for special reports, reviewer for the greenhouse gas inventories guidelines and a committee representative for his country. But these leadership attributes are not reflected in our data. Through these borderline cases, the model offers hints how we could improve the database in the future.

Through the combination of black-box modelling, explainable AI techniques and our expert knowledge of the IPCC, we have highlighted several shortcomings of machine learning. We have shown, in particular, that AI techniques are highly dependent on the data they are based upon and could end up not only transmitting, but also amplifying their biases. It is also challenging for AI techniques to explain phenomena that are extremely complex and influenced by contextual and contingent factors. This is not a reason to abandon these techniques, but an exhortation to make database biases also an object of research. Reflecting on the mistakes of black-box modelling can reveal asymmetries present in the data, such as the one between developed and developing countries, as well as the leanings of the techniques employed, such as the tendency of anomaly detection techniques to identify all outliers.

6.6.3 Relevant publications

- T. Blanke, et al. "A Peek Inside Two Black Boxes", 2023, *under review* [262].

6.6.4 Relevant software, datasets and other resources

- No additional resources published yet.

6.6.5 Relevance to AI4media use cases and media industry applications

This research speaks directly to UC4 "AI for Social Sciences and Humanities" since it offers support for an in-depth analysis of political debates within a complex international organisation that is nevertheless typical and is supposed to be replicated in other contexts like AI and biotechnology. We offer a perspective that is genuinely driven by research interests in the social sciences and applicable in a number of related fields. It is especially relevant here as it also showcases some shortcomings of these techniques that provide valuable lessons to the community.

6.7 Understanding the Ephemerality of the Public Discourse during the COVID-19 Pandemic

Contributing partners: BSC and UvA

This work-in-progress study investigates the ephemeral nature of public discourse during the COVID-19 pandemic, with a focus on understanding the emergence and dissemination of various topics on Twitter. Although the data was collected using the Twitter Public API and the now-defunct Academic API, the study remains significant as it sheds light on the role of news media and online social media platforms, particularly Twitter, in shaping and disrupting public discourse during the crisis. The research adopts a two-method approach, utilizing computational methods to analyze the topics that emerged over time and a quantitative and qualitative analysis to identify prominent topics and their patterns of emergence.

6.7.1 Background, Data and Methods

Public discourse during crises plays a critical role in shaping collective attitudes and actions. Social media platforms, especially Twitter, have emerged as influential spaces for public discourse during crises periods, especially health and political ones. The ephemerality of topics on these platforms raises questions about the longevity and impact of discussions on crucial matters.

The COVID-19 pandemic sparked intense public discourse across social media platforms, particularly Twitter, serving as a significant tool for information dissemination and consumption. Understanding the ephemerality of topics during this period is crucial for grasping the dynamics of public opinion formation. This study aims to explore how certain topics emerged and disappeared on Twitter during the pandemic and to identify potential disruptors of topic emergence, specifically news media and discussion on online social media platforms. However, there is a scarcity of research on the factors that contribute to topic ephemerality, and this study seeks to address this gap.

In this context, we examine the use of temporal dynamic patterns as a measure of discussion health on online social media. Our aim is to understand public discussions based on their volume and the timing of contributions. As a result, we analyze the potential use of our ephemerality concept for labeling online public discourses based on how desirable, healthy and constructive they are. Additionally, we aim to understand the relation between ephemerality and information source. Initially, our approach lies in the identification of discussion types in an unsupervised manner. Subsequently, we use the concept of ephemerality to characterize these types, which we define more formally.

Our data consists of tweets related to the COVID-19 pandemic that were collected using the Twitter Public API during the pandemic's peak period. The Academic API, which was available at the time, was also utilized to augment the dataset after its availability in 2021. Although the Academic API is no longer operational, the data collected before its discontinuation remains a valuable source for this study.

Computational methods, including Natural Language Processing (NLP) and topic modeling, were then employed to identify emerging topics during different phases of the pandemic. The collected tweets underwent preprocessing to ensure data quality by removing duplicates, spam, and irrelevant content. To address variance and noise while minimizing the loss of precision, we implement a smoothing technique on vectors.

After gathering and preprocessing the raw tweets, we proceeded to extract sets of tweets related to specific topics. To identify COVID-19-related controversies, misinformation topics, and rumors, we referred to Poynter's database. For each of these topics, we manually composed a Solr query based on their description. The purpose was to retrieve relevant tweets using their most characteristic keywords for each specific topic. During our timeframe, we successfully characterized a total of 68 different misinformation topics. By analyzing patterns of engagement and dissemination over time, our method enables a comprehensive understanding of topic ephemerality.

From the computationally generated topics, a list of more prominent ones was selected for further analysis. This is followed by a quantitative and qualitative analysis of more prominent topics to identify their patterns of emergence and to uncover the role of news media and online social media platforms in shaping and disrupting public discourse. The quantitative analysis involved measuring the volume and velocity of tweets for each topic, offering insights into their emergence and dissemination. Simultaneously, the qualitative analysis delved into the content of these prominent topics, identifying potential disruptors of topic emergence, such as news media narratives or influential social media accounts.

6.7.2 Results

This study is still a work in progress from a Young Fellow Exchange; hence, the preliminary findings suggest that public discourse during the COVID-19 pandemic was highly dynamic and influenced by both news media and online social media platforms. The computational analysis revealed the emergence and dissemination patterns of various topics throughout the pandemic. Quantitative and qualitative analysis of the selected prominent topics identified key factors contributing to their ephemerality, including media narratives and the amplification of certain discussions by influential Twitter users. The research emphasizes the significance of comprehending the role of these entities in shaping public opinion during crises and how they generate public engagement that influences the ephemerality of certain topics. Additionally, our findings delve into the implications of ephemeral discourse for information dissemination and societal responses to pandemics, along with its broader impact on public perception.

6.7.3 Conclusions

In conclusion, this study provides valuable insights into the ephemerality of public discourse on Twitter during the COVID-19 pandemic. The research showcases the importance of employing computational and qualitative analysis to comprehend the emergence and dissemination of topics. By identifying factors that disrupt the initiation of topics, such as news media and influential social media accounts, this study contributes to a better understanding of how public opinion evolves during crises. As a result, this knowledge can aid in developing computational methods to identify ephemeral characteristics of certain topics and offer guidance to media on how to counter the impact of misinformation by avoiding its amplification.

6.7.4 Relevant publications

- No relevant publications published yet.

6.7.5 Relevant software, datasets and other resources

- No additional resources published yet.

6.7.6 Relevance to AI4media use cases and media industry applications

This research aligns perfectly with UC4 “AI for Social Sciences and Humanities” as it provides valuable insights for conducting in-depth analyses of political debates. Its relevance lies in shedding light on the opportunities to detect ephemeral contents, offering valuable lessons for the research community. By exploring the ephemerality of public discourse during the COVID-19 pandemic on Twitter and identifying factors that influence topic emergence and dissemination, this study contributes to the advancement of computational applications in social sciences and humanities, while also highlighting areas for improvement for media and further research in these fields.

7 Perception of hyper-local news (T6.5)

Contributing partners: IDIAP

Local news are sources of both day-to-day and critical information, and part of individuals' and communities' life. In Europe, local news are produced by hundreds of sources and many languages. This task is focused on the analysis of local news and the understanding of their perception both by people and machines. This section summarizes the main work and results for the current period. Section 7.1 describes outcomes of the work initiated in the previous period and partly presented in the previous WP6 deliverable (D6.1). Sections 7.2 - 7.5 describe work initiated in the current period. Please note that, whenever possible, the material in this section is taken and adapted from the published papers listed on each subsection.

7.1 Analysis of Health News from Newspapers and Online Videos

Contributing partner: IDIAP

7.1.1 Analysis of Europe's Press Coverage of Covid-19 Vaccination News

Understanding how newspapers present local news contributes to characterizing the full information ecosystem. In the previous deliverable (D6.1), we presented a dataset of over 50,000 online articles about Covid-19 vaccination, published by 19 newspapers from 5 European countries (France, Italy, Spain, Switzerland, and UK) over 22 months in 2020-2021. We performed NLP analyses on headlines and full articles, including named entity recognition, topic modeling, and sentiment analysis, to identify actors, subtopics, and sentiment across countries. Our analysis first showed a few consistencies across countries and subtopics, including the prevalence of neutral sentiment and relatively more negative sentiment for non-neutral articles, with few exceptions like the case of vaccine brands. The analysis also revealed differences, like a high negative-to-positive sentiment ratio for the no-vax subtopic. The dataset also provided research resources for new work, presented in sections 7.2 and 7.3.

7.1.1.1 Relevant publications

- D. Alonso del Barrio and D. Gatica-Perez, "How Did Europe's Press Cover Covid-19 Vaccination News? A Five-Country Analysis". In Proc. ACM International Workshop on Multimedia AI against Disinformation, Newark, Jun. 2022. [263].
  Zenodo record: https://zenodo.org/record/6779422.

7.1.1.2 Relevant software, datasets and other resources

- A European news dataset was produced for academic research purposes. However, the dataset cannot be distributed to third parties. Newspapers were contacted to request permission to distribute the data, but authorization was not obtained as of today.

7.1.1.3 Relevance to AI4media use cases and media industry applications This work is of relevance to use cases that analyze multiple information sources on the same topic. This initial work was used as the basis of the work in Section 7.2, where it was used for UC4 (AI for Social Sciences and Humanities).

7.1.2 Analysis of Health Online Video

Practices related to health circulate widely on YouTube. In the previous deliverable (D6.1), we presented a study to understand health and wellbeing-related practices of a group of popular, professional YouTubers from their audio-visual content. The analysis involved 2,500 YouTube videos, polytextual thematic analysis to identify six health-related categories, manual labeling of videos, and automated analysis of speech transcriptions and visual content. The analysis showed that distinctive patterns exist for the health-related categories under analysis, and that is possible to classify videos according to health-related categories in a binary setting (best accuracy obtained for linguistic features, ranging between 74-87%, see Table 35.) The work resulted in the publication listed below.

Binary video class	Number of videos	Accuracy (lingusitic features)
Nutrition	2120	87%
Self-development	1364	75%
Bodycare	1328	78%
Physical activity	878	78%
Companionship	370	77%
Rest	316	74%

Table 35. Video classification results according to category.

7.1.2.1 Relevant publications

- T.-T. Phan, C. Michoud, L. Volpato, M. del Rio Carral and D. Gatica-Perez, "Health Talk: Understanding Practices of Popular Professional YouTubers". In Proc. Int. Conf. on Mobile and Ubiquitous Multimedia, Lisbon, Nov. 2022. [264].
  Zenodo record: https://zenodo.org/record/8045932.

7.1.2.2 Relevant software, datasets and other resources

- A dataset of video transcriptions was produced for academic research purposes, but the terms of YouTube do not allow to make them publicly available.

7.1.2.3 Relevance to AI4media use cases and media industry applications Given the close collaboration between computing and psychology involved in this research, the work is potentially relevant for UC4 (AI for Social Sciences and Humanities.)

7.2 Frame analysis of European News

Contributing partner: IDIAP In recent years, there has been a surge in concepts like data-driven journalism, computational journalism, and computer-aided reporting, all aimed at integrating journalism with technological advancements. The progress in the use of Natural Language Processing (NLP) in journalism is driven by transformer-based machine learning models, which are being integrated into the media sector to perform various tasks, including generating headlines using language models, summarizing news articles, or detecting misinformation.

The advent of Large Language Models (LLM) like GPT-4, BLOOM, and ChatGPT represent a trend toward facilitating human-machine interaction. Consequently, this has opened up a broad spectrum of possibilities. However, these models also suffer from a lack of transparency. Ongoing efforts are being made to enhance transparency, analyzing use cases to gauge their utility and limitations.

The objective of this line of work is to evaluate the effectiveness of GPT-3.5 in a specific application, namely, the analysis of frames in news articles. Frame analysis is a journalistic concept that involves studying the presentation of news stories, and examining which aspects are emphasized. It aims to determine whether an article adopts an informative approach, imparts moral lessons, showcases an economic perspective, or emphasizes human emotions, among other frames.

For the task of automatic frame classification, pre-trained models with fine-tuning techniques had been used. However, generative models now provide the possibility of employing prompt engineering techniques for frame classification, thus offering an alternative to the previous fine-tuning approach.

7.2.1 Methods

We followed three steps, using a subset of the European news dataset described earlier in this section [263]. The first step is data annotation of headlines sourced from articles centered around the Covid-19 anti-vaccine movement. The second step is the use of LLM with the fine-tunning technique. The final step is the use of LLM with prompt engineering.

Annotation. Reviewing the existing literature on framing, we prepared a codebook containing definitions and examples of frame types, following the framework proposed by [265]. This included 5 generic frames (attribution of responsibility, human interest, conflict, morality, and economic consequences), along with a 'no-frame' category. A total of 1,786 headlines were manually labeled. Two researchers independently labeled the entire dataset, each annotating 893 headlines, ensuring comprehensive coverage and reducing errors.

Fine-tuning approach. Previous works related to frame classification in the literature have used fine-tuning, BERT-based models. In our work, we implemented this approach as baseline, but we aimed to go one step further and investigated the use of fine-tuning of GPT-3.5.

Prompt-engineering with GPT-3.5. In this approach, instead of adapting pre-trained language models (LMs) for specific tasks, the downstream tasks are restructured to resemble the tasks tackled during the original LM training. This is achieved with the help of textual prompts. For instance, when determining the emotion of a social media post like "I missed the bus today.", we may use a prompt like "*I felt so* _" and prompt the LM to fill the blank with an emotion-bearing word. Alternatively, by using a prompt like "*English: I missed the bus today. French:* _", the LM can fill the blank with a French translation. By carefully selecting these prompts, we can influence the model's behavior, enabling the pre-trained LM to predict the desired output without any additional task-specific training. Adapting this idea to the classification of frames at headline level, we define the prompt as including the definitions of the different types of frames, and the headline to classify. We then asked the model to choose one of the definitions that best fit that headline.

7.2.2 Experiments

For the first step of our methodology, we sought to determine the dominant frames used in news headlines across five European countries when discussing the anti-vaccine movement. Through the process of human annotation, we examined the 1,786 headlines from articles on Covid-19 anti-vaccine movement. The analysis revealed that the prevailing frame observed in these headlines was 'human interest.' This frame involved personalizing events through the inclusion of individual statements or specific incidents. Additionally, a substantial number of headlines were classified as presenting neutral information without a distinct frame. Notably, we found remarkable consistency in the distribution of frame types across all countries, with 'human interest' and the absence of frames emerging as the most prominent patterns, as shown in Fig. 52.

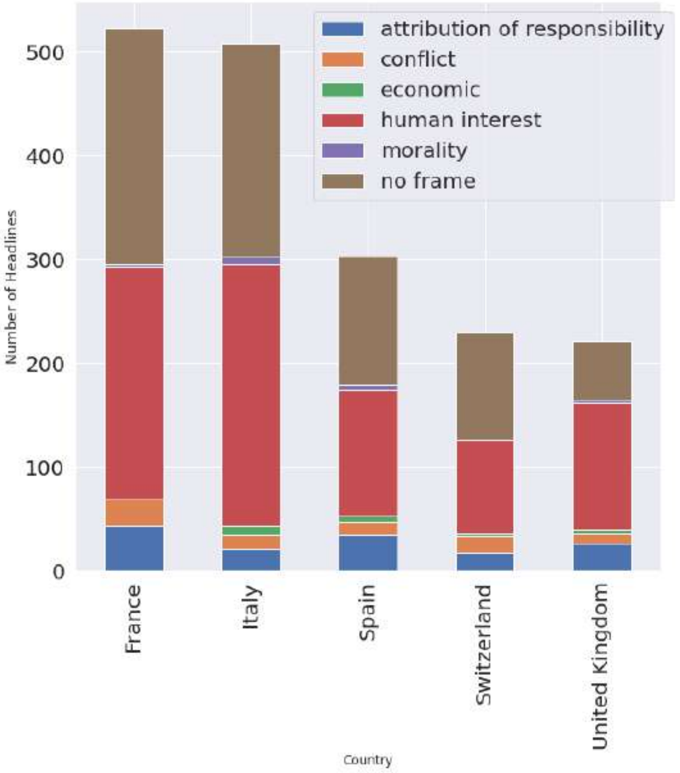


Figure 52. Non-normalized distribution of frames per country

For the classification steps, as mentioned before, we followed two approaches: a traditional fine-tuning approach and the prompt-engineering method. Our study aimed to investigate the feasibility of utilizing prompt engineering to classify headlines based on different frames. Prompt engineering offered an alternative approach that eliminated the need for labeled training data and instead relied on the capabilities of GPT-3.5. We conducted a thorough analysis of GPT-3.5's performance in frame classification. The findings indicated that the baseline based on fine-tuning two BERT models achieved an accuracy of 67% (BERT) and 70% (RoBERTa). In contrast, fine-tuning GPT-3.5 achieved an accuracy of 72%, thus surpassing the performance of the baseline methods. However, when employing prompt engineering, the accuracy dropped to 49%. It is worth noting that the subjective nature of human labeling played a significant role in influencing the achieved accuracy of the classification task. Because of this subjectivity, we did an additional experiment in which we asked the annotators if they agreed with the label given by GPT-3.5, and the agreement between humans and machine resulted in an increase up to 76% of the cases, which shows that the GPT-3.5 does not make random guesses, but tends to coincide with human annotation for this subjective task.

7.2.3 Conclusions

This work contributed in two ways. Firstly, we implemented a systematic human annotation process, enabling the identification of the dominant frames in European news headlines related to the anti-vaccine movement. This process revealed the prevalence of the “human interest” frame, as well as a notable number of headlines presenting neutral information without a specific frame. Secondly, we conducted an extensive performance analysis of GPT-3.5, comparing the traditional fine-tuning approach with the emerging prompt engineering method for frame classification. The outcomes of our work also shed light on the potential and limitations of these approaches, emphasizing the impact of human subjectivity on accuracy. More specifically, the intrinsic subjectivity of the annotation is something to deal with, either with multiple annotators, or by not choosing only one of the possible frames as valid, as we have observed cases in which more than one option makes sense in semantic terms.

7.2.4 Relevant publications

- D. Alonso del Barrio and D. Gatica-Perez : “Framing the News: From Human Perception to Large Language Model Inferences”. In Proc. ACM. International Conference on Mutimedia Retrieval (ICMR) 2023, Thessaloniki, Jun. 2023. [266].
  Zenodo record: https://zenodo.org/record/8045932.

7.2.5 Relevant software, datasets and other resources

- A dataset of labeled headlines with respect to frames, derived from the European news dataset, was produced. However, as mentioned in Section 7.1.1, we currently do not have authorization to make the articles publicly available.

7.2.6 Relevance to AI4Media use cases and media industry applications

Based on discussions with NISV (UC4: AI for Social Sciences and Humanities), the work on frame analysis was seen as valuable and timely, involving ML to study a concept from journalism – the identification of frames in a text. A concrete collaboration was undertaken to investigate the applicability of the frame analysis approach in the context of NISV content. This involved the selection of a dataset of NISV original content in Dutch; the manual labeling of frames according to the framing scheme investigated for European news; the application of the existing frame classification models to this dataset; the analysis of the quality of the results; and a further analysis related to its future applicability.

7.3 Analysis of No-Vax News in the European Press

Contributing partner: IDIAP

The anti-vaccine movement, fueled by the ease of spreading information, particularly through social media platforms - where rumors and conspiracy theories can contribute to vaccine hesitancy - has gained visibility [267] [268]. Both traditional and social media have played a pivotal role in informing the general public about vaccination, influencing public opinion by either promoting vaccination or amplifying doubts. Surveys conducted across different countries have consistently shown that individuals relying on traditional media like newspapers and television news are more inclined towards getting vaccinated, while those who primarily rely on social media as their primary information source exhibit more hesitancy towards vaccination [269] [270] [271].

Existing computing research has predominantly focused on identifying and analyzing false content, be it text, images, or videos circulated on social media platforms. However, comparatively less attention has been given to how reputable news outlets, which hold significant trust among the public, have addressed the issues of vaccine hesitancy, misinformation, and disinformation.

To bridge this research gap, our work studied how major European newspapers approached reporting on the movement against the Covid-19 vaccines, and how these media outlets addressed the issue of misinformation and disinformation surrounding the vaccines. By examining these aspects, we contributed to understanding the role of the press in shaping public opinion and countering vaccine-related disinformation.

7.3.1 Methods

In this work, we implemented several techniques, with the aim of extracting indicators of how the written press treated the issue of the anti-vaccine movement and disinformation, on the European news dataset described earlier [263]. The analysis included:

- **Subtopic modeling**. The goal was to identify the main sub-themes within the main No-Vax theme by using BERTopic [272].
- **Semantic similarity with word embeddings**. The objective was to identify relationships between terms to capture different aspects within no-vax and disinformation. We used Word2vec to encapsulate the words in vectors and then study their similarity and the relationship between terms.
- **Classification by country**. The aim was to study the similarity or differences among countries when dealing with a common topic. We used logistic regression, which allowed to see which words influenced the country classification of the articles, and also allowed to see what happened in the classification of each country.
- **Political orientation**. Based on the sentiment analysis of one of our previous studies [263], we examined the sentiment of the no-vax articles according to their political orientation.
- **Analysis of sentences related to disinformation**. The goal was to study in depth those words that contain the word disinformation, reveal what Named Entities appeared, and understand the relationship between them, in order to detect the main characters related to No-Vax and disinformation.

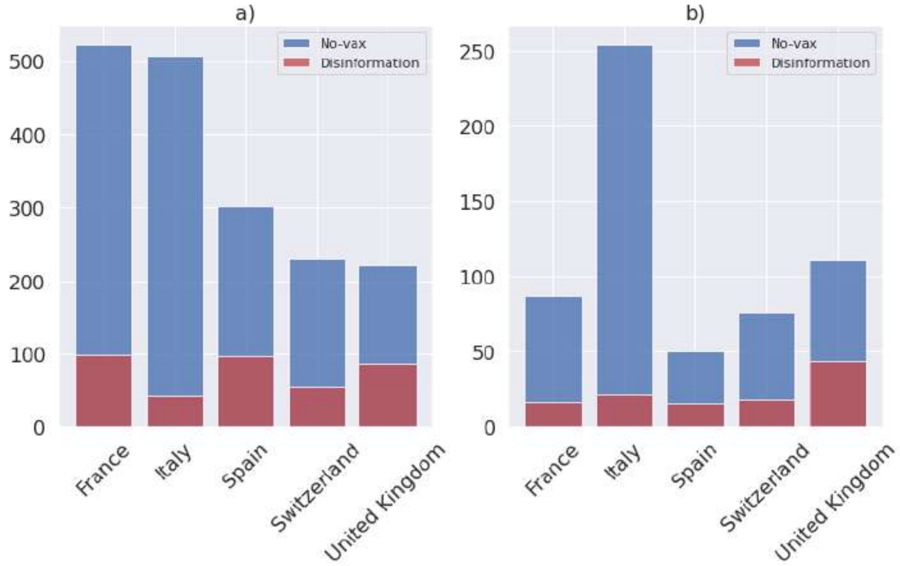


Figure 53. (a) Per-country number of articles about disinformation in the dataset of articles about no-Vax. (b) Per-country number of articles about disinformation in the dataset of articles about no-Vax normalized by the number of newspapers

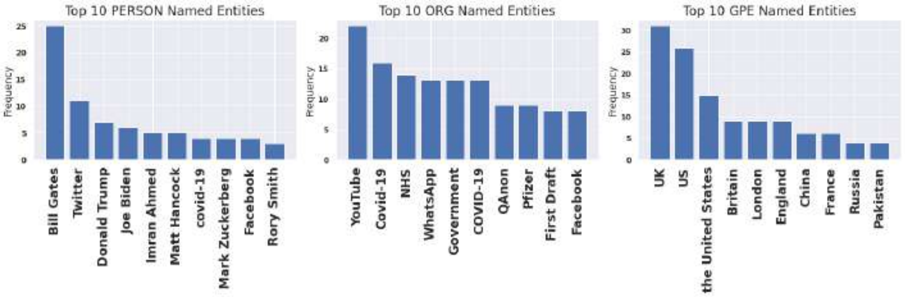


Figure 54. Most frequent named entities in disinformation sentences

7.3.2 Experiments

The above analyses provided insights into how major European newspapers presented news related to the movement against the Covid-19 vaccine. As a first illustration, the statistics of newspaper articles related to No-Vax and disinformation per country are shown in Fig. 53, which shows substantial differences in certain countries (e.g. Italy vs. Spain or Switzerland).

As a second example, we delved deeper about analyzing sentences that contained terms such as "disinformation" or related keywords in these articles. The results of the analysis is shown in Fig. 54, which shows the list of most frequent named entities in disinformation sentences. This reveals the main actors (persons, organizations, and places) involved in such articles, and clearly show the relevance of global actors (i.e., beyond Europe) in European news content.

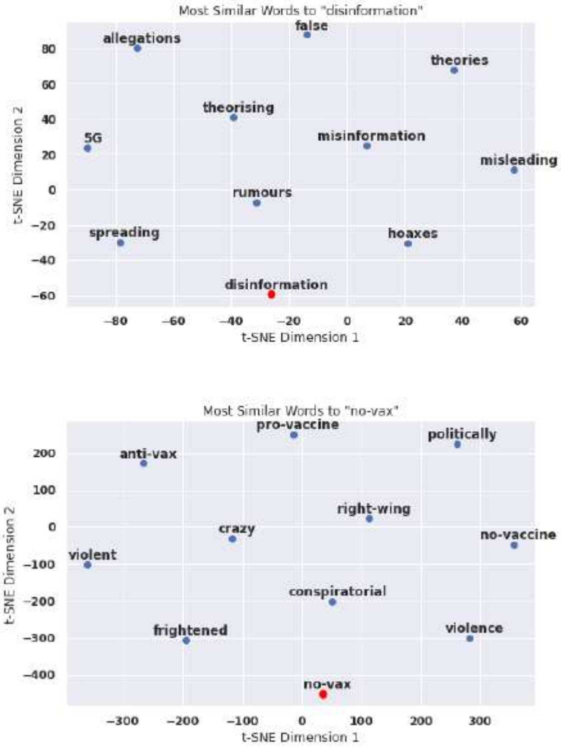


Figure 55. Word embeddings: no-vax and disinformation

As a third example of analysis, Fig. 55 shows the results of the word embedding for both no-vax and misinformation, which highlights the use of specific terms in the European press related to these themes. The full analysis can be found in the corresponding publication, cited at the end of this section.

7.3.3 Conclusions

Utilizing a dataset of European articles concerning the anti-vaccination movement, our initial step involved conducting a range of assessments, including subtopic modeling, investigating semantic connections via word embeddings, categorizing articles by their country of origin, and performing sentiment analysis grounded in political inclinations. To delve further into our primary examination, we explored and scrutinized sentences containing the term "disinformation" (or similar concepts) within these articles. This was complemented by applying various analytical methods such as named entity recognition, identifying keyword associations, sentiment analysis, and evaluating relationships between entities. This comprehensive analytical approach enabled to gain a better understanding of how the European press actively confronted and addressed the issue of Covid-19 vaccine-related disinformation associated with the anti-vaccine movement. As our analysis shows, the efforts of the European press played a significant role in providing information of quality towards countering the spread of disinformation and misinformation.

The main limitation of the work has to do with is access to more newspapers, with a larger coverage of the political spectrum. More specifically, there are some countries with only two newspapers in the dataset, while others have six. Expanding the data would provide a more complete description of the phenomena in each country.

7.3.4 Relevant publications

- D. Alonso del Barrio and D. Gatica-Perez : "Examining European Press Coverage of the Covid-19 No-Vax Movement: An NLP Framework". In Proc. ACM International Workshop on Multimedia AI against Disinformation, Thessaloniki, Jun. 2023. [273].
  Zenodo record: `https://zenodo.org/record/8045910`.

7.3.5 Relevant software, datasets and other resources

- This work used a subset of the European news dataset already described in Section 7.1.1.

7.3.6 Relevance to AI4Media use cases and media industry applications

This work is related to use case UC1: AI for Social Media and Against Disinformation, as we describe an important part of the ecosystem where disinformation is made visible and critiqued. This was the role played by some of the main European newspapers, regarding specifically the issue of the no-vax movement.

7.4 A Local Lens: Characterization of French-Speaking Swiss News

Contributing partner: IDIAP

In addition to examining trends at the European level for multiple countries, another research thread is the examination of news content at a more local scale. Over the past years, the distribution of newspapers has been challenging for the local press [274], notably in Europe [275]. The consumption of the physical form of the local news has declined, while people follow more closely this type of information [276]. In order to maintain and expand their audience and notoriety, the stakeholders in this sector have developed and implemented new strategies [277]. Also, there has been a resurgence of focus on local outlets in the wake of various events that have caused people to have doubts about global and national newspapers [278]. The local press, with its specific issues, has been subject to transformation, particularly through digitization. This research thread aimed to analyze the local media landscape in three French-speaking cantons of Switzerland, through the characterization of the content published by a local media outlet, ESH Medias, on their online platform [279]. Our work focuses on the computation of linguistic characteristics in the content, to extract 'local' features and to identify research questions to examine journalistic phenomena specific to the local press.

7.4.1 Experiments

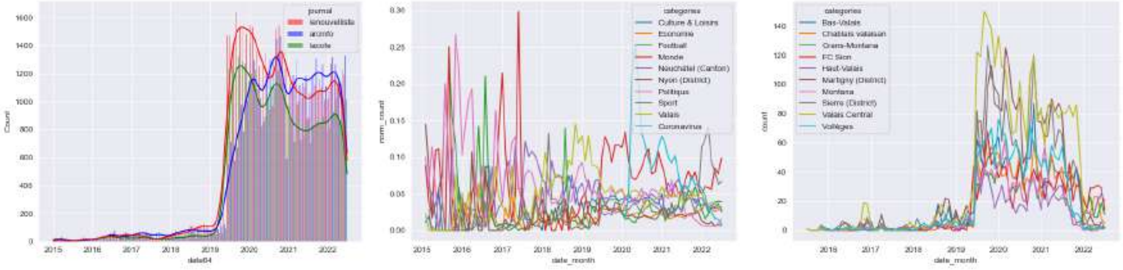
7.4.1.1 Data The data collected consists of online articles published in the three newspapers of the ESH Medias group, which have dedicated local audiences: Le Nouvelliste (in Valais canton), La Cote (in Vaud canton), and Arc Info (in Neuchatel canton). An agreement with the news company allowed for and facilitated the data collection process. The time window of data begins in January 2015 and ends on June 30, 2022. The number of articles collected is displayed in Table 36.

Articles published			
Le Nouvelliste	La Cote	Arc Info	Total
43 393	38 283	48 479	130 155
(21 581)	(12 618)	(22 830)	(83 243)

Table 36. Numbers of articles collected for each local newspaper. The numbers of unique articles are listed in parentheses. The final total of unique articles does not equal the sum of unique, because we kept the first occurrence of the duplicated.

For each article, we collected the title, the headline (or *chapeau* in French), and the content of articles as textual data. We have also added as metadata the authors, the date, the tags (defined by the authors), and the picture illustrating the article. In total, there are 56,744 unique articles with an illustration.

7.4.1.2 Descriptive Analysis For the initial part of this analysis, we evaluated the quality of the dataset to determine its relevance to the analysis and to understand the specifics of the articles. The time distribution in Figure 56a reveals a difference in the digitization of articles and the prevalence of published articles for each newspaper. We note a clear distinction between before and after June 2019. The main source of articles was Keystone-ATS, the national press agency of Switzerland, with more than 10,000 contributions over the years. We also observed journalists who wrote articles for the three newspapers. Taking this aspect into account may be useful to analyze potential differences between the 'local' features of the articles. Furthermore, self-reported topic tags were used as a pre-categorization for the articles, and we noticed several interesting elements. Although the metadata before June 2019 may be too sparse to work with, we still observe some temporal patterns that are representative of certain periods. One notable example connected to our previous work reported in this section is the Covid-19 pattern in the first half of 2020, with more than 25% of the articles having this tag in Figure 56b. The excluded tags, which were reported in the articles of one newspaper but not in the other, may also represent indicators of locality of the canton of distribution. As an illustration, we can see in Figure 56c references to FC Sion (the local Valais football club), as well as Martigny and Sierre (Valais cities), and Crans-Montana (a mountain resort.)



(a) temporal distribution of the number of articles published

(b) temporal distribution of the reported tags

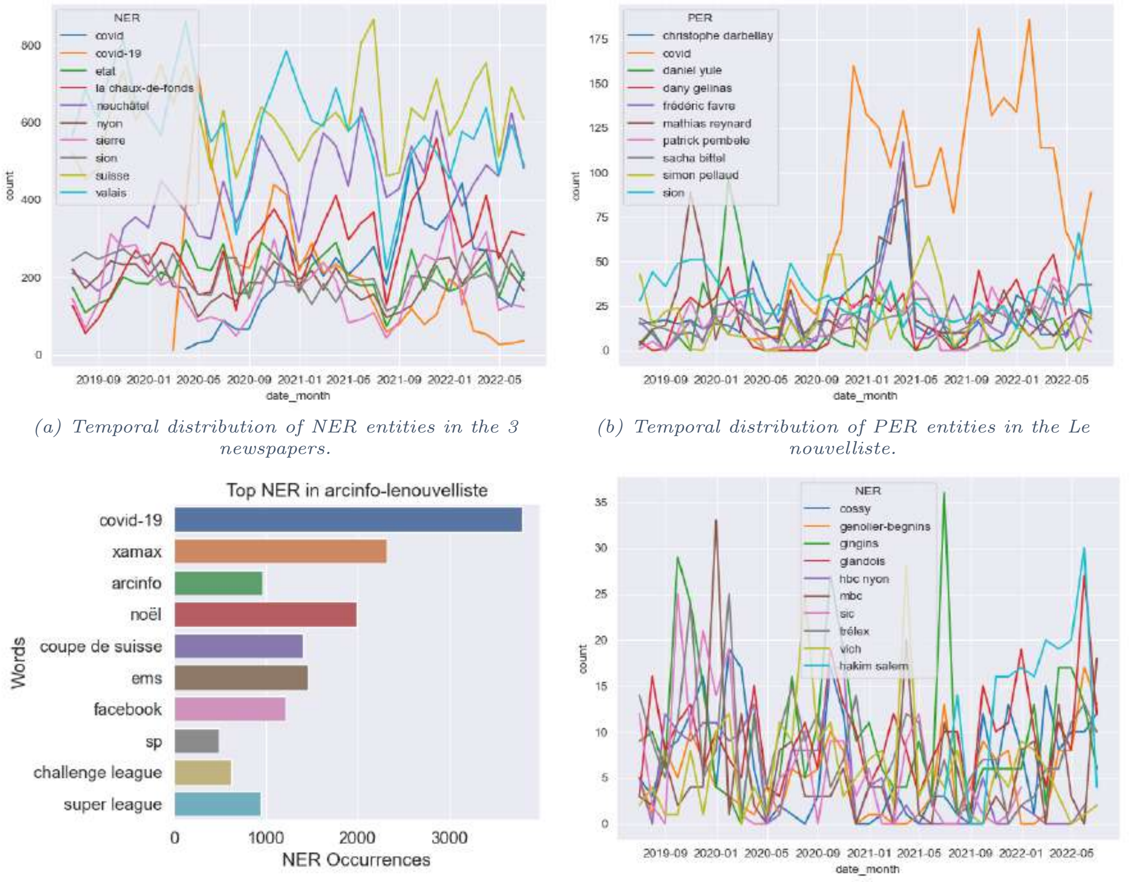
(c) temporal distribution of the tags reported only in Le Nouvelliste

Figure 56. Temporal distributions of the number of articles published in Arc Info, La Cote, and Le Nouvelliste, with their tags

7.4.1.3 Content Analysis After analyzing the metadata, we conducted a linguistic characterization of the articles published after June 2019 to extract the information conveyed by the text. To do this, we used the Named Entities Recognition (NER) approach of the SpaCy pipeline[280] for our French articles. This model retrieves entities according to 4 tags: *LOC* for location, *PERS* for persons, *ORG* for organization, and *MISC* for miscellaneous. The results in Figure 57a showed that the entities linked to location were predominant (excluding 'covid'), likely due to the nature of news reporting, which situates their article, and the interest of journalists and readers in this local dimension. The model also demonstrated its generalization performance by extracting very local and very specific entities, such as local personalities and national or local institutions. In Figure 57b, we note for PERS entities 'Christophe Darbellay' (member of the Executive Government of Valais) or 'Daniel Yule' (a Valais ski athlete.) To further analyze the results, we identified the entities specific to newspapers by looking at (1) the intersection of the entities of pairs or local newspapers, and (2) the entities of a single newspaper by excluding those of the others. The intersection approach shows more general entities, like “Noël” or “coupe de suisse” in Figure 57c. The exclusion approach revealed very local entities, which are in the distribution area of the newspaper, such as “Cossy” or “Genolier-Begnins” in Figure 57d.

The characterization of the articles was further studied by evaluating the lexical richness and readability of the texts, looking for notable differences between groups of articles. To conduct these measures, we used the lexical richness library [281] and the pyreadability library [282] to implement the evaluation of these metrics. As lexical richness metrics, we selected the *Moving Average Type-Token ratio*, the *Hypergeometric Distribution Diversity*, the Measure of Textual Lexical Diversity, and the Dugast's lexical diversity measure. For the Readability metrics, we selected *Flesch Reading Ease* (FRE), *Gunning-Fog Index* (GFI), *Coleman-Liau Index* (CLI), *Flesch-Kincaid Grade level* (F-K GL), and *Automated Readability Index* (ARI). These parameters of these metrics have been adapted to the French language. After evaluation, the lexical richness and readability measures were found to be consistent across the newspapers and measures. The readability measures were determined to be suitable for a pre-adult/adult audience, while still being accessible to the general public. In Table 37, the results evaluate the average readability to the 16-to-17 years old (GFI) or high-school level (FRE).

7.4.1.4 Unsupervised topic clustering In this part of the work, we conducted experiments to extract a self-organization of articles based on their topic. To do this, we used the BERTopic pipeline [283], with two embedding models: a multi-lingual BERT model with 128 tokens as input and 384 features as output [284], and a French BERT model (CAMEMBERT), with 512 tokens as input and 1024 as output [285]. We also used the HDBSCAN clustering algorithm, which was the



(a) Temporal distribution of NER entities in the 3 newspapers.

(b) Temporal distribution of PER entities in the Le nouvelliste.

(c) Occurrences of NER entities appearing in Arc Info and Le nouvelliste.

(d) Temporal distribution of NER entities in Le nouvelliste excluding entities in other newspapers.

Figure 57. Temporal distribution and bar graph of the NER entities in the articles of Le Nouvelliste, Arc Info, and La Cote.

Metrics	all	Arc Info	La Cote	Le Nouvelliste
FRE	54.84 (9.56)	54.84 (10.2)	54.60 (8.48)	54.96 (9.53)
GFI	13.76 (2.76)	13.60 (2.92)	13.94 (2.3)	13.82 (2.84)
CLI	10.07 (1.65)	10.24 (1.73)	9.961 (1.58)	9.975 (1.6)
F-K GL	11.37 (2.56)	11.25 (2.73)	11.53 (2.1)	11.38 (2.63)
ARI	11.95 (3.28)	11.87 (3.49)	12.09 (2.65)	11.94 (3.39)

Table 37. Results of the readability evaluation applied to the articles in each newspaper of the dataset

most suitable for our articles due to their sparse and varied nature. The two major results were (1) the inference of the French model with the entire articles as input (60 topics in results); and (2) the inference with the multilingual model with only the title and the headlines as input (84 topics). In Figure 58, we present the result of the hierarchy with the top 20 topics of the French model. Notably, the number of outliers with a minimum restriction of 50 neighbors was relatively large (around 20,000 articles). This may be explained by the nature of the data, with articles which can be relatively distinct according to the information reported, but also by the binarity of the classification, the articles are predicted for a single cluster in this implementation.

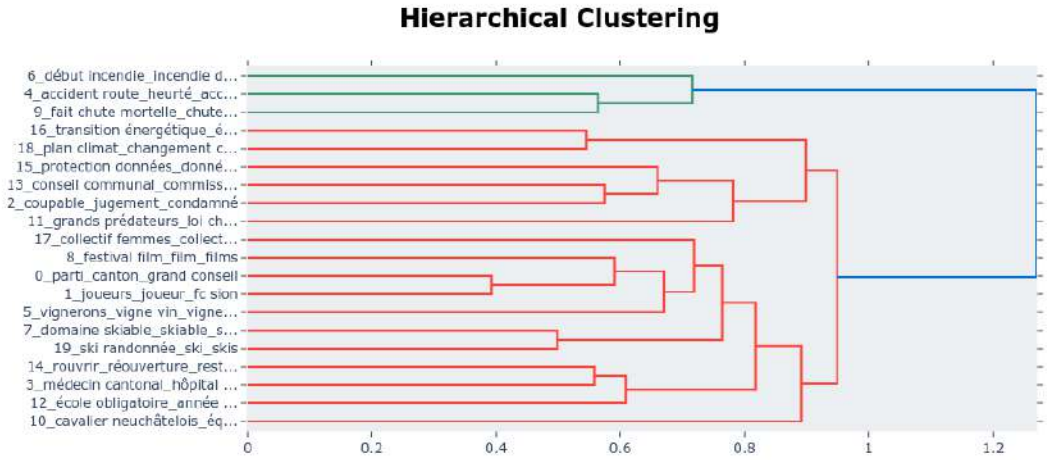


Figure 58. Hierarchy of the top 20 clusters of the French model applied to cluster the Swiss local dataset.

We used OCTIS[286] to measure the quality of the clustering, and the results of the two inferences are presented in Table 38. For topic diversity metrics, we selected the usual *Topic Diversity* with the ratio of words from the top representative words for a topic unseen in other topics, the *Inverted Ranked-Based Overlap* (Inverted RBO). For topic coherence, we used *U Mass coherence*, and *Normalized Pointwise Mutual Information* (NPMI). Finally, we used the *Kullback-Leibler divergence* (K-L) to quantify the significance of the topic clustering results. The results are shown in Table 38. These numerical results are relatively difficult to interpret regarding their specificity and the data we consider. For instance, the topic diversity measure only takes a partial representation of the topics, but they may be used as proxies to compare topic models. Further research could follow a comparative approach and investigate how these measures compare to other existing news datasets, such that certain baseline reference points can be established.

Metrics	French	Multi
Topic Diversity	0.8800	0.8663
Inverted RBO	0.9949	0.9973
Umass	-3.308	-8.056
NPMI	0.07857	-0.05649
K-L	2.289	2.964

Table 38. The results of the coherence, diversity, and significance metrics applied to the inferences of the two main clustering outputs.

7.4.2 Conclusions

This research provided a first analysis of the unique characteristics of local Swiss news data and the associated challenges. We proposed a framework for the systematic analysis of newspaper articles from multiple perspectives. Additionally, we were able to identify several features that are specific to local news, as well as those that differentiate the sources. Our work will be further developed by analyzing the particularities of the article clusters that were identified, and by refining and expanding the definitions of locality we proposed here.

7.4.3 Relevant publications

- A publication is under preparation.

7.4.4 Relevant software, datasets and other resources

- This study involves a dataset of news articles collected from the websites of three local newspapers, in agreement with the company who owns the newspapers. The dataset is currently private; we will investigate whether it is possible to make it publicly accessible in the future.

7.4.5 Relevance to AI4Media use cases and media industry applications

This study initiates an exploration of tools in NLP that align with the methodologies used in social sciences for news analysis. This is particularly relevant to UC4: AI for Social Sciences and Humanities. The research also focuses on the distinct features of local news, contributing to a deeper understanding of its added value compared to national or international news, including the specific type of stories covered and the framing of the articles. This line of work may influence how the media addresses local news topics and their biases by taking into consideration the studied characteristics.

7.5 Referencing in YouTube Knowledge Communication Videos

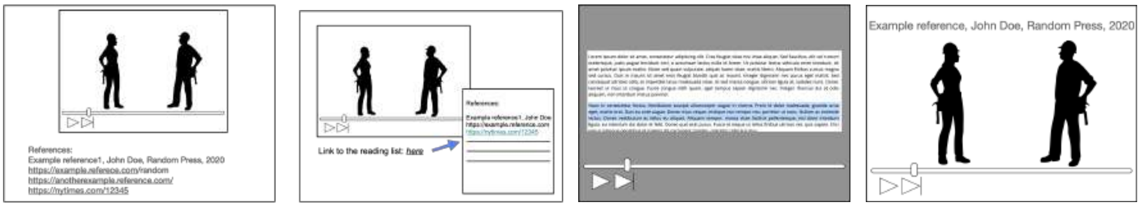
Contributing partner: IDIAP

As a final contribution in T6.5, we studied another important information channel beyond newspapers, namely YouTube, in connection to the quality of information of online videos. In recent years, there has been widespread concern about misinformation and hateful content on social media that are damaging societies. Being one of the most influential social media that practically serves as a new search engine, YouTube has accepted criticisms of being a major conduit of misinformation. However, it is often neglected that there exist communities on YouTube that aim to produce credible and informative content, usually falling under the educational category. This video material is also referred to in the media literature as knowledge communication videos [287]. One way to characterize this valuable content is to find references entailed to each video. While such citation practices function as a voluntary gatekeeping culture within the community, how they are actually done varies and remains unquestioned. Through the work of an AI4Media Junior Fellow hosted at Idiap, our research aimed to investigate common citation practices in major knowledge communication channels on YouTube using an in-depth manual selection and close analysis of videos [288]. Methodologically, this complements other methods based on computational techniques. After investigating 44 videos manually sampled from YouTube, we characterized two common referencing methods, namely *bibliographies* and *in-video citations*, as illustrated in Figure 59. We then selected 129 referenced resources, and assessed and categorized their availability as being *immediate*, *conditional*, and *absent*. After relating the observed referencing methods to the characteristics of the knowledge communication community, we showed that the usability of references could vary depending on viewers' user profiles. Furthermore, we witnessed the use of rich-text technologies that can enrich the usability of online video resources. In our paper [288], we discussed design implications for the platform to have a standardized referencing convention that can promote information credibility and improve user experience.

7.5.1 Experiments

We collected and qualitatively analyzed 44 English-speaking knowledge communication videos uploaded on YouTube. The three pillars of observation were 1) common referencing methods; 2) characteristics of the video creators; and 3) availability of the references. To characterize different referencing methods, we watched parts of the videos and observed different methods to put references in the in-frame elements and the video description section. We categorized the way how creators mention the existence of resources that were not produced by themselves, whether a commercial product, a document in any form, or a third person that supports or is related to what is being delivered in the video. To analyze the availability of referenced resources, we randomly selected up to 4 items per each video from the list of references in the metadata field or from the video in the predefined observed duration, and then searched each reference using Google search engine and took note of the reference availability (or lack thereof) for a general viewer.

Our analysis revealed that 39 among 44 videos used bibliography lists, where resource identifiers are listed in a text-based format. The analysis also showed that 23 among 44 videos used in-video citations, i.e., captions that are inserted in the visual body of a video and contain information needed to identify source materials.


References:
Example reference1, John Doe, Random Press, 2020
https://example.referece.com/random
https://anotherexample.reference.com/
https://nytimes.com/12345
Link to the reading list: here
References:
Example reference1, John Doe
https://example.reference.com
https://nytimes.com/12345
Example reference, John Doe, Random Press, 2020

(a) A bibliography in video metadata section (b) A bibliography as a file attachment (c) An in-video citation as a screenshot of an online article (d) An in-video citation in open captions

Figure 59. An illustrative example of video referencing methods.

Furthermore, among 44 investigated videos, 25 specified the names and roles of the collaborators (writer, narrator, editor, researcher and reviewer, visual effects and animation, etc.) in description sections. The number of people who participated as co-writers, whose roles were named as writer, researcher, host, or fact-checker, did not exceed 3. Among 44 creators-in-chief, 34 had university-level experience, 18 of which with graduate school experience.

Finally, among 129 references we fetched and reached back, we found 63 references whose entire content could be immediately reviewed and accessed by a viewer; 22 references where a viewer could immediately reach the item but did not have full access to the content; 29 references where a viewer could reach the online catalog page of the physical version of the document but no immediate online access; and 15 references where the material could not be found or accessed (see Table 39). These results show that actual access to the original scientific sources that videos refer to is not always possible. This situation sets limits on the ability of viewers to potentially verify the information provided in the video and improve or deepen their understanding of the presented concepts.

7.5.2 Conclusions

Our investigation first showed that many of the knowledge communication video creators not only specified (i.e., named) information sources, but also specified collaborators in video-making. Pro-

Category	Availability (# of videos)	Description
Immediate	Online Open Access (63)	Available online without any authorization.
Conditional	Conditional Online Access (22)	Available online after authorization/payment.
	Online Catalog (29)	Physically available after payment; delivery delay.
Absent	Not Found (14)	The resource is not found
	Not Available (1)	A message states that resource is not available.

Table 39. Summary of availability and accessibility of references.

viding this information leads to better interactions towards verifying the content, and also reduces the burden of transcribing the video while verifying the content. Considering that there is no reward system for such kinds of action in the YouTube platform, we could argue that the knowledge communication video community is more motivated and aware of the importance of making the video creation process more transparent. This could in turn be related to the high percentage of postgraduate education experience in the observed group, which reflects the importance of formal education and of training around digital media literacy.

Also, our investigation showed that references in YouTube knowledge communication videos adopt the conventions of text-based media to cue the audience and list the resource in rich-text format. By bridging the conventional gap between traditional media (scientific journals) and new media (social video platforms), the YouTube knowledge communication community is reinventing videos as a structured informational medium that requires active reviews and quality assessment by its viewers.

Finally, there still exists a possibility that viewers cannot review the resource immediately because of their affiliation or subscription status, or due to economic affluence. This imbalance in terms of access to the original scientific sources might affect the opportunity that viewers have to assess the content's credibility and, eventually, the effectiveness of referencing on YouTube for the general public.

7.5.3 Relevant publications

- HE. Kim and D. Gatica-Perez: "Referencing in YouTube Knowledge Communication Videos," in Proc. ACM International Conference on Interactive Media Experiences (IMX), Nantes, Jun. 2023. [288].
  Nominated for Best Paper Award.
  Zenodo record: https://zenodo.org/record/8190321.

7.5.4 Relevant software, datasets and other resources

- This work involved the manual coding of a small number of YouTube videos. This manual dataset could be made available to other interested researchers (but not the videos themselves, which are bound to YouTube terms.)

7.5.5 Relevance to AI4Media use cases and media industry applications

The research in this task has a natural alignment with UC4 "AI for Social Sciences and Humanities" as it proposed the used of qualitative methods for in-depth analysis of media content from a human-centered perspective.

8 Measuring and Predicting User Perception of Social Media (T6.6)

Contributing partners: UPB, QMUL, UvA

This section summarizes the contributions of participating partners to Task 6.6 - Measuring and Predicting User Perception of Social Media. This area of research joins researchers from many domains, including but not limited to computer vision, sociology, and psychologists working and analyzing human perception, targeting various concepts related to the way humans understand, perceive, and are affected by multimedia samples and posts in social media environments. Numerous studies have shown that the study of human perception of multimedia data represents one of the more difficult areas of research in computer vision, especially given the subjective nature of the annotations associated with datasets, and the evolving nature of human preferences and opinions [289], [290]. The work presented in this Section attempts to alleviate and study this problem, by proposing methods for the analysis of hard-to-classify visual samples, and proposing methods that bridge the gap between scientific datasets and in-the-wild scenarios.

The works summarized in this Deliverable are related to media memorability prediction in Sections 8.1 and 8.2, emotional content of media samples in Section 8.3, and viewer perception of Echo-chambers in Section 8.4.

8.1 Assessing the difficulty of media memorability prediction

Contributing partner: UPB

The study of human memory has been extensively studied in various fields, including psychology, cognition, physiology, and computer science. Memorability represents an essential aspect of data, closely related to learning, decision-making and creating lasting impressions. From a computational standpoint, memorability has been closely studied, with numerous papers dealing with image or video interestingness, targeting short- and long-term memorability. Interest in this domain has significantly grown, as benchmarking competitions like the MediaEval Predicting Video Memorability[16] (PVM), whose development is partly supported by AI4Media [291], provided common training and evaluation data, metrics, and tasks for interested participants. A large number of research papers attempting to predict memorability scores have been published during the five editions of this task. These range from using adapted large language models [292], convolutional and deep networks [293], to the use of transformer networks [294] and ensembles of various networks [295]. However, no study has yet been performed that analyzes the correlation between the performance of AI models and the visual aspect and content of the videos themselves.

In this context, we propose an in-depth analysis of the underlying visual and content-based factors that may influence how hard to classify from a memorability standpoint video samples are. From the 33 methods submitted by participants during the 2022 edition of PVM, we keep and analyze only 31, as the others are incomplete and may affect our analysis. This type of analysis is important not only for the insight it brings to the methods of classification and for understanding memorability, but it may also help in producing better results, as future training schemes may take the conclusions of this work into account, and augment the videos that may be harder to classify.

8.1.1 Experiments

8.1.1.1 Difficulty metrics and sample selection criteria We create a normalized distance metric, that has the role of measuring how difficult to predict and classify a video is from a

[16] https://multimediaeval.github.io/editions/2022/tasks/memorability/

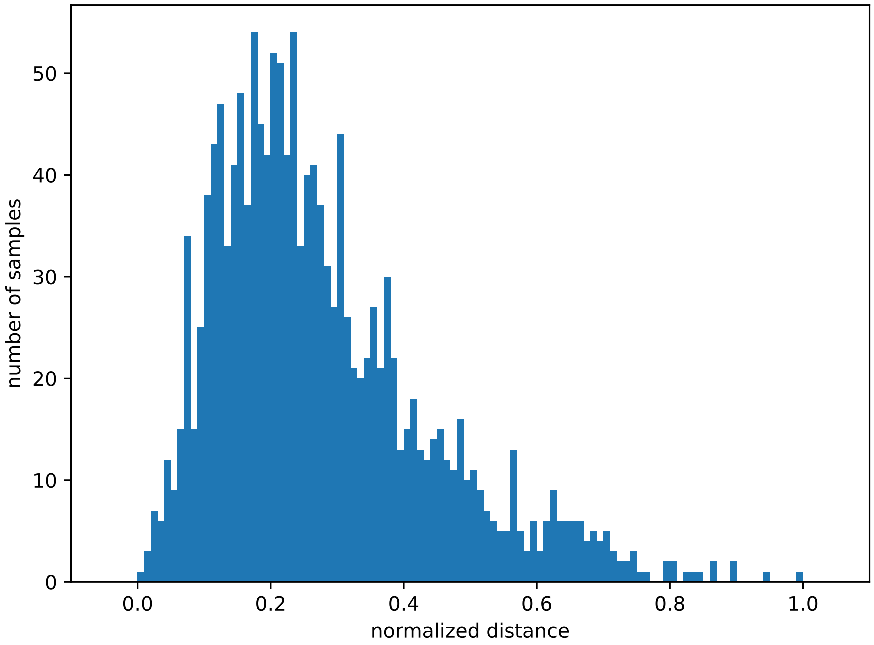


Figure 60. Distribution of the video samples in PVM according to the Distance metric.

memorability standpoint. Starting from the official metric, Spearman's rank correlation:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{21}$$

where N is the number of samples in two collections that are being compared, and d_i represents the difference between the two ranks for sample i, with $i \in [1, N]$. Therefore, for the given data and metric, the difference from the ground truth data ranks $G = \{g_1, g_2, ..., g_n\}$ is defined by the d_i distance. Given a set of M participant runs, $R = \{R_1, R_2, ..., R_M\}$, each run having the ranks of its predictions for the N samples in the dataset $R_j = \{p_{1,j}, p_{2,j}, ..., p_{N,j}\}$, the distance for a movie i can be expressed as:

$$D_i = \sum_{j=1}^{M} |p_{i,j} - g_i| \tag{22}$$

In the final step, we propose normalizing the set of distances for the N videos, obtaining a set of distances $\hat{D} = \{\hat{D_1}, \hat{D_2}, ..., \hat{D_N}\}$ with values $\hat{D_i} \in [0, 1]$. Figure 60 shows the distribution of the samples in the 2022 PVM, according to their normalized distance metric.

It is interesting to note that Figure 60 clearly shows that the results are skewed towards lower values of the distance metric. This indicates that the majority of video samples in the testset of PVM 2022 are easier to classify, on average, by the various AI models proposed by participants, with a median distance value of 0.2391 and an average value of 0.2774. We create two types of splits, based on two different criteria. The first one splits the videos into equal quartiles, denoted Q_1, Q_2, Q_3 and $Q4$, while the second one splits the videos using a 0.25 step according to the distance metric, denoted T_1, T_2, T_3 and T_4. Q_1 and T_1 respectively signify the easiest to predict videos, while Q_4 and T_4 represent the hardest to predict.

8.1.1.2 Features We propose a set of features that are able to describe these video samples according to their content. We apply and compute these features for each video samples, and then compute average values for each video category. We use three types of features, namely:

- visual features, implementing several traditional visual feature extractors;
- object-based features, implementing object detectors for detecting the objects in the video files;
- annotator-based features, looking for correlation between annotations and how hard to predict videos are.

Visual features

We compute the following visual features:

- sharpness via the Laplacian (f_1) operator [296];
- sharpness via the Canny (f_2) operator;
- colorfulness of images (f_3), based on the "psychophysical" experiments described in [297];
- contrast feature (f_4) that computes the contrast of images in RGB color space [298];
- average the pixels of the images transformed to HSL color space, resulting in a hue (f_5), saturation (f_6) and brightness (f_7);
- dynamism across the entire video (f_8) using a dense optical flow function computed via the Farneback method [299].

Object-based features

We use the architecture presented in [300], based on the MaskR-CNN [301] for automatically creating object annotations for the video samples. We use two features based on the objects detected in the videos, namely:

- top-5 most common objects (f_9);
- percentage of the frame size that is covered by detectable objects (f_{10}).

Annotator-based features

Finally, we wish to analyze the correlation between the created video categories and the annotated ground truth values. Starting from the four generated quartiles, we will compute and analyze a histogram that measures the distribution of each video quartile given the ground truth memorability scores annotated by participants to the PVM experiments.

8.1.1.3 Experimental results The results of the 8 visual features are presented in Table 40, where we present the percentage difference between the categories and a baseline composed of Q_1 for the quartile categories, and T_1 for the threshold categories. Several features present interesting results, while others are either inconclusive, or do not show any tendencies towards influencing the difficulty of memorability prediction. Experimental results show that videos with lower colorfulness, higher color saturation, lower brightness, higher average value component and less dynamism are harder to automatically predict.

The results for the object-based features are presented in Table 41. Overall the most common object in the entire video collection are "person" in 72.6% of all the videos, "chair" in 8.94%, "car" in 5.86%, "dining table" 4.26%, and "bird" in 3.93%, while 10.66% of the videos do not have any detectable objects in them. One of the most interesting observations regarding f_9 is related to videos without any discernible objects, where the top easiest to classify videos have less samples with no objects present. Major differences can also be seen for the "car", "table", and "bird"

Feature	Q2	Q3	Q4	T2	T3	T4
f_1	30.58%	40.82%	30.12%	20.67%	18.11%	32.48%
f_2	16.67%	18.59%	11.36%	5.67%	4.99%	31.25%
f_3	-6.54%	-4.25%	-4.11%	-3.62%	-1.53%	3.76%
f_4	15.54%	17.44%	18.65%	9.35%	3.27%	5.01%
f_5	-1.05%	-0.21%	-0.39%	-0.35%	2.42%	10.97%
f_6	0.54%	6.51%	2.59%	3.32%	1.55%	5.14%
f_7	-6.84%	-6.21%	-6.43%	-3.55%	-0.66%	-1.14%
f_8	-1.67%	-10.91%	-10.01%	-5.51%	-18.38%	-38.13%

Table 40. Visual feature analysis. Percentage change between the lower quartiles (Q2 - Q4) and the top quartile (Q1), and between the lower threshold intervals ($T2 - T4$) and the top interval (T1) for features f_1 - f_8.

Feature	Q2	Q3	Q4	T2	T3	T4
$f_9 - pers$	5.63%	2.63%	1.12%	-1.99%	-3.55%	9.11%
$f_9 - none$	27.69%	19.94%	47.97%	-3.77%	32.67%	-27.1%
$f_9 - chair$	-4.21%	-8.28%	-12.5%	5.73%	-29.67%	-
$f_9 - car$	-21.1%	-31.55%	-10.65%	-7.41%	-35.73%	-
$f_9 - table$	92.24%	199.46%	99.46%	25.36%	45.77%	94.4%
$f_9 - bird$	62.91%	138.02%	62.91%	92.14%	74.64%	-
f_{10}	-7.78%	-12.08%	-10.56%	-7.44%	-11.09%	-12.34%

Table 41. Object-based feature analysis. Percentage change between the lower quartiles (Q2 - Q4) and the top quartile (Q1), and between the lower threshold intervals ($T2 - T4$) and the top interval (T1) for features f_9 and f_{10}.

classes, however, given their low representation overall these may not actually represent significant changes in the video samples. Another interesting result is represented by the f_{10} feature, showing that easy to classify videos generally have a higher object coverage.

Finally, the results for the annotator-based features are presented in Figure 61. The most and least memorable videos generally belong to the Q_1 category, while videos with mid-level memorability are more likely representatives of the other categories. This is a result that we expected, as we theorize that videos that are harder to predict by AI models may be influenced by a lower agreement between human annotators.

8.1.2 Conclusions

This work presents an analysis of typical video features and attributes, and their correlation with how easy or hard they are to classify by AI models from a video memorability standpoint. We gathered a large collection of AI models used and submitted by participants to the 2022 MediaEval Predicting Video Memorability task. Our experiments show that videos that are harder to classify have the following attributes: (i) they have higher contrast and sharpness, but lower dynamism; (ii) they have lower brightness and colorfulness, but a higher saturation; (iii) they have fewer discernible objects in them, and the coverage of these objects is smaller; (iv) they tend to have mid-level memorability scores.

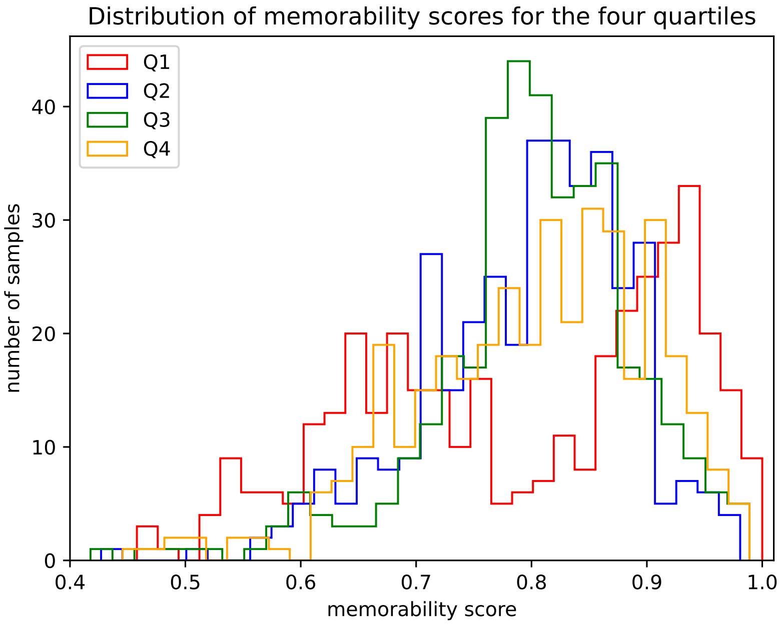


Figure 61. Distribution of the video samples in the four quatiles, according to their memorability score as annotated by human assessors, grouped in memorability score intervals of 0.02.

8.1.3 Relevant publications

- M.G. Constantin, M. Dogariu, A.-C. Jitaru, B. Ionescu: "Assessing the difficulty of predicting media memorability". 20th International Conference on Content-based Multimedia Indexing, CBMI 2023, September 2023. [302].

8.1.4 Relevant software, datasets and other resources

- No additional resources published yet.

8.1.5 Relevance to AI4Media use cases and media industry applications

This research can be applied to the training of systems proposed for UC3 "AI for Vision" and requirement 3C2-13 "Modality-dependent sentiment analysis". Concretely, this type of work can be applied and integrated into affective-content prediction AI models. Given the subjective nature of such data, which in turn leads to systems with lower accuracy, we believe that the analysis of the data itself could provide interesting insights at training time, especially through data augmentation of samples that are less easy-to-classify. This observation applies not only to the AI4Media use case, but to the entire subjective AI community.

8.2 Predicting Media Memorability Using Video Vision Transformers and Augmented Memorable Moments

Contributing partner: UPB

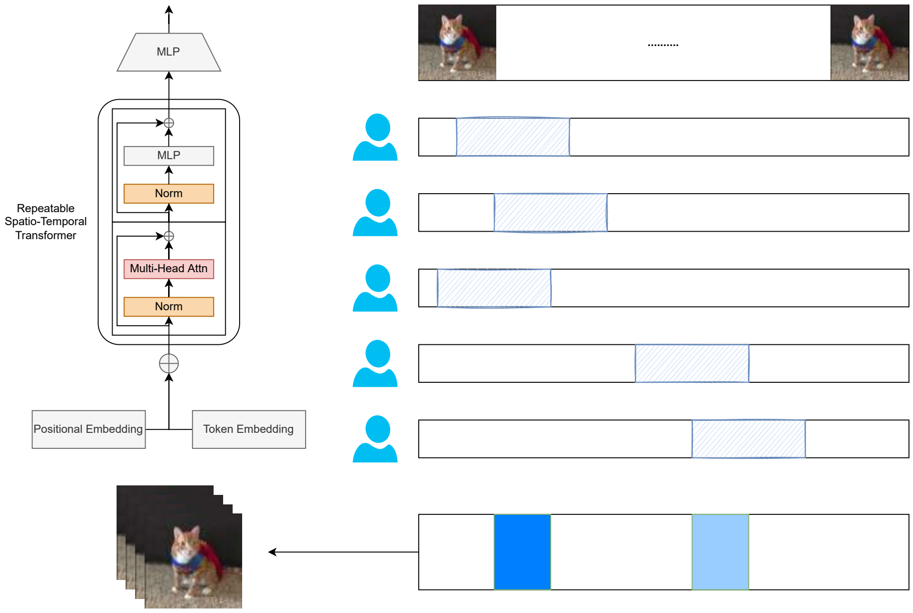


Figure 62. Memorability prediction model. The Memorable Moments Frame Selection phase computes the most representative video segments, represented by the blue segments. These frames, along with the ground truth memorability score for the entire video clip are then passed to the ViViT architecture. Annotator picks on the training set for the Memorable Moments are presented with blue sketched segments.

Media memorability prediction is a research domain that has gained considerable traction in the computer vision community, thanks to the development of large social media and multimedia visual repositories. A significant push for the prediction of video memorability has been recorded since the inception of the MediaEval Predicting Video Memorability task [291], a benchmarking task partly supported by AI4Media now at its fifth edition. The data offered by this task is extracted from the popular Memento10k dataset [303].

These experiments represent a continuation of previous experiments [304], presented in Deliverable D6.1. In the previous experiments, we proposed and proved that having a frame-selection method is a positive addition to the overall performance of a memorability prediction AI model, as this would help the model ignore frames and sequences that are not important when dealing with memorability prediction. While our previous approach used two image processing deep neural networks, based on the DeiT [305] and BEiT [306] models, for this update we propose using a video-processing network. On the other hand, while our previous approach dealt with only one segment of frames, for the novel Augmented Memorable Moment approach, we propose selecting several memorable moments as representatives for video samples at training time.

8.2.1 Methods

The proposed method is presented in Figure 62. For processing the data we use the popular ViViT [307] vision transformer-based architecture, while testing several configurations of the network. We vary the following network parameters: the number of parallel self-attention blocks, using values $4, 8, 16, 32$, and the number of repeatable blocks, testing the same values. We use the tubelet embedding, as defined by the authors, which creates a set of 3-dimensional tubes representing slices of the entire video.

For the Augmented Memorable Moments, we propose several variations for choosing represen-

Nr. Heads	Spearman
4	0.5831
8	*0.5942*
16	0.5876
32	0.5879

Nr. Repeats	Spearman
4	0.5831
8	*0.6054*
16	0.5980
32	0.5601

α	Spearman
0.95	0.5974
0.90	0.6376
0.85	*0.6410*
0.75	0.6248

Table 42. Study performed on the validation set, concerning the three variable parameters of the proposed method, namely the number of heads, the number of repeats and the α parameter.

Method	Spearman	Pearson	MSE
AIMultimediaLab-subtask1-Single	0.618	0.622	0.007
AIMultimediaLab-subtask1-Double	0.648	0.650	0.006
AIMultimediaLab-subtask1-Multi	**0.665**	**0.669**	**0.006**

Table 43. Final results for the proposed method, under the Single, Double and Multi Memorable Moments configuration.

tative segments of larger videos. Given a clip composed of a set of N frames $V = \{f_1, f_2, ..., f_N\}$ and a set of M annotators $A = \{a_1, a_2, ..., a_M\}$, each annotator will, according to the established annotation protocol, push a button when they recall seeing a video clip. Given a response delay of approximately 500 milliseconds, obtained in our previous experiments, we allocate a score of 1 for a number of 15 frames around the frame of recall. Furthermore, we sum up all the annotations, obtaining a score for each frame $S_i = \sum_{j=1}^{M} s_i^j$. Finally, we propose three methods for frame selection, called *Single*, *Double* and *Multi*. The *Single* approach takes the highest value of S_i and uses the 15-frame interval around it as a representative of the entire video, the *Double* approach takes the top 2 highest values of S_i, while the *Multi* approach takes all the S_i values that are higher than a percentage of the top S_i value, searching for the best threshold value among the following: $\{0.75, 0.85, 0.90, 0.95\}$.

8.2.2 Experimental results

In our experiments we use the 2022 version of the MediaEval Predicting Video Memorability task [291]. The 2022 PVM dataset uses a set of 10,000 short soundless videos, depicting in-the-wild scenes. In total 7,000 videos are used for training, 1,500 for validation and 1,500 for testing the proposed methods.

We start with a set of preliminary experiments, attempting to determine the top performing architecture by using just the training and the validation sets. The results of these experiments are presented in Table 42. We obtain the best results for a number of 8 self-attention heads and a number of 8 repeatable blocks in the ViViT structure, while a threshold of 0.85 for selecting the number of *Multi* interval centroids obtains the best result.

The final results, obtained on the testset of the competition, are presented in Table 43. The best result is obtained by using a *Multi* configuration, with a final Spearman value of 0.665. While we can observe a significant rise in performance when going from the *Single* to the *Double* configuration, an even better result is achieved with the *Multi* configuration. We theorize that each larger Memorable Moments scheme adds more data to the training set, generating progressively more video segments associated with each video clip, while also ensuring that these segments are representative for the entire video from a memorability standpoint.

8.2.3 Conclusions

Our updated approach for selecting Memorable Moments from video clips uses and implements a video processing vision transformer, and tests several schemes for Memorable Moment selection. Our experimental results show that schemes that select more than one representative video segment tend to out-perform single segment selection methods. One of the most important limitation for this approach relates to predicting longer videos, where the subject of the video itself may change, perhaps even several times per larger clip. Solving this limitation would first involve creating or using a dataset with longer videos, and applying video splitting methods based on changes in the subject, in order to correctly select the Memorable Moments from each video split.

8.2.4 Relevant publications

- M.G. Constantin, B. Ionescu : "AIMultimediaLab at MediaEval 2022: Predicting Media Memorability Using Video Vision Transformers and Augmented Memorable Moments". In Working Notes Proceedings of the MediaEval 2022 Workshop, Bergen, Norway, 2023. [294].

8.2.5 Relevant software, datasets and other resources

- No additional resources published yet.

8.2.6 Relevance to AI4Media use cases and media industry applications

This research can be applied to UC3 "AI for Vision" and feature 3C2-13 - Modality-dependent sentiment analysis. The prediction of media memorability is one of the most important tasks in the analysis of subjective multimedia data. The application of AI models that are able to predict how memorable certain media samples are is of use for content creators and news agencies, as they try to create memorable experiences for their viewers and readers, but also in other domains like education, where educators and professors could create a better and easier-to-memorize experience for their students.

8.3 Learning from Label Relationships in Human Affect

Contributing partner: QMUL

Understanding human affect and mental state is a challenging and actively researched field with diverse applications in education [308] and healthcare [309]. It can be approached through classification using basic human emotions [310] or continuous labels along the Arousal-Valence axes [311]. However, regardless of the chosen labeling approach, certain challenges make the estimation of human affect and mental state difficult.

One of the challenges is the presence of in-the-wild datasets, which often consist of long videos with limited or no temporal label resolution. This means that a set of labels describes the entire video instead of specific moments within it. Additionally, publicly available datasets for human affect estimation tend to be small, leading to overfitting issues during training. Therefore, it is crucial to develop methods that can improve representations even with a limited number of samples, as they are essential for achieving success in the final regression task.

To address these challenges and enhance feature representations in human affect analysis, we propose two novel approaches. First, we introduce an attention-based video-clip encoder that builds upon previous work [312]. This encoder utilizes the temporal dimension of input clips and generates clip-level predictions that leverage contextual clip information. Second, we present a novel relational regression loss function that aligns the distances in the latent clip-level representations/features with the distances of the corresponding labels. These approaches aim to alleviate the difficulties associated with long video inputs and multi-label regression problems, contributing to improved continuous affect estimation.

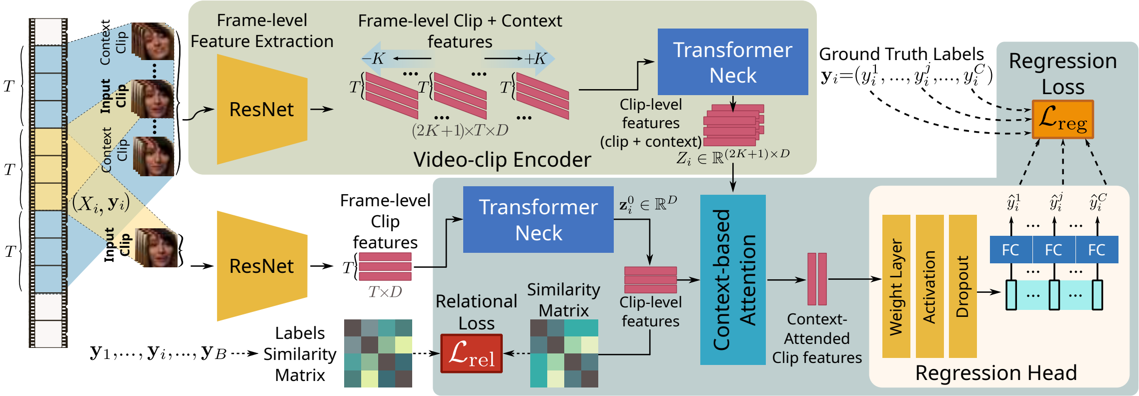


Figure 63. Overview of the proposed framework: (a) The bottom branch uses the proposed video-clip encoder (comprising of a ResNet frame-level and a Transformer clip-level feature extractors) to extract clip-level features from the input video clips, which subsequently feed the context-based attention block and are further used to construct the intra-batch similarity matrix for calculating the proposed relational loss. (b) The upper branch uses the proposed video-clip encoder to extract clip-level features from the input video clips and a set of context clips from each of the input clips, which subsequently feed the context-based attention block in order to infer the desired values and calculate the regression loss. The context-based attention block, fuses clip-level and context features and passes the context attended clip features to the regression head that estimates the desired continuous values. Error is back-propagated only through the shaded region of the bottom branch.

An overview of the proposed framework for the problem of multi-label regression from a sequence of clips is given in Fig. 63. In a nutshell, the proposed architecture consists of two branches with shared weights, that incorporate two main components: a) a video-clip encoder employing a convolutional backbone network for frame-level feature extraction and b) a Transformer-based network leveraging the temporal relationships of the spatial features for clip-level feature extraction. The clip and context features produced by the aforementioned branches are passed to a context-based attention block and a regression head. The proposed method uses the context-based attention block to incorporate features from the two branches before passing them to the regression head, as shown in Fig. 63. The bottom branch uses the proposed video-clip encoder to extract clip-level features from the input video clips, which subsequently feed the context-based attention block and are further used to construct the intra-batch similarity matrix for calculating the proposed relational loss.

The goal of the proposed relational loss, as an additional auxiliary task to the main regression, is to obtain a more discriminative set of latent clip-level features, by aligning the label distances in the mini-batch to the latent feature distances. At each training iteration, after having calculated the clip-level features for the clips in a mini-batch, i.e., $\mathbf{z}_i^0 \in \mathbb{R}^D$, $i = 1, \dots, B$, we calculate the proposed relational loss as follows:

$$\mathcal{L}_{\text{rel}} = \sqrt{\frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}\left(\hat{M}_{i,j} - M_{i,j}\right)^2} \tag{23}$$

where $\hat{M} \in \mathbb{R}^{B\times B}$ denotes the cosine similarity matrix calculated on the clip-level features, and $M \in \mathbb{R}^{B\times B}$ denotes the cosine similarity matrix calculated on the ground truth labels.

8.3.1 Experiments

8.3.1.1 Datasets The **"OMG-Emotion Dataset"** [313] consists of in-the-wild videos of recorded monologues and acting auditions, collected from YouTube and annotated for Arousal and Valence by expert annotators. As a number of videos have been removed since the publication of the dataset, we train and validate on a subset of the original set. The **AMIGOS** dataset [314] consists of audio-visual and physiological responses of participants (either alone or in a group) to a video stimulus, and annotated for Arousal Valence by three expert annotators. We trained the network following a leave-one-subject-out cross validation scheme.

8.3.1.2 Results For the experiments conducted on the AMIGOS dataset [314], we compared the performance of the proposed methodology against previous state-of-the-art [315] for the face modality and we show the results in Table 44. The proposed methodology leads to a clear improvement against both baselines, trained with an RMSE regression loss ($\mathcal{L}_{RMSE}$) and a CCC loss ($\mathcal{L}_{CCC}$). We also outperform previous state-of-the-art by a large margin for both Arousal and Valence.

	Arousal		Valence	
	PCC	CCC	PCC	CCC
Proposed	**0.69**	**0.68**	**0.75**	**0.74**
Proposed $\mathcal{L}_{CCC}$ w/o K w/o$\mathcal{L}_{rel}$	0.59	0.49	0.64	0.54
Proposed $\mathcal{L}_{RMSE}$ w/o K w/o $\mathcal{L}_{rel}$	0.60	0.39	0.55	0.40
Mou *et al.* [315]	0.60	0.59	0.62	0.61

Table 44. Performance of the proposed method against baseline and other uni-modal architectures (AMIGOS).

The results of our study on the OMG dataset [313] are presented in Table 45.Comparing the proposed methodology against its baseline (i.e., "w/o K w/o $\mathcal{L}_{rel}$"), we observe that the proposed relational loss improves the performance of the regression measured in terms of CCC, for both Arousal and Valence. Further incorporating the contextual features improved the CCC score for Valence, but lowered slightly the CCC for Arousal.

8.3.2 Conclusions

The proposed architecture is novel in the domain of affect and mental state analysis and leads to smaller training times in comparison to state-of-the-art. Furthermore, we introduced a novel relational regression loss that aims at learning from the label relationships of the samples during training. The proposed novel loss uses the distance between label vectors to learn intra-batch latent representation similarities in a supervised manner. The improved latent representations obtained with the addition of the relational regression loss lead to improved regression output, without the use of large datasets. We demonstrated the effectiveness of the proposed method on two datasets for continuous affect estimation, and we showed that our method achieves results outperforming previous state-of-the-art. As the context used in this work is bi-directional, the method assumes prior knowledge of the entire video and is therefore limited in offline evaluation. Furthermore, as affect datasets are typically small, the method's generalisation is limited by the training size and would inherit the bias of the dataset. Finally, as with all human-centric methods, there are ethical and privacy concerns, when used outside research settings. To address the context limitation, additional experiments need to be conducted to evaluate the method's online capabilities, as well as generalisation abilities on different datasets. To address the dataset bias, additional research on mitigating bias in affective computing needs to be conducted.

	Arousal	Valence
Proposed	0.26	**0.48**
Proposed w/o K	**0.29**	0.46
Proposed w/o K w/o $\mathcal{L}_{rel}$	0.24	0.44
Proposed w/ $\mathcal{L}_{cont.}$	0.15	0.32

Table 45. Performance (CCC) of the proposed method against baseline (OMG).

8.3.3 Relevant publications

- N. M. Foteinopoulou and I. Patras: "Learning from Label Relationships in Human Affect". ACM International Conference on Multimedia 2022 (MM '22) [316].
  Zenodo record: https://zenodo.org/record/8028376.

8.3.4 Relevant software, datasets and other resources

- Code is available at: https://github.com/NickyFot/ACMMM22_LearningLabelRelationships

8.3.5 Relevance to AI4Media use cases and media industry applications

This work contributes to UC3 "AI in Vision - High Quality Video Production and Content Automation" and feature 3C2-13 "Modality-dependent sentiment analysis" as it provides valuable findings on human affect in context. As emotion is not static and is related to a wider range of behaviours, the wider temporal context needs to be included for more accurate and informative understanding of sentiment across multiple modalities. This research can be applied to the training of systems for sentiment analysis. Such system could be useful in media and advertising, as the reaction of viewers to media stimuli can be measured rather than relying on self-reporting.

8.4 User Perception and Measuring Disinformation Echo-chambers on Facebook

Contributing partner: UvA

The landscape of information has experienced significant transformations with the rapid expansion of the internet and the emergence of online social networks. Initially, there was optimism that these platforms would encourage a culture of active participation and diverse communication. However, recent events have brought to light the negative effects of social media platforms, leading to the creation of echo chambers, where users are exposed only to content that aligns with their existing beliefs. Furthermore, malicious individuals exploit these platforms to deceive people and undermine democratic processes. To gain a deeper understanding of these phenomena, this study introduces a computational method designed to identify coordinated inauthentic user behavior and perception within Facebook groups. The method focuses on analyzing posts, URLs, and images, revealing that certain Facebook groups engage in orchestrated campaigns. These groups simultaneously share identical content, which may expose users to repeated encounters with false or misleading narratives, effectively forming "disinformation echo chambers." This study concludes by discussing the theoretical and empirical implications of these findings.

8.4.1 Background, Data and Methods

During crises, public discourse plays a vital role in influencing collective attitudes and actions. In the context of health crises, the spread of disinformation becomes even more concerning. Past experiences with public health crises, such as the Ebola outbreak in 2014 and the H1N1 epidemic in 2009, have highlighted the abundance of misinformation and false information propagated through social media [317].

The COVID-19 pandemic has highlighted the disruptive nature of false and misleading information and the impact of political disinformation on public health policy outcomes. Disinformation surrounding COVID-19, particularly concerning the safety and effectiveness of vaccines, has been widespread and problematic [317], [318].

The spread of false information about COVID-19 vaccines has led to the circulation of various myths and misconceptions. These myths have created hesitancy and doubts among the public, hindering vaccination efforts and potentially compromising the success of public health policies designed to control the pandemic.

In the literature [319], [320], collusive users are characterized as accounts deliberately engaged in disseminating false perceptions to influence public debate. These users often have a significant number of retweets, followers, or likes, which gives them a certain level of influence. They may be backed or made up of a network of individuals who engage in mutual follower exchanges to enhance their visibility.

These collusive users pose a significant challenge as they undermine the trust and credibility of online platforms, much like spam accounts do. By spreading disinformation and manipulating social metrics, they can distort public discourse and mislead others, potentially causing harm to public opinion and decision-making processes. It is crucial for platforms and users to remain vigilant and combat these deceptive practices to maintain the integrity of online discussions and information dissemination.

It is crucial to address and combat these myths through accurate and evidence-based communication to ensure that the public receives reliable information about COVID-19 vaccines and can make informed decisions regarding their health and well-being. Since 2018, Facebook has been using the concept of "coordinated inauthentic behavior" to remove content from its platform [321]. Instead of relying on a clear-cut distinction between problematic and non-problematic information, Facebook adopted a more ambiguous approach, encompassing not only bots and trolls that spread false content but also unwitting individuals and polarized groups recruited to actively influence society.

This "ill-defined concept of coordinated inauthentic behavior" has faced criticism, as some argue that Facebook is attempting to enforce its policies without clear guidelines [322]. However, the company has justified its approach by establishing a link between coordinated behavior and the dissemination of problematic information. By targeting such behavior, Facebook aims to mitigate manipulation attempts on its platform and maintain the integrity of its information ecosystem.

In this study, the focus is on exploring the role of users' content exhibiting signs of coordinated inauthentic behavior within Facebook groups, specifically through the lens of echo chambers. To analyze disinformation narratives related to COVID-19 vaccines on Facebook, the researchers gathered fact-checked stories from two Brazilian fact-checking initiatives (Agência Lupa and Aos Fatos) published between January 2020 and June 2021. These stories were used to source keywords for queries on CrowdTangle and identify active false narratives on Facebook. A total of 276 debunked stories were utilized in the study to uncover the disinformation narratives being propagated on the platform.

The computational method employed in this research aims to predict instances of coordinated behavior among Facebook groups that engage in inauthentic tactics to increase the visibility and reach of problematic content on the platform. The method involves analyzing the frequency and similarity of content to detect potential traces of coordinated inauthentic behavior. This may involve the replication of widely available narratives within the Facebook groups or the sharing of common links in a short timeframe, leading to external websites. The study also takes into account the coordinated dissemination of visual content, such as memes, which can be easily manipulated but challenging to detect using conventional computational methods. To address this, the researchers utilized a computer vision algorithm provided by Facebook to analyze the textual content within images and identify if multiple images shared the same message within a short period.

By employing this computational strategy, the study aims to shed light on the extent and nature of coordinated inauthentic behavior on Facebook and its potential role in amplifying problematic information within echo chambers.

8.4.2 Results

Findings reveal that coordinated efforts on Facebook have taken a disconcerting turn, manipulating public discourse with a strategic intent that we described as the creation of "disinformation echo chambers." As illustrated in Figure 64, these calculated endeavors have given rise to an alarming pattern of information dissemination, where false narratives are propagated widely, leading to a troubling cascade of consequences.

The orchestrated campaigns orchestrated on the social media platform have demonstrated an uncanny ability to penetrate various groups across the digital landscape. These efforts generate high levels of interaction with false narratives across various groups, many of which have political affiliations under their names.

Furthermore, these false facts can reinforce existing biases, undermine public health efforts, and lead to negative consequences regarding COVID-19 vaccines.

Upon closer examination of Figure 64, the intricate web of connections between different Facebook groups becomes evident. This networked dissemination of false narratives is deliberately coordinated to replicate and reinforce disinformation across a multitude of platforms. The manipulative design of these efforts seeks to establish a sense of legitimacy through repeated exposure, making it increasingly difficult for users to discern fact from fiction.

8.4.3 Conclusions

This study concludes by emphasizing the significant risks posed by these inauthentic coordinated efforts, which can potentially create confusion and erode trust among the public, ultimately hindering effective public health responses.

The dissemination of such content within different groups creates beliefs that are amplified and reinforced through repetition, forming cohesive and insulated communities that perpetuate echo chambers. These coordinated efforts pose specific risks by deceiving users into replicating these false narratives offline, leading to vaccine hesitancy and avoidance.

It aims to contribute to the literature on disinformation and platforms by demonstrating how coordinated inauthentic information can foster echo chambers that amplify false or misleading narratives. Additionally, the analysis of the structural properties of these Facebook groups, with their strong coordination, highlights the potential risks posed by disinformation in influencing people's decisions about vaccines and jeopardizing the development and implementation of public health policies, including vaccination campaigns .

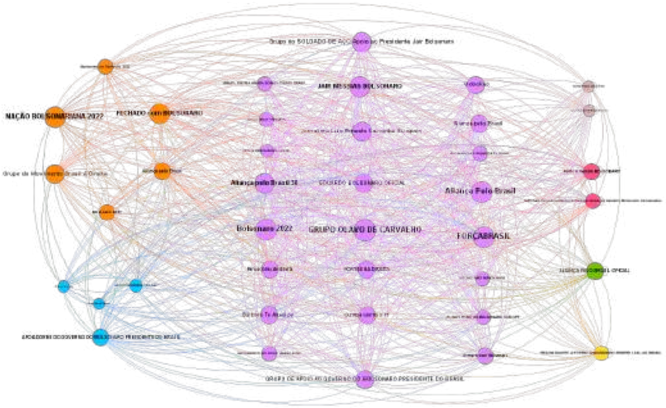

Grupo do SOLDADO DE AÇO Apoio ao Presidente Jair Bolsonaro
JAIR MESSIAS BOLSONARO
NAÇÃO BOLSONARIANA 2022
FECHADO com BOLSONARO
Aliança pelo Brasil 38
EDUARDO BOLSONARO OFICIAL
Aliança Pelo Brasil
Bolsonaro 2022
GRUPO OLAVO DE CARVALHO
FORÇABRASIL
JUSTIÇA PARA BRASIL OFICIAL
APOIADORES DO GOVERNO DO BOLSONARO PRESIDENTE DO BRASIL
GRUPO DE APOIO AO GOVERNO DO BOLSONARO PRESIDENTE DO BRASIL

Figure 64. This graph shows only Facebook groups with degrees above 100, that is, only those that have shared at least 100 coordinated posts. Most of these groups adopt political names. Source: Authors

8.4.4 Relevant publications

- de-Lima-Santos, M.F and Ceron, W. (in press). Disinformation Echo-chambers on Facebook In P.Seitz,; M. Eisenegger, & M. M. Bergman (Eds.). *Fighting Fake Facts.* MPDI Books [323].
- de-Lima-Santos, M.F and Ceron, W. (in press). Coordinated Amplification, Coordinated Inauthentic Behavior, Orchestrated Campaigns? A Systematic Literature Review of Coordinated Inauthentic Content on Online Social Networks In T. Filibeli & M. Ö. Özbek (Eds.). *Mapping Lies in the Global Media Sphere.* London: Taylor & Francis [324].

8.4.5 Relevant software, datasets and other resources

- No additional resources published yet.

8.4.6 Relevance to AI4Media use cases and media industry applications

This study aligns perfectly with the objective of measuring and predicting user perception on social media, offering valuable insights into the influence of user behaviors and consumption of disinformation content. The research also introduces an innovative computational method that uncovers the existence of disinformation echo chambers within public Facebook groups. This can help fact-checkers and civil society actors to combat coordinated efforts to disseminate disinformation in online spaces. Similarly, it can help these organizations to track the efforts made by tech platforms in moderating such content.

These disinformation echo chambers play a significant role in spreading false or misleading narratives aimed at discrediting vaccines and posing a threat to public health. By undermining the efforts of public authorities to combat the health crisis, these groups can have serious implications for public well-being.

The findings of this study have the potential to contribute to the development of new strategies to address online disinformation. By examining groups exhibiting traces of coordinated inauthentic behavior, the study suggests new measures for detection and possible mitigation of the tactics used to amplify problematic content and increase their reach.

We hope this research sheds light on effective ways to tackle online disinformation, providing valuable insights for policymakers, platform operators, and researchers to combat the spread of false narratives and enhance the integrity of information shared on social media. Ultimately, this can contribute to better public health outcomes by promoting accurate and reliable information dissemination.

9 Real-life effects of private content sharing (T6.7)

Contributing partners: CEA

Users are entitled to know how the data they share online can be leveraged by online social networks (OSNs) and third parties. Despite its importance for an informed online participation, the proposal of efficient user feedback remains difficult due to a combination of usability and technical challenges. Efforts to increase transparency and trust, such as Facebook's Privacy Checkup [325], were recently made under public and regulatory pressure but their effectiveness is strongly debated [326], [327]. The availability of on-device deep learning models [328], [329] enables the creation of AI-assisted tools which provide user feedback before sharing. This is important insofar as feedback is most efficient before sharing, when data are still on the user's device [330].

Task 6.7 examines the effects of data sharing, developing new techniques that predict how the user might be affected by publicly sharing their personal data (text, images, video, etc.) on social media, in the context of real-life situations like e.g. applying for a job or a loan. In the following, we report new work on predicting the consequences of online photo sharing, proposing a method for rating photographic profiles.

9.1 Raising user awareness about the consequences of online photo sharing

Contributing partner: CEA

Users should be able to understand the potential effects of their data-sharing practices. This topic was explored during the ImageCLEF 2021 Aware Task [331], which introduced a dataset which includes photographic user profiles, associating object detections and human ratings of these profiles in four real life situations. The modeled situations include search for: an accommodation, a bank loan, an IT job and a waiter job. Several situations are needed since the effects of data sharing vary depending on the context in which they are used, as illustrated in Figure 65. The objective of the Aware task was to automatically rate a set of photographic user profiles per situation in order to reproduce human ratings. The focus was on lightweight algorithms which can be run on users' devices. Inferences are thus done before the actual sharing, and this is important insofar as OSNs gain control of the data once they are shared.

We summarize the proposed method for rating photographic profiles in Figure 65. We note that the rating of these profiles varies in the three illustrated situations. For instance, the analysis of their shared photos is likely to place U_2 at the top of the user profile rankings for situations accommodation and IT job search because their shared photos are more appealing than those of U_1 and U_3 in the two situations. Inversely, U_2 will be low ranked for waiter job search because their photos are either unrelated to this situation or can be judged negatively. Note that the ratings of the profiles encode social biases of the people who perform the ratings [331], [332]. The automatic rating process will inherit these biases, but this is not considered problematic in our approach, because it is meant to simulate real-life situations, including biases. The rating process is based on Graph Neural Networks (GNNs) [333], whose ability to learn from entities and their relations in a flexible way [334] makes them particularly suited to our use case. We explore different ways to structure data as graphs and three GNN models to adapt the automatic rating of profiles to each situation.

9.1.1 Experiments

9.1.1.1 Dataset and metrics The ImageCLEF 2021 Aware dataset [331] includes 500 photographic user profiles sampled from the YFCC dataset [335], with 100 images per profile. Manual

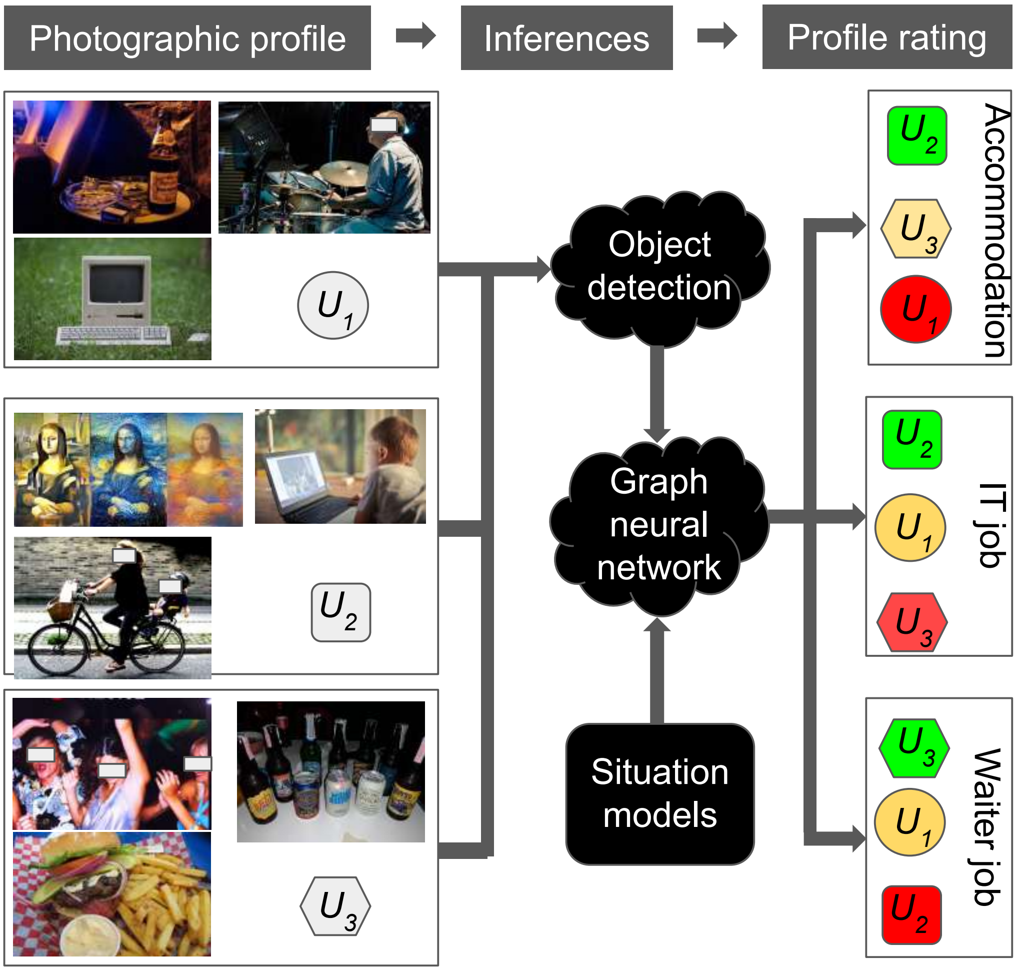


Figure 65. Potential consequences of personal data sharing for 3 users in 3 real-life situations. Users' photos include information about their interests and lifestyle, which can be inferred and aggregated to compute the appeal of a profile in context. The aggregation is done with a graph neural net trained to rate profiles. The network is trained with object detections and ratings per situation and uses human ratings of training profiles as ground truth. Modeling several situations is useful because the same shared information can be interpreted differently. For instance, the two top row photos of U_1 and U_3 might be assessed negatively when searching an accommodation, but positively for a waiter job. Top row photos of U_2 are positive for IT job, but their photos with young kids might be problematic if searching for a waiter job. A ranking of profile ratings is used to raise awareness about the consequences of data sharing.

Model	ACC	BANK	IT	WAIT	Average
BASE_η [332]	0.31	0.34	0.41	0.46	0.38
LERVUP^{fr} [332]	0.43	0.36	0.54	0.59	0.48
AUTO_{SKL}	0.27	**0.44**	0.56	0.68	0.49
GUAR_{SAG}	0.38	0.27	0.44	0.62	0.43
GUAR_{GIN}	0.44	0.20	0.54	0.59	0.44
$\text{GUAR}_{\text{MEAN}}$	**0.54**	0.36	**0.57**	**0.71**	**0.55**

Table 46. Pearson correlation coefficients ρ on the test split for the ImageCLEF 2021 Aware dataset [331] with the baselines and the proposed method. ρ interpretation is done using the ranges from [338]: weak for $[0.1, 0.3]$, moderate for $[0.3, 0.5]$, strong for $[0.5, 1]$. Best results for individual methods are in bold.

ratings of profiles per situation are obtained through crowdsourcing, and are used as ground truth. Profiles are characterized by two main elements, i.e. object detections and object ratings per situation. The dataset is split in 360, 40, 100 profiles for train, validation, and test. The evaluation objective is to produce an automatic ranking of profiles per situation which is as close as possible to the manual ground truth. The Pearson correlation coefficient is used to measure the correlation of human and automatic profile ratings for the test set. Correlations in individual situations are aggregated to provide a global score.

9.1.1.2 Baselines Two methods from [332] are tested here, along with a new baseline AUTO_{SKL}. BASE_η - represents photographic profiles as 269-dimensional vectors which combine object detections and ratings in each situation. Test profiles are ranked by summing individual object detections weighted by situation-specific object ratings. LERVUP^{fr} - compresses full profile vectors using object rating positivity, negativity, and average detection confidence. A focal rating component gives more weight to objects with have high ratings, which are likely to be the most influential. Profiles are automatically rated using a random forest model for regression with the compressed representations. AUTO_{SKL} - uses auto-sklearn [336] to search suitable models in Scikit-learn [337]. Similar to $BASE_\eta$, profiles are represented as 269-dimensional vectors. AUTO_{SKL} is interesting because it finds an optimal classical learning model.

9.1.1.3 Results The results from Table 46 indicate that the performance of $\text{GUAR}_{\text{MEAN}}$ is interesting because the correlation between manual and automatic profile rating is strong for ACC, IT and WAIT and moderate for BANK, following the intervals given in [338]. Equally important, $\text{GUAR}_{\text{MEAN}}$ provides a consistent performance improvement compared to the baselines. Globally, gains of 6 and 7 Pearson correlation coefficient ρ points are obtained compared to AUTO_{SKL} and LERVUP^{fr}, respectively. The results vary significantly between situations. WAIT is the easiest situation and BANK the most difficult one for $\text{GUAR}_{\text{MEAN}}$. The authors of [332] explained this variation, also observed for LERVUP^{fr}, by the fact that WAIT is better represented in the object detection dataset than the other situation. This explanation is related to the automatic inferences which are associated to a user's photos. A further explanation might be related to the quality of the ground truth data associated to each situation. Their judgments about the appeal of each profile in a situation is based on an aggregation of weak signals from profile photos. These signals are probably more interpretable in some situations than in others. $\mathrm{GUAR}_{\mathrm{MEAN}}$ has the best results among the three types of GNN models tested. $\mathrm{GUAR}_{\mathrm{SAG}}$ and $\mathrm{GUAR}_{\mathrm{GIN}}$ lag consistently behind, with average performance drops of 12 and 11 points with respect to $\mathrm{GUAR}_{\mathrm{MEAN}}$. This is interesting insofar as SAG and GIN aggregation components were proposed to improve over MEAN but are not efficient here. This finding might be explained by the size of the available training set. A simpler GNN architecture seems better suited here but further comparisons should be performed if the dataset is extended.

9.1.2 Conclusion

We have introduced a new GNN-based method which raises user awareness about the consequences of personal data sharing in impactful situations. Our results indicate that it is possible to automatically rate users' photographic profiles in an effective way. These ratings can then be aggregated into ratings using a community of reference in order to provide feedback to users about how well they stand in the crowd. The proposed model leverages graph flexibility to represent user profiles and an adapted data structuring process to regress the corresponding ratings to improve results compared to competitive baselines. The experiments show that the use of GNNs results in a consequent improvement of performance compared to existing methods.

9.1.3 Relevant publications

- Schindler, Hugo, et al. "Raising User Awareness about the Consequences of Online Photo Sharing." Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. 2023. .
  Zenodo record: https://zenodo.org/record/7940055.

9.1.4 Relevant software, datasets and other resources

- ImageCLEF Aware 2022 and 2023 datasets https://www.aicrowd.com/challenges/imageclef-2022-aware#data

9.1.5 Relevance to AI4Media use cases and media industry applications

This research is used in a standalone application prototype which gives feedback to social media users about the potential effects of online photo sharing. A video demo of the prototype is available at https://ydsyo.app/. More generally, the approach could be adapted to other types of personal data, such as political preferences, to give feedback about the political leaning of the news sources which they consult online.

10 Other relevant activities

Throughout this deliverable, we presented instances where reported contributions to WP6 represented the result of cooperation between partners, work packages, or use case integration, as well as Junior Fellow Exchanges, showing our interest in developing stronger collaborations between the organizations involved in this project. In addition to these, WP6 partners have collaborated for the organisation of special issues in journals, scientific workshops, special sessions in conferences, benchmarking activities and other events. This section summarises these contributions, and provides details regarding their intended purpose and topics.

Two **Multimedia Against Disinformation (MAD) workshops** were organized by partners participating in T6.2 at the 2022[17] and 2023[18] International Conference on Multimedia Retrieval. This series of workshops target several key aspects in disinformation detection, such as: multimodality, the analysis of disinformation campaigns, explaining disinformation, temporal and cultural aspects, multimedia verification systems, and ensembling techniques for disinformation detection. In the latest edition of this workshop, i.e. MAD'23 held in Thessaloniki, Greece on 12 June 2023, 7 papers were accepted, and 2 keynote speakers were invited. The workshop included three sessions on AI for audio analysis, Improving AI generalization, and AI for (Dis-)Information Analysis and attracted a diverse audience of researchers on AI and multimedia. The event was co-organised by AI4Media and vera.ai.[19]

Realistic Synthetic Data: Generation, Learning, Evaluation special issue in ACM Transactions on Multimedia Computing, Communications, and Applications[20], is a special issue call seeking innovative algorithms and approaches for generative AI models. The call for contributions for this special issue is currently over, with a large number of proposed works submitted (over 20 papers). The final tentative publication date is announced as December 2023. The special issue launched a call for contributions on topics like: synthetic data generation for various modalities, transfer learning in generative networks, addressing bias, limitations and trustworthiness, and ethical aspects of synthetic data generation. The special issue was endorsed by AI4Media, with guest editors including researchers from UPB, QMUL, HES-SO and UNIFI.

Predicting Video Memorability. The Predicting Video Memorability task represents a continuing initiative, first starting out as a dataset and task for the MediaEval Benchmarking Initiative[21], and evolving each edition and each year into a more complex task. This task deals with predicting the short- and long-term memorability of various types of social media videos, while also starting the development of an electroencephalogram signal-based sub-task. While the task and dataset are presented in more detail in WP4, T4.6 - Benchmarking of AI Systems, this work is closely related and connects to some of the experiments shown in T6.6. Finally, the development of this task has allowed the organization of the "**Computational Memorability of Imagery**" special session[22] at the 2023 International Conference on Content-based Multimedia Indexing.[23]

[17] https://mad2022.aimultimedialab.ro/

[18] https://mad2023.idmt.fraunhofer.de/

[19] https://www.veraai.eu/

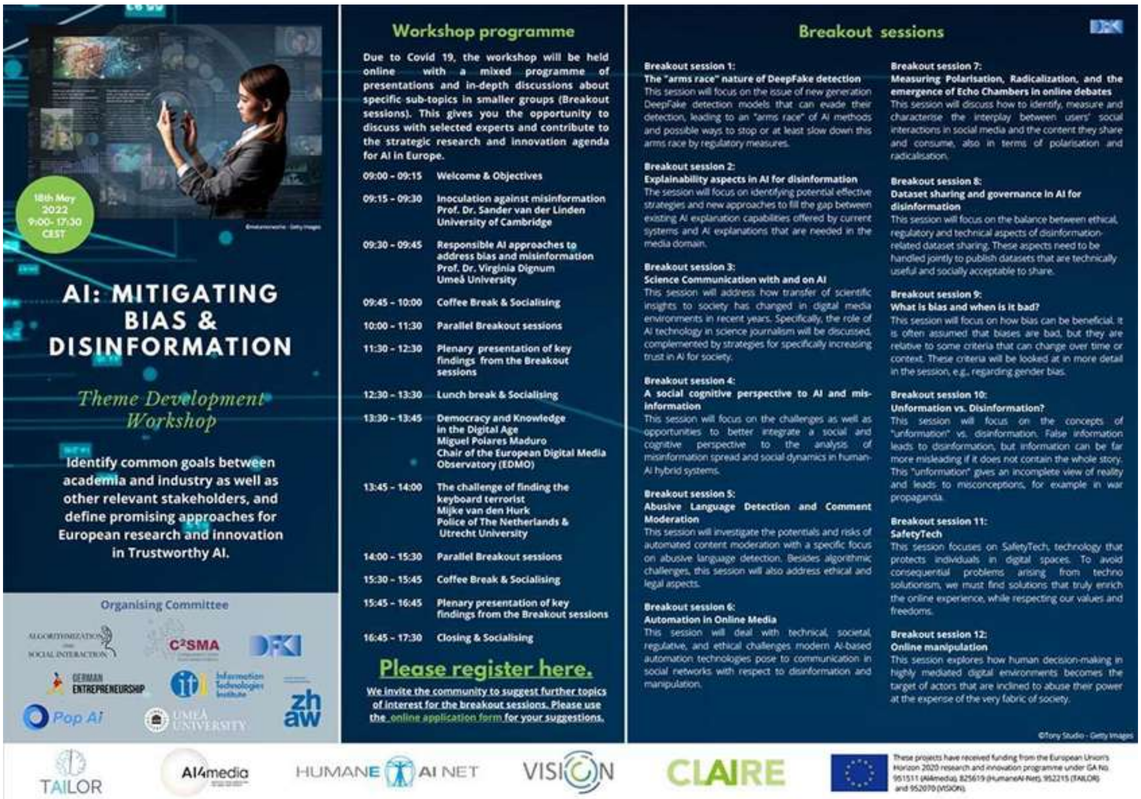
[20] https://dl.acm.org/pb-assets/static_journal_pages/tomm/pdf/CfP-TOMM-SI-RealisticSyntheticData-2023-1673383635690.pdf

[21] https://multimediaeval.github.io/editions/2023

[22] https://cbmi2023.org/special-sessions/

[23] https://cbmi2023.org/

The **Cross-cutting Theme Development Workshop on "AI: Mitigating Bias & Disinformation"**[24] was co-organised by AI4Media, Humane-AI-Net, TAILOR, and CLAIRE AISBL, under the lead of VISION CSA on May 18th, 2022, aiming to study several aspects related to finding common goals between industry professionals and academia in fighting and identifying bias and disinformation in the media landscape. Talks were held on topics like responsible AI approaches, inoculation against misinformation, and the challenges of identifying misinformation sources, while 12 breakout sessions were held, where experts from academia, media industry and politics discussed topics like the arms race of nature deepfake detection, explainable AI for misinformation detection, measuring radicalization, polarization and echo-chambers, abusive language detection, and incomplete information among others. Figure 66 presents an overview of the event program. The key findings of the workshop have been summarised in a public report.[25]



18th May 2022 9:00- 17:30 CEST

AI: MITIGATING BIAS & DISINFORMATION

Theme Development Workshop

Identify common goals between academia and industry as well as other relevant stakeholders, and define promising approaches for European research and innovation in Trustworthy AI.

Organising Committee

ALGORITHMIZATION AND SOCIAL INTERACTION
C²SMA
DFKI
GERMAN ENTREPRENEURSHIP
iti Information Technologies Institute
Pop AI
UMEÅ UNIVERSITY
zh aw

Workshop programme

Due to Covid 19, the workshop will be held online with a mixed programme of presentations and in-depth discussions about specific sub-topics in smaller groups (Breakout sessions). This gives you the opportunity to discuss with selected experts and contribute to the strategic research and innovation agenda for AI in Europe.

09:00 – 09:15	Welcome & Objectives
09:15 – 09:30	Inoculation against misinformation Prof. Dr. Sander van der Linden University of Cambridge
09:30 – 09:45	Responsible AI approaches to address bias and misinformation Prof. Dr. Virginia Dignum Umeå University
09:45 – 10:00	Coffee Break & Socialising
10:00 – 11:30	Parallel Breakout sessions
11:30 – 12:30	Plenary presentation of key findings from the Breakout sessions
12:30 – 13:30	Lunch break & Socialising
13:30 – 13:45	Democracy and Knowledge in the Digital Age Miguel Poiares Maduro Chair of the European Digital Media Observatory (EDMO)
13:45 – 14:00	The challenge of finding the keyboard terrorist Mijke van den Hurk Police of The Netherlands & Utrecht University
14:00 – 15:30	Parallel Breakout sessions
15:30 – 15:45	Coffee Break & Socialising
15:45 – 16:45	Plenary presentation of key findings from the Breakout sessions
16:45 – 17:30	Closing & Socialising

Please register here.

We invite the community to suggest further topics of interest for the breakout sessions. Please use the online application form for your suggestions.

Breakout sessions

Breakout session 1:
The "arms race" nature of DeepFake detection
This session will focus on the issue of new generation DeepFake detection models that can evade their detection, leading to an "arms race" of AI methods and possible ways to stop or at least slow down this arms race by regulatory measures.

Breakout session 2:
Explainability aspects in AI for disinformation
The session will focus on identifying potential effective strategies and new approaches to fill the gap between existing AI explanation capabilities offered by current systems and AI explanations that are needed in the media domain.

Breakout session 3:
Science Communication with and on AI
This session will address how transfer of scientific insights to society has changed in digital media environments in recent years. Specifically, the role of AI technology in science journalism will be discussed, complemented by strategies for specifically increasing trust in AI for society.

Breakout session 4:
A social cognitive perspective to AI and misinformation
This session will focus on the challenges as well as opportunities to better integrate a social and cognitive perspective to the analysis of misinformation spread and social dynamics in human-AI hybrid systems.

Breakout session 5:
Abusive Language Detection and Comment Moderation
This session will investigate the potentials and risks of automated content moderation with a specific focus on abusive language detection. Besides algorithmic challenges, this session will also address ethical and legal aspects.

Breakout session 6:
Automation in Online Media
This session will deal with technical, societal, regulative, and ethical challenges modern AI-based automation technologies pose to communication in social networks with respect to disinformation and manipulation.

Breakout session 7:
Measuring Polarisation, Radicalization, and the emergence of Echo Chambers in online debates
This session will discuss how to identify, measure and characterise the interplay between users' social interactions in social media and the content they share and consume, also in terms of polarisation and radicalisation.

Breakout session 8:
Dataset sharing and governance in AI for disinformation
This session will focus on the balance between ethical, regulatory and technical aspects of disinformation-related dataset sharing. These aspects need to be handled jointly to publish datasets that are technically useful and socially acceptable to share.

Breakout session 9:
What is bias and when is it bad?
This session will focus on how bias can be beneficial. It is often assumed that biases are bad, but they are relative to some criteria that can change over time or context. These criteria will be looked at in more detail in the session, e.g., regarding gender bias.

Breakout session 10:
Unformation vs. Disinformation?
This session will focus on the concepts of "unformation" vs. disinformation. False information leads to disinformation, but information can be far more misleading if it does not contain the whole story. This "unformation" gives an incomplete view of reality and leads to misconceptions, for example in war propaganda.

Breakout session 11:
SafetyTech
This session focuses on SafetyTech, technology that protects individuals in digital spaces. To avoid consequential problems arising from techno solutionism, we must find solutions that truly enrich the online experience, while respecting our values and freedoms.

Breakout session 12:
Online manipulation
This session explores how human decision-making in highly mediated digital environments becomes the target of actors that are inclined to abuse their power at the expense of the very fabric of society.

©Tony Studio - Getty Images

These projects have received funding from the European Union's Horizon 2020 research and innovation programme under GA No 951911 (AI4media), 825619 (HumaneAI-Net), 952215 (TAILOR) and 952070 (VISION).

Figure 66. Program of the Cross-cutting Theme Development Workshop on "AI: Mitigating Bias & Disinformation".

AI4Media together with EC-funded projects vera.ai, AI4TRUST, and TITAN - in cooperation with the European Commission - organised a hybrid event titled "**Meet the Future of AI - Countering sophisticated & advanced disinformation**"[26] in Brussels, on 29 June 2023. During the event, various aspects surrounding the development and use of AI and its relationship to the disinformation sphere were discussed. The event included four panels on: i) threats and

[24] https://www.vision4ai.eu/ai-mitigating-bias-disinformation/

[25] https://www.vision4ai.eu/wp-content/uploads/2023/01/Report-on-the-key-findings-from-the-Theme-Development-Workshop-_AI_-Mitigating-Bias-Disinformation.pdf

[26] https://agenda.euractiv.com/events/meet-future-ai-countering-sophisticated-advanced-disinformation-250623

opportunities of generative AI for mis- and disinformation; ii) policy implications and challenges to fight disinformation; iii) the role of critical thinking in addressing future AI tools to fight disinformation; iv) technological and strategic approaches to detecting and countering AI-generated content across four projects. CERTH presented AI4Media's work on AI against disinformation (i.e. the outcomes of Task 6.2 and use cases 1 and 2). The event attracted more than 90 in-person participants and more than one-hundred online attendees.

11 Summary and Conclusion

This deliverable provides and overview of the work and research conducted in WP6: *Human- and Society-centered AI* and presents the work of the partners in this WP, covering months 17 to 36, and representing a number of at least 32 conference and journal publications, as well as 18 publicly available repositories containing the associated methods, models and resources.

We show two different perspectives on the *Manipulation and synthetic content detection in multimedia* (T6.2), analyzing aspects related to both data creation and manipulation, and the detection of such data. We present experiments and work done in subjects like temporal and spatial considerations for content manipulation, gaze correction, and text-driven image manipulation, as well as manipulated content detection from three different perspectives: video- and image-based, audio-based, and text-based tweet detection.

We present the main directions in studying *Hybrid, privacy-enhanced recommendation* systems (T6.3), with work done in explainable knowledge graph-based news recommendations. We continue with the work done in *AI for healthier political debate* (T6.4), analyzing aspects related to sentiment classification in tweets, argumentative, propagandist and fallacious content, subjective and objective perspectives in political news, and other politically-related subjects. We show the advances made in the *Perception of hyper-local news* (T6.5), with contributions to domains related to health-related news items, YouTube knowledge communication channels and videos, local news items, and the evaluation of large language models in news article setups.

We continue with contributions brought to *Measuring and predicting user perception of social media* (T6.6), analyzing works related to media memorability, the prediction of affective content, and user perception of echo-chambers. Next, we analyze the work done in the study of *Real-life effects of private content sharing*, describing methods that predict the way social media items can negatively impact users in certain scenarios. Finally, we look at some *Other relevant activities* related to WP6 like the organisation of special issues, special sessions, workshops, etc.

Many of the contributions presented in this deliverable have already been published at various relevant conferences or in journals, and additionally we present the relevant repositories containing software, datasets, or other resources, that are publicly available for each technical contribution. We also analyze the relevance of each technical contribution, both to AI4Media use cases, at to the media industry as a whole. The next update of this deliverable will be the final one (D6.4 – Final version of Human- and Society-centered AI algorithms), presenting the final version of AI techniques developed in the context of WP6.

References

[1] Y. Mirsky and W. Lee, “The creation and detection of deepfakes: A survey,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.

[2] H. Alqahtani, M. Kavakli-Thorne, and G. Kumar, “Applications of generative adversarial networks (gans): An updated review,” *Archives of Computational Methods in Engineering*, vol. 28, pp. 525–552, 2021.

[3] Y. Yu, G. Liu, and J.-M. Odobez, “Improving few-shot user-specific gaze adaptation via gaze redirection synthesis,” in *CVPR*, 2019.

[4] Y. Yu and J.-M. Odobez, “Unsupervised representation learning for gaze estimation,” in *CVPR*, 2020.

[5] J. Zhang, J. Chen, H. Tang, W. Wang, Y. Yan, E. Sangineto, and N. Sebe, “Dual in-painting model for unsupervised gaze correction and animation in the wild,” in *ACM MM*, 2020.

[6] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,” in *CVPR*, 2018.

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *ICCV*, 2017.

[8] Z. He, A. Spurr, X. Zhang, and O. Hilliges, “Photo-realistic monocular gaze redirection using generative adversarial networks,” in *ICCV*, 2019.

[9] K. A. Funes Mora, F. Monay, and J.-M. Odobez, “Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras,” in *Proceedings of the Symposium on Eye Tracking Research and Applications*, 2014.

[10] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, “Gaze locking: Passive eye contact detection for human-object interaction,” in *ACM symposium on User interface software and technology*, 2013.

[11] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, “Mpiigaze: Real-world dataset and deep appearance-based gaze estimation,” *TPAMI*, 2017.

[12] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, “Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation,” *ECCV*, 2020.

[13] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, “Gaze360: Physically unconstrained gaze estimation in the wild,” in *CVPR*, 2019.

[14] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, “Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation,” in *ECCV*, 2020.

[15] X. Cai, B. Chen, J. Zeng, J. Zhang, Y. Sun, X. Wang, Z. Ji, X. Liu, X. Chen, and S. Shan, “Gaze estimation with an ensemble of four architectures,” *arXiv preprint arXiv:2107.01980*, 2021.

[16] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *CVPR*, 2015.

[17] D. E. King, “Dlib-ml: A machine learning toolkit,” *JMLR*, 2009.

[18] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, “Maskgan: Towards diverse and interactive facial image manipulation,” in *CVPR*, 2020.

[19] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, “Attgan: Facial attribute editing by only changing what you want,” *TIP*, 2019.

[20] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, “Stgan: A unified selective transfer network for arbitrary image attribute editing,” in *CVPR*, 2019.

[21] P.-W. Wu, Y.-J. Lin, C.-H. Chang, E. Y. Chang, and S.-W. Liao, “Relgan: Multi-domain image-to-image translation via relative attributes,” in *CVPR*, 2019.

[22] J. gi Kwak, D. K. Han, and H. Ko, “Cafe-gan: Arbitrary face attribute editing with complementary attention feature,” in *ECCV*, 2020.

[23] W. Chu, Y. Tai, C. Wang, J. Li, F. Huang, and R. Ji, “Sscgan: Facial attribute editing via style skip connections,” in *ECCV*, 2020.

[24] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, 2003.

[25] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *CVPR*, 2018.

[26] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a nash equilibrium,” in *NeurIPS*, 2017.

[27] J. Zhang, J. Chen, H. Tang, E. Sangineto, P. Wu, Y. Yan, N. Sebe, and W. Wang, “Unsupervised high-resolution portrait gaze correction and animation,” *IEEE Transactions on Image Processing*, vol. 31, no. 7, pp. 5272–5286, 2022.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *NeurIPS*, 2014.

[29] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[30] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[31] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, “Multimodal unsupervised image-to-image translation,” in *ECCV*, 2018.

[32] R. Abdal, P. Zhu, N. Mitra, and P. Wonka, “Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows,” *ACM TOG*, 2020.

[33] Y. Shen, J. Gu, X. Tang, and B. Zhou, “Interpreting the latent space of gans for semantic face editing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9243–9252.

[34] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, “Designing an encoder for stylegan image manipulation,” *ACM TOG*, 2021.

[35] Y. Shen and B. Zhou, “Closed-form factorization of latent semantics in gans,” in *CVPR*, 2021.

[36] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in *CVPR*, 2019.

[37] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein gans,” in *NeurIPS*, 2017.

[38] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” in *ICLR*, 2018.

[39] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *CVPR*, 2019.

[40] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan,” in *CVPR*, 2020.

[41] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, “Training generative adversarial networks with limited data,” in *NeurIPS*, 2020.

[42] Z. He, M. Kan, and S. Shan, “Eigengan: Layer-wise eigen-learning for gans,” in *ICCV*, 2021.

[43] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, “Alias-free generative adversarial networks,” in *NeurIPS*, 2021.

[44] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y.-L. Yang, “Hologan: Unsupervised learning of 3d representations from natural images,” in *ICCV*, 2019.

[45] T. Nguyen-Phuoc, C. Richardt, L. Mai, Y.-L. Yang, and N. Mitra, “Blockgan: Learning 3d object-aware scene representations from unlabelled images,” in *NeurIPS*, 2020.

[46] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum, “Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling,” in *NeurIPS*, 2016.

[47] M. Gadelha, S. Maji, and R. Wang, “3d shape induction from 2d views of multiple objects,” in *3DV*, 2017.

[48] J.-Y. Zhu, Z. Zhang, C. Zhang, J. Wu, A. Torralba, J. B. Tenenbaum, and W. T. Freeman, “Visual object networks: Image generation with disentangled 3D representations,” in *NeurIPS*, 2018.

[49] X. Chen, D. Cohen-Or, B. Chen, and N. J. Mitra, “Towards a neural graphics pipeline for controllable image generation,” *CGF*, 2021.

[50] Y. Liao, K. Schwarz, L. Mescheder, and A. Geiger, “Towards unsupervised learning of generative models for 3d controllable image synthesis,” in *CVPR*, 2020.

[51] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis,” in *Computer Vision – European Conference of Computer Vision (ECCV)*, 2020.

[52] A. Jain, M. Tancik, and P. Abbeel, “Putting nerf on a diet: Semantically consistent few-shot view synthesis,” in *ICCV*, 2021.

[53] T. DeVries, M. A. Bautista, N. Srivastava, G. W. Taylor, and J. M. Susskind, “Unconstrained scene generation with locally conditioned radiance fields,” in *ICCV*, 2021.

[54] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou, “Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans,” in *CVPR*, 2021.

[55] S. Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, H. Bao, and X. Zhou, “Animatable neural radiance fields for human body modeling,” in *ICCV*, 2021.

[56] C. Reiser, S. Peng, Y. Liao, and A. Geiger, “Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps,” in *ICCV*, 2021.

[57] E. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, "Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis," in *arXiv*, 2020.

[58] K. Schwarz, Y. Liao, M. Niemeyer, and A. Geiger, "Graf: Generative radiance fields for 3d-aware image synthesis," in *NeurIPS*, 2020.

[59] M. Niemeyer and A. Geiger, "GIRAFFE: Representing scenes as compositional generative neural feature fields," in *CVPR*, 2021.

[60] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312 .6114*, 2013.

[61] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *CVPR*, 2016.

[62] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NeurIPS*, 2017.

[63] J. Zhang, E. Sangineto, H. Tang, A. Siarohin, Z. Zhong, N. Sebe, and W. Wang, "3d-aware semantic-guided generative model for human synthesis," in *European Conference on Computer Vision (ECCV)*, 2022.

[64] H. Tang, S. Bai, P. Torr, and N. Sebe, "Bipartite graph reasoning gans for person pose and facial image synthesis," *International Journal of Computer Vision*, vol. 131, no. 3, pp. 644–658, 2023.

[65] Y. Liu, M. De Nadai, D. Cai, H. Li, X. Alameda-Pineda, N. Sebe, and B. Lepri, "Describe what to change: A text-guided unsupervised image-to-image translation approach," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1357–1365.

[66] B. Wang and C. R. Ponce, "A geometric analysis of deep generative image models and its applications," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=GH7QRzUDdXG.

[67] R. Abdal, P. Zhu, N. J. Mitra, and P. Wonka, "Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 3, pp. 1–21, 2021.

[68] X. Li, S. Zhang, J. Hu, L. Cao, X. Hong, X. Mao, F. Huang, Y. Wu, and R. Ji, "Image-to-image translation via hierarchical style disentanglement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8639–8648.

[69] A. Gabbay, N. Cohen, and Y. Hoshen, "An image is worth more than a thousand words: Towards disentanglement in the wild," in *Neural Information Processing Systems (NeurIPS)*, 2021.

[70] A. Gabbay and Y. Hoshen, "Scaling-up disentanglement for image translation," in *International Conference on Computer Vision (ICCV)*, 2021.

[71] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: Discovering interpretable gan controls," in *Proc. NeurIPS*, 2020.

[72] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in gans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1532–1540.

[73] A. Voynov and A. Babenko, "Unsupervised discovery of interpretable directions in the gan latent space," in *International Conference on Machine Learning*, PMLR, 2020, pp. 9786–9796.

[74] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[75] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[76] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 2085–2094.

[77] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *CoRR*, vol. abs/2103.00020, 2021. arXiv: 2103.00020. [Online]. Available: https://arxiv.org/abs/2103.00020.

[78] R. Gal, O. Patashnik, H. Maron, G. Chechik, and D. Cohen-Or, *Stylegan-nada: Clip-guided domain adaptation of image generators*, 2021. arXiv: 2108.00946 [cs.CV].

[79] D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or, "Pivotal tuning for latent-based editing of real images," *arXiv preprint arXiv:2106.05744*, 2021.

[80] Z. Wu, D. Lischinski, and E. Shechtman, "Stylespace analysis: Disentangled controls for stylegan image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 863–12 872.

[81] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "Tedigan: Text-guided diverse face image generation and manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2256–2265.

[82] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "Towards open-world text-guided face image generation and manipulation," *arxiv preprint arxiv: 2104.08910*, 2021.

[83] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, Dec. 2015.

[84] Z. Xu, T. Lin, H. Tang, F. Li, D. He, N. Sebe, R. Timofte, L. Van Gool, and E. Ding, "Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[85] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 28, 2015, pp. 91–99. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

[86] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 318–10 327.

[87] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhöfer, C. Lassner, and C. Theobalt, "Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video," in *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[88] J. Ost, F. Mannan, N. Thuerey, J. Knodt, and F. Heide, "Neural scene graphs for dynamic scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2856–2865.

[89] W. Yuan, Z. Lv, T. Schmidt, and S. Lovegrove, “Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[90] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine, “Stochastic adversarial video prediction,” *ArXiv*, vol. abs/1804.01523, 2018.

[91] M. Kumar, M. Babaeizadeh, D. Erhan, C. Finn, S. Levine, L. Dinh, and D. Kingma, “Videoflow: A conditional flow-based model for stochastic video generation,” in *International Conference on Learning Representations (ICLR)*, 2020. [Online]. Available: https://openreview.net/forum?id=rJgUfTEYvH.

[92] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, “Mocogan: Decomposing motion and content for video generation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[93] Y. Tian, J. Ren, M. Chai, K. Olszewski, X. Peng, D. N. Metaxas, and S. Tulyakov, “A good image generator is what you need for high-resolution video synthesis,” in *International Conference on Learning Representations (ICLR)*, 2021.

[94] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “Animating arbitrary objects via deep motion transfer,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[95] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “First order motion model for image animation,” in *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[96] A. Siarohin, O. Woodford, J. Ren, M. Chai, and S. Tulyakov, “Motion representations for articulated animation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[97] W. Menapace, S. Lathuiliere, S. Tulyakov, A. Siarohin, and E. Ricci, “Playable video generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 061–10 070.

[98] *Minecraft*, https://www.minecraft.net, Accessed: 2021-11-12.

[99] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[100] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017, pp. 6626–6637. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074%20926d-Paper.pdf.

[101] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, “Towards accurate generative models of video: A new metric & challenges,” *arXiv preprint arXiv:1812.01717*, 2018.

[102] W. Menapace, A. Siarohin, C. Theobalt, V. Golyanik, S. Tulyakov, S. Lathuiliere, and E. Ricci, “Playable environments: Video manipulation in space and time,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’22)*, 2022.

[103] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, and Z. Liu, "Forgerynet: A versatile benchmark for comprehensive forgery analysis," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4358–4367. DOI: 10.1109/CVPR46437.2021.00434.

[104] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *ICCV*, Oct. 2019.

[105] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. DOI: 10.1109/CVPR42600.2020.00028.

[106] S. Baxevanakis, G. Kordopatis-Zilos, P. Galopoulos, L. Apostolidis, K. Levacher, I. B. Schlicht, D. Teyssou, I. Kompatsiaris, and S. Papadopoulos, "The mever deepfake detection service: Lessons learnt from developing and deploying in the wild," in *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, 2022.

[107] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7824–7833. DOI: 10.1109/CVPR.2019.00802.

[108] I. Petrov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, R. Luis, J. Jiang, S. Zhang, P. Wu, B. Zhou, and W. Zhang, "Deepfacelab: A simple, flexible and extensible face swapping framework," *ArXiv*, vol. abs/2005.05535, 2020.

[109] Y. Deng, J. Yang, D. Chen, F. Wen, and X. Tong, "Disentangled and controllable face image generation via 3d imitative-contrastive learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5153–5162. DOI: 10.1109/CVPR42600.2020.00520.

[110] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala, *Text-based editing of talking-head video*, Jun. 2019.

[111] Y. Jo and J. Park, "Sc-fegan: Face editing generative adversarial network with user's sketch and color," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1745–1753. DOI: 10.1109/ICCV.2019.00183.

[112] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8107–8116. DOI: 10.1109/CVPR42600.2020.00813.

[113] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5548–5557. DOI: 10.1109/CVPR42600.2020.00559.

[114] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5073–5082. DOI: 10.1109/CVPR42600.2020.00512.

[115] Y. Nirkin, Y. Keller, and T. Hassner, "Fsgan: Subject agnostic face swapping and reenactment," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7183–7192. DOI: 10.1109/ICCV.2019.00728.

[116] A. Siarohin, S. Lathuiliere, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *ArXiv*, vol. abs/2003.00196, 2019.

[117] D. A. Coccomini, G. K. Zilos, G. Amato, R. Caldelli, F. Falchi, S. Papadopoulos, and C. Gennaro, *Mintime: Multi-identity size-invariant video deepfake detection*, 2022. DOI: 10.48550/ARXIV.2211.10996. [Online]. Available: https://arxiv.org/abs/2211.10996.

[118] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *ICCV*, 2021. DOI: 10.1109/ICCV48922.2021.01477.

[119] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining efficientnet and vision transformers for video deepfake detection," in *Image Analysis and Processing (ICIAP 2022) - Part III*, Lecce, Italy: Springer, 2022, ISBN: 978-3-031-06432-6. DOI: 10.1007/978-3-031-06433-3_19. [Online]. Available: https://doi.org/10.1007/978-3-031-06433-3_19.

[120] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *arXiv preprint arXiv:2102.11126*, 2021.

[121] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, 2016. DOI: 10.1109/LSP.2016.2603342.

[122] S. Seferbekov. "Dfdc 1st place solution." (2020), [Online]. Available: https://github.com/selimsef/dfdc_deepfake_challenge.

[123] A. Buslaev, A. Parinov, E. Khvedchenya, V. I. Iglovikov, and A. A. Kalinin, "Albumentations: fast and flexible image augmentations," *ArXiv e-prints*, 2018. eprint: 1809.06839.

[124] D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, and G. Amato, "Cross-forgery analysis of vision transformers and cnns for deepfake image detection," in *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, ser. MAD '22, Newark, NJ, USA: ACM, 2022, ISBN: 9781450392426. DOI: 10.1145/3512732.3533582. [Online]. Available: https://doi.org/10.1145/3512732.3533582.

[125] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[126] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," 2021. DOI: 10.48550/ARXIV.2104.00298. [Online]. Available: https://arxiv.org/abs/2104.00298.

[127] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6559–6568.

[128] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5001–5010.

[129] A. Haliassos, R. Mira, S. Petridis, and M. Pantic, "Leveraging real talking faces via self-supervision for robust forgery detection," in *CVPR*, 2022. DOI: 10.1109/CVPR52688.2022.01453.

[130] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

[131] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[132] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *International conference on machine learning*, PMLR, 2020, pp. 3247–3258.

[133] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 772–781.

[134] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 317–16 326.

[135] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George, "Deepfake detection using spatiotemporal convolutional networks," *arXiv preprint arXiv:2006.14749*, 2020.

[136] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5039–5049.

[137] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 044–15 054.

[138] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 720–18 729.

[139] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[140] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3207–3216.

[141] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton-Ferrer, "The deepfake detection challenge (dfdc) preview dataset," *ArXiv*, vol. abs/1910.08854, 2019.

[142] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset.," *arXiv: Computer Vision and Pattern Recognition*, 2020.

[143] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[144] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 023–15 033.

[145] A. V. Nadimpalli and A. Rattani, "On improving cross-dataset generalization of deepfake detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 91–99.

[146] A. Jain, P. Korshunov, and S. Marcel, "Improving generalization of deepfake detection by training for attribution," in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2021, pp. 1–6.

[147] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information fusion*, vol. 76, pp. 243–297, 2021.

[148] A. N. Angelopoulos, S. Bates, *et al.*, "Conformal prediction: A gentle introduction," *Foundations and Trends® in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023.

[149] X. Wang, H. Guo, S. Hu, M.-C. Chang, and S. Lyu, "Gan-generated faces detection: A survey and new perspectives," *arXiv preprint arXiv:2202.07145*, 2022.

[150] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, "Are gan generated images easy to detect? a critical analysis of the state-of-the-art," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2021.

[151] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[152] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," in *Proceedings of the Advances in Neural Information Processing Systems*, 2021.

[153] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[154] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[155] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, 2017.

[156] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "De-fake: Detection and attribution of fake images generated by text-to-image diffusion models," *arXiv preprint arXiv:2210.06998*, 2022.

[157] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," *arXiv preprint arXiv:2211.00680*, 2022.

[158] T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani, "A comprehensive survey on model compression and acceleration," *Artificial Intelligence Review*, vol. 53, no. 7, 2020.

[159] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "Model compression and acceleration for deep neural networks: The principles, progress, and challenges," *IEEE Signal Processing Magazine*, vol. 35, no. 1, 2018.

[160] P. Dogoulis, G. Kordopatis-Zilos, I. Kompatsiaris, and S. Papadopoulos, "Improving synthetically generated image detection in cross-concept settings," in *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, 2023, pp. 28–35.

[161] I. Sarridis, C. Koutlis, S. Papadopoulos, and I. Kompatsiaris, "Indistill: Transferring knowledge from pruned intermediate layers," *arXiv preprint arXiv:2205.10003*, 2022.

[162] L. Cuccovillo, C. Papastergiopoulos, A. Vafeiadis, A. Yaroshchuk, P. Aichroth, K. Votis, and D. Tzovaras, "Open challenges in synthetic speech detection," in *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, 2022, pp. 1–6.

[163] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[164] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, *Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech*, 2020. arXiv: 1911.01601 [eess.AS].

[165] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, *Conformer: Convolution-augmented transformer for speech recognition*, 2020. arXiv: 2005.08100 [eess.AS].

[166] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[167] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T.-Y. Liu, "Understanding and improving transformer from a multi-particle dynamic system point of view," *arXiv preprint arXiv:1906.02762*, 2019.

[168] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[169] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*, ISCA, Sep. 2019. DOI: 10.21437/interspeech.2019-2680. [Online]. Available: https://doi.org/10.21437%2Finterspeech.2019-2680.

[170] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal processing letters*, vol. 24, no. 3, pp. 279–283, 2017.

[171] B. McFee, J. Salamon, and J. P. Bello, *Adaptive pooling operators for weakly labeled sound event detection*, 2018. arXiv: 1804.10070 [cs.SD].

[172] M. Huzaifah, *Comparison of time-frequency representations for environmental sound classification using convolutional neural networks*, 2017. arXiv: 1706.07156 [cs.CV].

[173] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "Biometric score normalization: Eer-based score normalization techniques," in *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems*, IEEE, 2007, pp. 1–9.

[174] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, *T-dcf: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification*, 2019. arXiv: 1804.09618 [eess.AS].

[175] C. Krätzer, A. Oermann, J. Dittmann, and A. Lang, "Digital audio forensics: A first practical evaluation on microphone and environment classification," in *ACM Workshop on Multimedia & Security (MM&Sec)*, Dallas, TX, USA, 2007, pp. 63–74.

[176] L. Cuccovillo, S. Mann, M. Tagliasacchi, and P. Aichroth, "Audio tampering detection via microphone classification," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Pula, Italy, Sep. 2013, pp. 177–182.

[177] L. Cuccovillo and P. Aichroth, “Open-set microphone classification via blind channel analysis,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shangai, China, 2016, pp. 2074–2078.

[178] D. Luo, P. Korus, and J. Huang, “Band energy difference for source attribution in audio forensics,” *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 13, no. 9, pp. 2179–2189, 2018.

[179] M. A. Qamhan, H. Altaheri, A. H. Meftah, G. Muhammad, and Y. A. Alotaibi, “Digital audio forensics: Microphone and environment classification using deep learning,” *IEEE Access*, vol. 9, pp. 62 719–62 733, 2021.

[180] L. Cuccovillo, A. Giganti, P. Bestagini, P. Aichroth, and S. Tubaro, “Spectral denoising for microphone classification,” in *ACM International Workshop on Multimedia AI against Disinformation (MAD)*, Newark, NJ, USA, 2022, pp. 10–17. DOI: 10.1145/3512732.3533586.

[181] L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992. DOI: 10.1016/0167-2789(92)90242-F.

[182] A. Buades, B. Coll, and J.-M. Morel, “A non-local algorithm for image denoising,” in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, San Diego, CA, USA: IEEE, 2005, pp. 60–65. DOI: 10.1109/CVPR.2005.38.

[183] C. Tomasi and R. Manduchi, “Bilateral filtering for gray and color images,” in *IEEE International Conference on Computer Vision*, Bombay, India: IEEE, 1998, pp. 839–846. DOI: 10.1109/ICCV.1998.710815.

[184] S. G. Chang, B. Yu, and M. Vetterli, “Adaptive wavelet thresholding for image denoising and compression,” *IEEE Transactions on Image Processing*, vol. 9, no. 9, pp. 1532–1546, 2000. DOI: 10.1109/83.862633.

[185] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising,” *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[186] A. Défossez, G. Synnaeve, and Y. Adi, “Real time speech enhancement in the waveform domain,” in *Annual Conference of the International Speech Communication Association (Interspeech)*, Shanghai, China: ISCA, 2020, pp. 3291–3295. DOI: 10.21437/Interspeech.2020-2409.

[187] C. Kotropoulos and S. Samaras, “Mobile phone identification using recorded speech signals,” in *IEEE International Conference on Digital Signal Processing*, Hong Kong, China, 2014, pp. 586–591.

[188] A. Giganti, L. Cuccovillo, P. Bestagini, P. Aichroth, and S. Tubaro, “Speaker-independent microphone identification in noisy conditions,” in *European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, 2022, pp. 1047–1051. DOI: 10.23919/EUSIPCO55093.2022.9909800.

[189] N. D. Gaubitch, M. Brookes, and P. A. Naylor, “Blind channel magnitude response estimation in speech using spectrum classification,” *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 21, no. 10, pp. 2162–2171, 2013. DOI: 10.1109/TASL.2013.2270406.

[190] D. U. Leonzio, L. Cuccovillo, P. Bestagini, M. Marcon, P. Aichroth, and S. Tubaro, “Audio splicing detection and localization based on acquisition device traces,” *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 18, pp. 4157–4172, 2023. DOI: 10.1109/TIFS.2023.3293415.

[191] G. Baldini, I. Amerini, and C. Gentile, “Microphone identification using convolutional neural networks,” *IEEE Sensors Letters*, vol. 21, no. 10, pp. 1–4, 2019. DOI: 10.1109/TASL.2013.2270406.

[192] D. Capoferri, C. Borrelli, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, “Speech audio splicing detection and localization exploiting reverberation cues,” in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2020, pp. 1–6. DOI: 10.1109/WIFS49906.2020.9360900.

[193] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[194] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, *Language models are unsupervised multitask learners*, 2019.

[195] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, “CTRL: A conditional transformer language model for controllable generation,” *arXiv:1909.05858 [cs]*, 2019. arXiv: 1909.05858.

[196] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[197] J. Rodríguez-Ruiz, J. I. Mata-Sánchez, R. Monroy, O. Loyola-González, and A. López-Cuevas, “A one-class classification approach for bot detection on twitter,” *Computers & Security*, vol. 91, 2020.

[198] N. Chavoshi, H. Hamooni, and A. Mueen, “DeBot: Twitter bot detection via warped correlation,” in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, IEEE, 2016, pp. 817–822.

[199] M. Heidari, J. H. Jones, and O. Uzuner, “Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter,” in *2020 International Conference on Data Mining Workshops (ICDMW)*, IEEE, 2020, pp. 480–487.

[200] T. Fagni, F. Falchi, M. Gambini, A. Martella, and M. Tesconi, “TweepFake: About detecting deepfake tweets,” *PLOS ONE*, vol. 16, no. 5, 2021.

[201] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *Roberta: A robustly optimized bert pretraining approach*, 2019. arXiv: 1907.11692 [cs.CL].

[202] J. Tourille, B. Sow, and A. Popescu, “Automatic detection of bot-generated tweets,” in *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, 2022, pp. 44–51.

[203] G. Balloccu, L. Boratto, G. Fenu, and M. Marras, “Hands on explainable recommender systems with knowledge graphs,” in *Proceedings of the 16th ACM Conference on Recommender Systems*, 2022, pp. 710–713.

[204] Y. Xian, Z. Fu, S. Muthukrishnan, G. De Melo, and Y. Zhang, "Reinforcement knowledge graph reasoning for explainable recommendation," in *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, 2019, pp. 285–294.

[205] Y. Xian, Z. Fu, H. Zhao, Y. Ge, X. Chen, Q. Huang, S. Geng, Z. Qin, G. De Melo, *et al.*, "Cafe: Coarse-to-fine neural symbolic reasoning for explainable recommendation," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1645–1654.

[206] B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," vol. 11, Jan. 2010.

[207] A. Bermingham and A. Smeaton, "On using Twitter to monitor political sentiment and predict election results," in *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, Chiang Mai, Thailand, Nov. 2011, pp. 2–10. [Online]. Available: https://aclanthology.org/W11-3702.

[208] B. Bansal and S. Srivastava, "On predicting elections with hybrid topic based sentiment analysis of tweets," *Procedia Computer Science*, vol. 135, pp. 346–353, 2018.

[209] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," *Proceedings of the International AAAI Conference on Web and Social Media*, 2010.

[210] L. Wang and J. Q. Gan, "Prediction of the 2017 french election based on twitter data analysis," in *2017 9th Computer Science and Electronic Engineering (CEEC)*, 2017, pp. 89–93. DOI: 10.1109/CEEC.2017.8101605.

[211] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017.

[212] M. I. Patsiouras E. Koroni I. and P. I., "Greekpolitics: Sentiment analysis on greek politically charged tweets," *European Signal Processing Conference (EUSIPCO)*, 2023.

[213] I. Pitas and A. Kaimakamidis, "Political tweet sentiment analysis for public opinion polling," *Technical Report*,

[214] L. Zewen, L. Fan, Y. Wenjie, P. Shouheng, and P. Jun, "A survey of Convolutional Neural Networks: Analysis, applications, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, pp. 6999–7019, 12 2021.

[215] A. Gillioz, J. Casas, E. Mugellini, and O. Abou K., "Overview of the Transformer-based models for NLP tasks," in *Proceedings of the Conference on Computer Science and Information Systems (FedCSIS)*, IEEE, 2020.

[216] P. Bojanowski, E. Grave, A. Joulin, and M. T, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[217] Q. Xie, Z. Dai, E. Hovy, M. T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2020.

[218] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2015.

[219] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019.

[220] S. Haddadan, E. Cabrio, and S. Villata, "Yes, we can! mining arguments in 50 years of us presidential campaign debates," in *Proceedings of ACL 2019*, 2019, pp. 4684–4690.

[221] I. Habernal, P. Pauli, and I. Gurevych, "Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices," in *Proceedings of LREC 2018*, ELRA, 2018. [Online]. Available: https://aclanthology.org/L18-1526.

[222] I. Habernal, H. Wachsmuth, I. Gurevych, and B. Stein, "Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation," en, in *Proceedings of NAACL 2018*, ACL, 2018, pp. 386–396. DOI: 10.18653/v1/N18-1036. [Online]. Available: http://aclweb.org/anthology/N18-1036 (visited on 03/19/2021).

[223] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov, "Fine-grained analysis of propaganda in news article," in *Proceedings of EMNLP-IJCNLP 2019*, ACL, 2019, pp. 5636–5646.

[224] D. Walton, *Informal Fallacies : Towards a Theory of Argument of Criticisms*. Philadelphia: John Benjamins Publishing Company, 1987, ISBN: 978-90-272-7890-6. [Online]. Available: http://ebookcentral.proquest.com/lib/unilu-ebooks/detail.action?docID=802001.

[225] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North*, Association for Computational Linguistics, 2019, pp. 4171–4186.

[226] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv:1907.11692 [cs]*, Jul. 26, 2019. arXiv: 1907.11692. [Online]. Available: http://arxiv.org/abs/1907.11692 (visited on 04/02/2021).

[227] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *CoRR*, vol. abs/2004.05150, 2020. arXiv: 2004.05150. [Online]. Available: https://arxiv.org/abs/2004.05150.

[228] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *CoRR*, vol. abs/1901.02860, 2019. arXiv: 1901.02860. [Online]. Available: http://arxiv.org/abs/1901.02860.

[229] V. Vorakitphan, E. Cabrio, and S. Villata, ""Don't discuss": Investigating Semantic and Argumentative Features for Supervised Propagandist Message Detection and Classification," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, Held Online: INCOMA Ltd., Sep. 2021, pp. 1498–1507. [Online]. Available: https://aclanthology.org/2021.ranlp-1.168.

[230] M. Boudry, F. Paglieri, and M. Pigliucci, "The fake, the flimsy, and the fallacious: demarcating arguments in real life," eng, *ARGUMENTATION*, vol. 29, no. 4, 431–456, 2015, ISSN: 0920-427X. [Online]. Available: %7Bhttp://dx.doi.org/10.1007/s10503-015-9359-1%7D.

[231] P. Goffredo, S. Haddadan, V. Vorakitphan, E. Cabrio, and S. Villata, "Fallacious argument classification in political debates," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed., Main Track, International Joint Conferences on Artificial Intelligence Organization, Jul. 2022, pp. 4143–4149. DOI: 10.24963/ijcai.2022/575. [Online]. Available: https://doi.org/10.24963/ijcai.2022/575.

[232] H. D. Lasswell, “Propaganda technique in the world war,” 1938. eprint: https://hdl.handle.net/2027/mdp.39015000379902.

[233] H. Koppang, “Social influence by manipulation: A definition and case of propaganda,” *Middle East Critique*, vol. 18, pp. 117–143, 2009.

[234] J. P. Dillard and M. Pfau, *The Persuasion Handbook: Developments in Theory and Practice*. Sage Publications, Inc., 2009, pp. 371–380, ISBN: 9781412976046.

[235] L. Longpre, E. Durmus, and C. Cardie, “Persuasion of the undecided: Language vs. the listener,” in *Proceedings of the 6th Workshop on Argument Mining*, Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 167–176. DOI: 10.18653/v1/W19-4519. [Online]. Available: https://www.aclweb.org/anthology/W19-4519.

[236] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov, “Fine-Grained Analysis of Propaganda in News Article,” en, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 5635–5645. DOI: 10.18653/v1/D19-1565. [Online]. Available: https://www.aclweb.org/anthology/D19-1565 (visited on 04/26/2022).

[237] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov, “SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles,” in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1377–1414. DOI: 10.18653/v1/2020.semeval-1.186. [Online]. Available: https://aclanthology.org/2020.semeval-1.186.

[238] P. J. Stone, D. C. Dunphy, and M. S. Smith, “The general inquirer: A computer approach to content analysis.,” 1966.

[239] M. Brysbaert, A. Warriner, and V. Kuperman, “Concreteness ratings for 40 thousand generally known english word lemmas,” *Behavior research methods*, vol. 46, Oct. 2013. DOI: 10.3758/s13428-013-0403-5.

[240] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada: Association for Computational Linguistics, Oct. 2005, pp. 347–354. [Online]. Available: https://www.aclweb.org/anthology/H05-1044.

[241] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://www.aclweb.org/anthology/N19-1423.

[242] S. Baccianella, A. Esuli, and F. Sebastiani, “SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining,” in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta: European Language Resources Association (ELRA), May 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf.

[243] O. Araque, L. Gatti, J. Staiano, and M. Guerini, “Depechemood++: A bilingual emotion lexicon built through simple yet powerful techniques,” *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[244] A. Warriner, V. Kuperman, and M. Brysbaert, “Norms of valence, arousal, and dominance for 13,915 english lemmas,” *Behavior research methods*, vol. 45, Feb. 2013. DOI: 10.3758/s13428-012-0314-x.

[245] S. Feng, J. S. Kang, P. Kuznetsova, and Y. Choi, “Connotation lexicon: A dash of sentiment beneath the surface meaning,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 1774–1784. [Online]. Available: https://www.aclweb.org/anthology/P13-1174.

[246] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts, “A computational approach to politeness with application to social factors,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 250–259. [Online]. Available: https://www.aclweb.org/anthology/P13-1025.

[247] Y. Tsvetkov, L. Boytsov, A. Gershman, E. Nyberg, and C. Dyer, “Metaphor detection with cross-lingual model transfer,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 248–258. DOI: 10.3115/v1/P14-1024. [Online]. Available: https://www.aclweb.org/anthology/P14-1024.

[248] A. Ferreira Cruz, G. Rocha, and H. Lopes Cardoso, “On sentence representations for propaganda detection: From handcrafted features to word embeddings,” in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 107–112. DOI: 10.18653/v1/D19-5015. [Online]. Available: https://www.aclweb.org/anthology/D19-5015.

[249] C. Stab and I. Gurevych, “Annotating argument components and relations in persuasive essays,” in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 1501–1510. [Online]. Available: https://www.aclweb.org/anthology/C14-1142.

[250] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html.

[251] N. Mapes, A. White, R. Medury, and S. Dua, “Divisive language and propaganda detection using multi-head attention transformers with deep learning BERT-based language models for binary classification,” in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 103–106. DOI: 10.18653/v1/D19-5014. [Online]. Available: https://www.aclweb.org/anthology/D19-5014.

[252] S. Yoosuf and Y. Yang, “Fine-grained propaganda detection with fine-tuned BERT,” in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 87–91. DOI: 10.18653/v1/D19-5011. [Online]. Available: https://www.aclweb.org/anthology/D19-5011.

[253] S. Yoosuf and Y. Yang, "Fine-grained propaganda detection with fine-tuned BERT," in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 87–91. DOI: 10.18653/v1/D19-5011. [Online]. Available: https://www.aclweb.org/anthology/D19-5011.

[254] D. Jurkiewicz, Ł. Borchmann, I. Kosmala, and F. Graliński, "ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1415–1424. [Online]. Available: https://www.aclweb.org/anthology/2020.semeval-1.187.

[255] J.-P. Cointet, D. Cardon, A. Mogoutov, B. Ooghe-Tabanou, G. Plique, and P. Morales, "Uncovering the structure of the french media ecosystem," *arXiv preprint arXiv:2107.12073*, 2021.

[256] J. Cagé, M. Hengel, N. Hervé, and C. Urvoy, "Hosting media bias: Evidence from the universe of french broadcasts, 2002-2020," *SSRN Electronic Journal*, 2022. DOI: 10.2139/SSRN.4036211.

[257] F. Caravaca, J. González-Cabañas, Á. Cuevas, and R. Cuevas, "Estimating ideology and polarization in european countries using facebook data," *EPJ Data Science*, vol. 11, no. 1, p. 56, 2022.

[258] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "An open-source speaker gender detection framework for monitoring gender equality," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 5214–5218.

[259] F. Hamborg, K. Donnay, P. Merlo, *et al.*, "Newsmtsc: A dataset for (multi-) target-dependent sentiment classification in political news articles," Association for Computational Linguistics (ACL), 2021.

[260] H. Döring and P. Manow, "Parliament and government composition database (parlgov)," *An infrastructure for empirical information on parties, elections and governments in modern democracies. Version*, vol. 12, no. 10, 2012.

[261] F. T. Asr, M. Mazraeh, A. Lopes, V. Gautam, J. Gonzales, P. Rao, and M. Taboada, "The gender gap tracker: Using natural language processing to measure gender bias in media," *PloS one*, vol. 16, no. 1, e0245533, 2021.

[262] T. Venturini, T. Blanke, and K. De Pryck, "Similarity sampling by machine learning a social science experiment with artificial intelligence and ipcc leadership," 2023.

[263] D. Alonso del Barrio and D. Gatica-Perez, "How did europe's press cover covid-19 vaccination news? a five-country analysis," in *Proc. ACM International Workshop on Multimedia AI against Disinformation (MAD)*, Newark, 2022, pp. 52–59.

[264] T.-T. Phan, C. Michoud, L. Volpato, M. del Rio Carral, and D. Gatica-Perez, "Health talk: Understanding practices of popular professional youtubers," in *Proc. Int. Conf. on Mobile and Ubiquitous Multimedia (MUM)*, Lisbon, 2022.

[265] H. Semetko and P. Valkenburg, "Framing european politics: A content analysis of press and television news," *Journal of Communication*, vol. 50, pp. 93–109, Jun. 2000. DOI: 10.1111/j.1460-2466.2000.tb02843.x.

[266] D. Alonso del Barrio and D. Gatica-Perez, “Examining european press coverage of the covid-19 no-vax movement: An nlp framework,” in *Proc. ACM International Workshop on Multimedia AI against Disinformation (MAD)*, Thessaloniki, Greece, 2023, pp. 52–59. DOI: 10.1145/3592572.3592845.

[267] M. S. Islam, A.-H. M. Kamal, A. Kabir, D. L. Southern, S. H. Khan, S. M. Hasan, T. Sarkar, S. Sharmin, S. Das, T. Roy, *et al.*, “Covid-19 vaccine rumors and conspiracy theories: The need for cognitive inoculation against misinformation to improve vaccine adherence,” *PloS one*, vol. 16, no. 5, e0251605, 2021.

[268] D. Freeman, B. S. Loe, A. Chadwick, C. Vaccari, F. Waite, L. Rosebrock, L. Jenner, A. Petit, S. Lewandowsky, S. Vanderslott, *et al.*, “Covid-19 vaccine hesitancy in the uk: The oxford coronavirus explanations, attitudes, and narratives survey (oceans) ii,” *Psychological medicine*, vol. 52, no. 14, pp. 3127–3141, 2022.

[269] R. Fletcher, A. Cornia, L. Graves, and R. K. Nielsen, “Measuring the reach of” fake news” and online disinformation in europe,” *Australasian Policing*, vol. 10, no. 2, 2018.

[270] M. Dupuis, K. Chhor, and N. Ly, “Misinformation and disinformation in the era of covid-19: The role of primary information sources and the development of attitudes toward vaccination,” in *Proceedings of the 22nd Annual Conference on Information Technology Education*, 2021, pp. 105–110.

[271] R. Piltch-Loeb, E. Savoia, B. Goldberg, B. Hughes, T. Verhey, J. Kayyem, C. Miller-Idriss, and M. Testa, “Examining the effect of information channel on covid-19 vaccine acceptance,” *Plos one*, vol. 16, no. 5, e0251095, 2021.

[272] M. Grootendorst, *Bertopic: Neural topic modeling with a class-based tf-idf procedure*, 2022. arXiv: 2203.05794 [cs.CL].

[273] D. Alonso del Barrio and D. Gatica-Perez, “Framing the news: From human perception to large language model inferences,” in *ACM International Conference on Multimedia Retrieval (ICMR)*, Thessaloniki, Greece, 2023, pp. 627–635. DOI: 10.1145/3591106.3592278.

[274] P. Pourashraf and B. Mobasher, “Using user’s local context to support local news,” in *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, ACM, Jul. 2022. DOI: 10.1145/3511047.3537657. [Online]. Available: https://doi.org/10.1145%2F3511047.3537657.

[275] K. E. Matsa and K. Worden. “Local newspapers fact sheet,” Pew Research Center’s Journalism Project. (), [Online]. Available: https://www.pewresearch.org/journalism/fact-sheet/local-newspapers/ (visited on 08/02/2023).

[276] A. Mitchell. “2. publics around the world follow national and local news more closely than international,” Pew Research Center’s Global Attitudes Project. (Jan. 11, 2018), [Online]. Available: https://www.pewresearch.org/global/2018/01/11/publics-around-the-world-follow-national-and-local-news-more-closely-than-international/ (visited on 08/02/2023).

[277] N. Pignard-Cheynel and L. Amigo, “(re)connecting with audiences. an overview of audience-inclusion initiatives in european french-speaking local news media,” *Journalism*, Apr. 25, 2023, Publisher: SAGE Publications, ISSN: 1464-8849. DOI: 10.1177/14648849231173299. [Online]. Available: https://doi.org/10.1177/14648849231173299 (visited on 05/09/2023).

[278] A. Mitchell. "Publics globally want unbiased news coverage, but are divided on whether their news media deliver," Pew Research Center's Global Attitudes Project. (Jan. 11, 2018), [Online]. Available: https://www.pewresearch.org/global/2018/01/11/publics-globally-want-unbiased-news-coverage-but-are-divided-on-whether-their-news-media-deliver/ (visited on 08/02/2023).

[279] "Home," ESH Médias. (), [Online]. Available: https://www.eshmedias.ch/ (visited on 08/02/2023).

[280] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020. DOI: 10.5281/zenodo.1212303.

[281] L. Shen, *LexicalRichness: A small module to compute textual lexical richness*, 2022. DOI: 10.5281/zenodo.6607007. [Online]. Available: https://github.com/LSYS/lexicalrichness.

[282] *py-readability-metrics: Score the readability of text using popular readability formulas and metrics*, 2019. [Online]. Available: https://github.com/cdimascio/py-readability-metrics.

[283] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[284] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Nov. 2019. [Online]. Available: http://arxiv.org/abs/1908.10084.

[285] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot, "Camembert: A tasty french language mode," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[286] S. Terragni, E. Fersini, B. G. Galuzzi, P. Tropeano, and A. Candelieri, "OCTIS: Comparing and optimizing topic models is simple!" In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Apr. 2021, pp. 263–270. [Online]. Available: https://www.aclweb.org/anthology/2021.eacl-demos.31.

[287] H. Xia, H. X. Ng, Z. Chen, and J. Hollan, "Millions and billions of views: Understanding popular science and knowledge communication on video-sharing platforms," in *Proceedings of the Ninth ACM Conference on Learning @ Scale*, 2022, pp. 163–174.

[288] H. Kim and D. Gatica-Perez, "Referencing in YouTube Knowledge Communication Videos," in *Proc. ACM International Conference on Interactive Media Experiences (IMX) (to appear)*, 2023. [Online]. Available: https://publications.idiap.ch/attachments/papers/2023/HeKim_ACMIMX_2023.pdf.

[289] M. G. Constantin, M. Redi, G. Zen, and B. Ionescu, "Computational understanding of visual interestingness beyond semantics: Literature survey and analysis of covariates," *ACM Computing Surveys (CSUR)*, vol. 52, no. 2, pp. 1–37, 2019.

[290] M. G. Constantin, L.-D. Ştefan, B. Ionescu, N. Q. Duong, C.-H. Demarty, and M. Sjöberg, "Visual interestingness prediction: A benchmark framework and literature review," *International Journal of Computer Vision*, vol. 129, pp. 1526–1550, 2021.

[291] L. Sweeney, M. G. Constantin, C.-H. Demarty, C. Fosco, A. G. S. de Herrera, S. Halder, G. Healy, B. Ionescu, A. Matran-Fernandez, A. F. Smeaton, *et al.*, "Overview of the mediaeval 2022 predicting video memorability task," in *Working Notes Proceedings of the MediaEval 2022 Workshop*, 2023.

[292] C. Guinaudeau and A. G. Xalabarder, “Textual analysis for video memorability prediction,” in *Working Notes Proceedings of the MediaEval 2022 Workshop*, 2023.

[293] R. G. Prakash, J. Bhuvana, E. A. Chodisetty, A. Mukesh, and T. Mirnalinee, “Multi-model estimators and ensemble-based regressors for predicting video memorability,” in *Working Notes Proceedings of the MediaEval 2022 Workshop*, 2023.

[294] M. G. Constantin and B. Ionescu, “Aimultimedialab at mediaeval 2022: Predicting media memorability using video vision transformers and augmented memorable moments,” in *Working Notes Proceedings of the MediaEval 2022 Workshop*, 2023.

[295] M. M. Ali Usmani, S. Zahid, and M. A. Tahir, “Modelling of video memorability using ensemble learning and transformers,” in *Working Notes Proceedings of the MediaEval 2022 Workshop*, 2023.

[296] J. Wan, X. He, and P. Shi, “An iris image quality assessment method based on laplacian of gaussian operation.,” in *MVA*, 2007, pp. 248–251.

[297] D. Hasler and S. E. Suesstrunk, “Measuring colorfulness in natural images,” in *Human vision and electronic imaging VIII*, SPIE, vol. 5007, 2003, pp. 87–95.

[298] Y. Ke, X. Tang, and F. Jing, “The design of high-level features for photo quality assessment,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, IEEE, vol. 1, 2006, pp. 419–426.

[299] G. Farnebäck, “Two-frame motion estimation based on polynomial expansion,” in *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, Springer, 2003, pp. 363–370.

[300] M. Dogariu, L.-D. Stefan, M. G. Constantin, and B. Ionescu, “Human-object interaction: Application to abandoned luggage detection in video surveillance scenarios,” in *2020 13th International Conference on Communications (COMM)*, IEEE, 2020, pp. 157–160.

[301] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, “Mask R-CNN,” in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, IEEE Computer Society, 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.322. [Online]. Available: https://doi.org/10.1109/ICCV.2017.322.

[302] M. G. Constantin, M. Dogariu, A.-C. Jitaru, and B. Ionescu, “Assessing the difficulty of predicting media memorability,” in *20th International Conference on Content-based Multimedia Indexing*, 2023.

[303] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, and A. Oliva, “Multimodal memorability: Modeling effects of semantics and decay on video memorability,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham: Springer International Publishing, 2020, pp. 223–240, ISBN: 978-3-030-58517-4.

[304] M. G. Constantin and B. Ionescu, “Using vision transformers and memorable moments for the prediction of video memorability,” in *MediaEval Multimedia Benchmark Workshop Working Notes*, Dec. 2021. [Online]. Available: http://ceur-ws.org/Vol-3181/.

[305] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *International Conference on Machine Learning*, PMLR, 2021, pp. 10 347–10 357.

[306] H. Bao, L. Dong, and F. Wei, “Beit: Bert pre-training of image transformers,” *arXiv preprint arXiv:2106.08254*, 2021.

[307] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836–6846.

[308] E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, and N. B. Hussin, "Affective computing in education: A systematic review and future research," *Computers & Education*, vol. 142, p. 103 649, 2019, ISSN: 0360-1315. DOI: https://doi.org/10.1016/j.compedu.2019.103649.

[309] E. Smith, E. A. Storch, H. Lavretsky, J. L. Cummings, and H. A. Eyre, "Affective computing for brain health disorders," in *Handbook of Computational Neurodegeneration*, P. Vlamos, I. S. Kotsireas, and I. Tarnanas, Eds. Cham: Springer International Publishing, 2020, pp. 1–14, ISBN: 978-3-319-75479-6. DOI: 10.1007/978-3-319-75479-6_36-1.

[310] P. Ekman and W. V. Friesen, *Facial action coding system: Investigator's guide*. Consulting Psychologists Press, 1978.

[311] J. A. Russell, "A circumplex model of affect.," en, *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980, ISSN: 0022-3514. DOI: 10.1037/h0077714.

[312] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krähenbühl, and R. Girshick, "Long-Term Feature Banks for Detailed Video Understanding," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.

[313] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The OMG-Emotion Behavior Dataset," in *2018 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2018. DOI: 10.1109/IJCNN.2018.8489099.

[314] J. A. Miranda Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018, ISSN: 1949-3045, 2371-9850. DOI: 10.1109/TAFFC.2018.2884461.

[315] W. Mou, H. Gunes, and I. Patras, "Alone versus In-a-group: A Multi-modal Framework for Automatic Affect Recognition," en, *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 2, Jun. 2019, ISSN: 1551-6857, 1551-6865. DOI: 10.1145/3321509.

[316] N. M. Foteinopoulou and I. Patras, "Learning from label relationships in human affect," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 80–89.

[317] W. Ceron, M.-F. De-Lima-Santos, and M. G. Quiles, "Fake news agenda in the era of COVID-19: Identifying trends through fact-checking content," *Online Social Networks and Media*, vol. 21, p. 100 116, Jan. 2021, ISSN: 24686964. DOI: 10.1016/j.osnem.2020.100116. arXiv: 2012.11004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2468696420300562%20http://arxiv.org/abs/2012.11004%20http://dx.doi.org/10.1016/j.osnem.2020.100116.

[318] W. Ceron, G. G. Sanseverino, M.-F. De-Lima-Santos, and M. G. Quiles, "COVID-19 fake news diffusion across Latin America," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 47, Dec. 2021, ISSN: 1869-5450. DOI: 10.1007/s13278-021-00753-z. [Online]. Available: https://link.springer.com/article/10.1007/s13278-021-00753-z.

[319] D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, "Uncovering Coordinated Networks on Social Media: Methods and Case Studies," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, pp. 455–466, May 2021, ISSN: 2334-0770. DOI: 10.1609/ICWSM.V15I1.18075. arXiv: 2001.05658. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/18075.

[320] A. Gruzd, P. Mai, and F. B. Soares, “How coordinated link sharing behavior and partisans’ narrative framing fan the spread of COVID-19 misinformation and conspiracy theories,” *Social Network Analysis and Mining*, vol. 12, no. 1, p. 118, Dec. 2022, ISSN: 1869-5450. DOI: 10.1007/s13278-022-00948-y. [Online]. Available: https://link.springer.com/article/10.1007/s13278-022-00948-y%20https://link.springer.com/10.1007/s13278-022-00948-y.

[321] N. Gleicher, *Coordinated inauthentic behavior explained*, Dec. 2018. [Online]. Available: https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/%7B%5C%%7D0Ahttps://newsroom.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/ (visited on 10/11/2021).

[322] F. Giglietto, N. Righetti, L. Rossi, and G. Marino, “It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections,” *Information Communication and Society*, vol. 23, no. 6, pp. 867–891, 2020, ISSN: 14684462. DOI: 10.1080/1369118X.2020.1739732. [Online]. Available: https://www.tandfonline.com/action/journalInformation?journalCode=rics20.

[323] M. F. De-Lima-Santos and W. Ceron, “Disinformation Echo-chambers on Facebook,” in *Fighting Fake Fact*, P. Seitz, M. Eisenegger, and M. M. Bergman, Eds., 1st, Basel, 2023.

[324] M.-F. De-Lima-Santos and W. Ceron, “Coordinated Amplification, Coordinated Inauthentic Behavior, Orchestrated Campaigns? A Systematic Literature Review of Coordinated Inauthentic Content on Online Social Networks,” in *Mapping Lies in the Global Media Sphere*, T. Filibeli and M. Ö. Özbek, Eds., 1st, London: Taylor & Francis, 2024, ch. 10.

[325] Facebook, *Facebook privacy checkup*, https://www.facebook.com/help/443357099140264, Accessed: 2021-11-10, 2021.

[326] P. C. Bauer, F. Gerdon, F. Keusch, F. Kreuter, and D. Vannette, “Did the gdpr increase trust in data collectors? evidence from observational and experimental data,” *Information, Communication & Society*, vol. 0, no. 0, pp. 1–21, 2021.

[327] Y. Zhang, T. Wang, and C. Hsu, “The effects of voluntary gdpr adoption and the readability of privacy statements on customers’ information disclosure intention and trust,” *Journal of Intellectual Capital*, vol. 21, no. 2, pp. 145–163, Jan. 2020, ISSN: 1469-1930. DOI: 10.1108/JIC-05-2019-0113. [Online]. Available: https://doi.org/10.1108/JIC-05-2019-0113.

[328] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, June 18-22, 2018*, Salt Lake City, UT, USA: Computer Vision Foundation / IEEE Computer Society, 2018, pp. 4510–4520. DOI: 10.1109/CVPR.2018.00474. [Online]. Available: http://openaccess.thecvf.com/content%5C_cvpr%5C_2018/html/Sandler%5C_MobileNetV2%5C_Inverted%5C_Residuals%5C_CVPR%5C_2018%5C_paper.html.

[329] M. Tan, R. Pang, and Q. V. Le, “Efficientdet: Scalable and efficient object detection,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, June 13-19, 2020*, Seattle, WA, USA: Computer Vision Foundation / IEEE, 2020, pp. 10 778–10 787. DOI: 10.1109/CVPR42600.2020.01079. [Online]. Available: https://openaccess.thecvf.com/content%5C_CVPR%5C_2020/html/Tan%5C_EfficientDet%5C_Scalable%5C_and%5C_Efficient%5C_Object%5C_Detection%5C_CVPR%5C_2020%5C_paper.html.

[330] E. Spyromitros-Xioufis, S. Papadopoulos, A. Popescu, and Y. Kompatsiaris, "Personalized privacy-aware image classification," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '16, New York, New York, USA: Association for Computing Machinery, 2016, pp. 71–78, ISBN: 9781450343596. DOI: 10.1145/2911996.2912018. [Online]. Available: https://doi.org/10.1145/2911996.2912018.

[331] B. Ionescu, H. Müller, R. Peteri, A. Ben Abacha, D. Demner-Fushman, S. Hasan, M. Sarrouti, O. Pelka, C. Friedrich, A. Herrera, J. Jacutprakart, V. Kovalev, S. Kozlovski, V. Liauchuk, Y. Dicente Cid, J. Chamberlain, A. Clark, A. Campello, H. Moustahfid, and A. Popescu, "The 2021 imageclef benchmark: Multimedia retrieval in medical, nature, internet and social media applications," *Lecture Notes in Computer Science*, vol. 0, no. 0, Jan. 2021.

[332] V. Nguyen, A. Popescu, and J. Deshayes-Chossart, "Unveiling real-life effects of online photo sharing," *CoRR*, vol. abs/2012.13180, 2020. arXiv: 2012.13180. [Online]. Available: https://arxiv.org/abs/2012.13180.

[333] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009. DOI: 10.1109/TNN.2008.2005605. [Online]. Available: https://doi.org/10.1109/TNN.2008.2005605.

[334] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. F. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, Ç. Gülçehre, H. F. Song, A. J. Ballard, J. Gilmer, G. E. Dahl, A. Vaswani, K. R. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu, "Relational inductive biases, deep learning, and graph networks," *CoRR*, vol. abs/1806.01261, 2018. arXiv: 1806.01261. [Online]. Available: http://arxiv.org/abs/1806.01261.

[335] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[336] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter, "Auto-sklearn 2.0: The next generation," *CoRR*, vol. abs/2007.04074, 2020. arXiv: 2007.04074. [Online]. Available: https://arxiv.org/abs/2007.04074.

[337] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[338] J. Cohen, *Statistical power analysis for the behavioral sciences*. New York: Academic press, 2013.