# D6.4

# Final Version of Human- and Society-centered AI algorithms

| | |
|---|---|
| **Project Title** | Artificial Intelligence for the Media and Society |
| **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

info@ai4media.eu www.ai4media.eu

| | |
|---|---|
| **Deliverable title** | Final Version of Human- and Society-centered AI algorithms |
| **Deliverable number** | D6.4 |
| **Deliverable version** | 1.0 |
| **Previous version(s)** | - |
| **Contractual date of delivery** | 31/08/2024 |
| **Actual date of delivery** | 06/09/2024 |
| **Deliverable filename** | D6.4_AI4MEDIA.pdf |
| **Nature of deliverable** | Report |
| **Dissemination level** | Public |
| **Number of pages** | 152 |
| **Work Package** | WP6 |
| **Task(s)** | T6.2, T6.3, T6.4, T6.5, T6.6 |
| **Parner responsible** | CEA |
| **Author(s)** | Nicu Sebe (UNITN), Marco Formentini (UNITN), Thomas Köllmer (FhG-IDMT), David Alonso del Barrio (IDIAP), Victor Bros (IDIAP), Daniel Gatica-Perez (IDIAP), Christoforos Papastergiopoulos (CERTH), John Violos (CERTH), Symeon Papadopoulos (CERTH), Spiros Baxevanakis (CERTH), Christos Koutlis (CERTH), Nikolaos Giatsoglou (CERTH), Yiannis Kompatsiaris(CERTH), Ioannis Patras (QMUL), Ioannis Maniadis Metaxas (QMUL), Mihai Gabriel Constantin (UNSTPB), Cristian Stanciu (UNSTPB), Roberto Caldelli (UNIFI), Noémie Krack (KUL), Anna Schjøtt Hansen (UvA), Tobias Blanke (UvA), Natali Helberger (UvA), Pierpaolo Goffredo (UCA), Mariana Chaves (UCA), Serena Villata (UCA), Rasa Bočytė (NISV), Adrian Popescu (CEA) |
| **Editor** | Adrian Popescu (CEA) |
| **Project Officer** | Evangelia Markidou |

| | |
|---|---|
| **Abstract** | This document presents the final outcomes of the research on human- and society-centered AI algorithms, reporting the progress in tasks T6.2, T6.3, T6.4, T6.5, and T6.6 during the M37-M48 period. Specifically, this document builds upon and presents the updates of the work presented in deliverables D6.1 and D6.3. 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 present ongoing work and conclusions. |
| **Keywords** | Artificial intelligence, media, synthetic content creation, synthetic content detection, deepfake, manipulation detection, online political debate, privacy-aware content recommendation, hyper-local news, user perception measurement, content moderation |

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

# Copyright

## Contributors

| NAME | ORGANIZATION |
|---|---|
| Nicu Sebe | UNITN |
| Marco Formentini | UNITN |
| Thomas Köllmer | FhG-IDMT |
| David Alonso del Barrio | IDIAP |
| Victor Bros | IDIAP |
| Daniel Gatica-Perez | IDIAP |
| Christoforos Papastergiopoulos | CERTH |
| John Violos | CERTH |
| Symeon Papadopoulos | CERTH |
| Spiros Baxevanakis | CERTH |
| Christos Koutlis | CERTH |
| Nikolaos Giatsoglou | CERTH |
| Yiannis Kompatsiaris | CERTH |
| Ioannis Patras | QMUL |
| Ioannis Maniadis Metaxas | QMUL |
| Mihai Gabriel Constantin | UNSTPB |
| Cristian Stanciu | UNSTPB |
| Roberto Caldelli | UNIFI |
| Noémie Krack | KUL |
| Anna Schjøtt Hansen | UvA |
| Tobias Blanke | UvA |
| Natali Helberger | UvA |
| Pierpaolo Goffredo | UCA |
| Mariana Chaves | UCA |
| Serena Villata | UCA |
| Rasa Bočytė | NISV |
| Adrian Popescu | CEA |

## Peer Reviews

| NAME | ORGANIZATION |
|---|---|
| Chris Georgiev | IMG |
| Fabrizio Sebastiani | CNR |

# Revision History

| Version | Date | Reviewer | Modifications |
|---|---|---|---|
| 1 | 10.07.2024 | Adrian Popescu | First draft with contributions from all partners |
| 2 | 25.07.2024 | Adrian Popescu | Second draft with contributions from all partners |
| 3 | 30.07.2024 | Adrian Popescu | Draft sent to internal reviewers |
| 4 | 01.09.2024 | Adrian Popescu | Updated version based on internal reviews |
| 5 | 06.09.2024 | Adrian Popescu | Final version |

## Table of Abbreviations and Acronyms

| Abbreviation | Meaning |
|---|---|
| 1D | Uni-dimensional |
| 2D | Two-dimensional |
| 3D | Tri-dimensional |
| ACC | Accuracy |
| AE | Auto-encoder |
| AI | Artificial Intelligence |
| AP | Average Precision |
| API | Application Programming Interface |
| AUC | Area Under Curve |
| BA | Balanced Accuracy |
| BERT | Bidirectional Encoder Representations from Transformers |
| BiLSTM | Bidirectional Long Short-Term Memory |
| CC | Common Crawl |
| CIT | Cloth Interactive Transformer |
| CLIP | Contrastive Language-Image Pretraining |
| CLS | Classify Token |
| CNN | Convolutional Neural Network |
| CPU | Central Processing Unit |
| CUDA | Compute Unified Device Architecture |
| CV | Computer Vision |
| DF | Deep Fake |
| DFER | Dynamic Facial Expression Recognition |
| DNN | Deep Neural Network |
| DoF | Depth of Field |
| DPM | Diffusion Probabilistic Models |
| DSA | Digital Services Act |
| DSP | Digital Signal Processing |
| EC | European Commission |
| ECGAN | Edge-guided Generative Adversarial Network |
| EDMO | European Digital Media Observatory |
| EU | European Union |
| EER | Equal Error Rate |
| F1 | F1-score |
| FAR | False Acceptance Rate |
| FER | Facial Expression Recognition |
| FFN | Feed-Forward Networks |
| FFT | Fast Fourier Transform |

| | |
|---|---|
| FID | Fréchet Inception Distance |
| FLOP | Floating-point Operations per Second |
| FPS | Frames per Second |
| FS | Face Swap |
| FSH | Face Shift |
| FT | Fine Tuning |
| FVD | Fréchet Video Distance |
| GAN | Generative Adversarial Network |
| GauGaN | Gaugain GAN |
| GenAI | Generative AI |
| GMAC | Giga Multiply-Add Operations per Second |
| GNN | Graph Neural Network |
| GPT | Generative Pre-trained Transformer |
| GPU | Graphical Processing Unit |
| GSD | Global Surface Descriptor |
| HQ | High Quality |
| IFFT | Inverse Fast Fourier Transform |
| IPCC | Intergovernmental Panel on Climate Change |
| IoU | Intersection over Union |
| IS | Inception Score |
| JS | Jensen–Shannon divergence |
| KD | Knowledge Distillation |
| KL | Kullback-Leibler |
| LLM | Large Language Model |
| LMM | Large Multimedia Model |
| LPIPS | Learned Perceptual Image Patch Similarity |
| LSTM | Long-Short Term Memory |
| MAD | Multimedia Against Disinformation |
| MAE | Mean Absolute Error |
| mAP | mean Average Precision |
| mIoU | mean Intersection-over-Union |
| MLM | Masket Language Modeling |
| MLP | Multi-Layer Perceptron |
| MSE | Mean Squared Error |
| MTCNN | Multi-Task Cascaded Convolutional Neural Network |
| NER | Named Entity Recognition |
| NeRF | Neural Radiance Field |
| NTXent | Normalized Temperature-scaled Cross Entropy Loss |
| NLP | Natural Language Processing |
| OSN | Online Social Network |

| | |
|---|---|
| P | Precision |
| PE | Playable Environment |
| PIL | Python Imaging Library |
| PSNR | Peak Signal to Noise Ratio |
| PVG | Playable Video Generation |
| PVM | Predicting Video Memorability |
| PVT | Pyramid Vision Transformer |
| RGB | Red Green Blue |
| RL | Reinforcement Learning |
| RINE | Representations from Intermediate Encoder-blocks |
| ROC | Receiver Operating Characteristic |
| SGAN | Semantic-guided Generative Model |
| SGD | Stochastic Gradient Descent |
| SNR | Signal to Noise Ratio |
| SOTA | State of the Art |
| SRCC | Spearman's Rank Correlation Coefficient |
| SSIM | Structural Similarity |
| STFT | Short-Time Fourier Transform |
| SupCon | Supervised Constrastive Loss |
| SVM | Support Vector Machine |
| TIE | Trainable Importance Estimator |
| TPS | Thin-plate Spline |
| t-SNE | t-distributed Stochastic Neighbor Embedding |
| TTS | Text-to-Speech |
| UAR | Unweighted Average Recall |
| UC | Use Case |
| URL | Uniform Resource Locator |
| UMAP | Uniform Manifold Approximation and Projection |
| US | United States |
| VAD | Valence, Arousal, Domination |
| VAE | Variational Auto-Encoder |
| VC | Voice Conversion |
| ViT | Vision Transformer |
| ViViT | Video Vision Transformer |
| VTON | Virtual try-on |
| VLM | Visual-Language Model |
| VLOP | Very Large Online Platform |
| VLP | Vision-Language Pre-training |
| WAR | Weighted Average Recall |
| 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 final status of Work Package 6 (WP6): *Human and Society-centered AI*, presenting the results attained and the work published between months 37 (September 2023) and 48 (August 2024) of the AI4Media project. WP6 has an integral role in the AI4Media project, creating citizen- and society-oriented methods and algorithms that can then fuel the use cases developed in WP8. This deliverable presents progress in Tasks 6.2-6.6, covering a number of different topics, ranging from deepfake detection to social media analysis to user perception of media and more. Note that the T6.7 work was completed in the previous periods and has been documented in D6.3.

Sections 5 and 6 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 cloth interactive virtual try-on, semantic image synthesis, and human reaction generation (Section 5). The second addresses synthetic and manipulated content detection, presenting techniques for different modalities, i.e. for video- and image-based detection, and audio-based detection (Section 6).

Section 7 presents the work covered in Task 6.3: *Hybrid, privacy-enhanced recommendation*, describing an LLM-based zero-shot learning approach that detects opinions in news to diversify recommendations. This method has the potential to diversify the recommended news and thus contribute to countering filter bubbles.

Section 8 describes the work done in Task 6.4: *AI for Healthier Political Debate*, focusing on the development of AI models to analyse political debates. The contributions explore the objective and subjective analysis of political news in two EU countries, the reassembling of digital archives focused on international politics to elicit marginalized entities and viewpoints, and the detection and classification of fallacies in political debates.

Section 9 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 frame analysis in Dutch TV, local news analysis from the perspective of migrants, local news analysis in German and French-speaking regions of Switzerland, and an event-based analysis of Swiss news.

Progress in Task 6.6: *Measuring and predicting user perception of social media* is presented in Section 10. The works presented here include algorithms for memorability difficulty prediction and large-multimedia-model-based facial expression recognition in photos and videos.

The impact of WP6 work is highlighted by the fact that the contributions presented in this deliverable have led to publications in top-tier conferences and journals in the field. In summary, the work presented in this deliverable resulted in:

- 8 conference papers (ICLR, EMNLP, ACM IMX, BMVC) and 4 journal articles (IEEE TPAMI, IEEE ToM, Nature HSSC),
- 8 open-source software and tools publicly available (e.g., in GitHub),
- 4 contributions that are currently prepared for publication.

Some of the contributions were integrated into the AI4Media use cases, particularly for synthetic content detection and news analysis.

# 2. Introduction

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

**Generative AI** technologies make it possible to produce synthetic content at an unprecedented scale. Such content can be used as a complement to or instead of real training data whenever the latter 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 assist journalists with text writing and summarisation or 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 to counter such disinformation campaigns efficiently. The current contributions of AI4Media regarding synthetic data creation and data manipulation are presented in Section 5. The automatic detection of such data is discussed in Section 6. In total, four contributions were accepted to peer-reviewed conferences and four were accepted to peer-reviewed journals.

**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 that give users feedback about why content is recommended to them. The contributions of AI4Media regarding the usability of LLMs in content recommendation are discussed in Section 7.

A **healthy political debate** is a condition 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 on 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. The current contribution of AI4Media regarding a comprehensive analysis of political texts, the reuse of digital archives, and LLM-based fallacious argument detection are discussed in Section 8. One work was accepted to a peer-reviewed conference, one was accepted to a peer-reviewed journal, and one is currently under review.

News play an important role in shaping our understanding of the world. While many research efforts focus on automatic news analysis, local news is currently understudied. This is problematic insofar as local news provides up-to-date critical information about communities' lives. 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 analyze local news along different dimensions, which is important for professional and general public stakeholders. The current contributions of AI4Media regarding the **perception of hyper-local news** are presented in Section 9. One paper was accepted to a peer-reviewed conference, and two are under preparation.

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 this research area properly, inputs from different disciplines, including computer vision, linguistics, psychology, and sociology, are required. They were integrated into the proposed works, which address content memorability and adapting vision-language models for social media. Media professionals can use the proposed components to increase the impact of their content. The current contributions of AI4Media regarding the measuring and the prediction of user perception of social media are presented in Section 10. Two contributions were accepted to peer-reviewed conferences.

A part of the contributions proposed during the final period build on and extend the works discussed in D6.1 and D6.3, but also encompass new areas of interest covered by WP6. In particular, works on synthetic content generation and detection address challenges brought by the rapid development of generative AI and propose innovative answers to tackle them. The analysis of political debate is deeper and addresses understanding political news in different countries and languages, using digital archives to understand international politics, and effectively classifying fallacies using LLMs.

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 analyzing the political debate (T6.4) are 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 introduce innovative solutions to open topics, such as synthetic content generation and detection, the automatic analysis of political texts, or social media perception. These contributions led to a significant number of publications in leading journals and conferences in the respective areas of research.

# 3. Concise descriptions of the presented works

In the following, we briefly summarise the outcomes of each WP6 task for the period M37-M48. These works are then presented in detail in sections 4-10.

## 3.1. Policy recommendations for content moderation (Task 6.1)

### 3.1.1. Introduction

The rapid development of AI raises numerous questions regarding their use for content moderation, which has important practical implications given users' ubiquitous use of Web platforms. For instance, effective tools are needed to counter disinformation campaigns while also ensuring respect for freedom of speech. Such questions require a combination of legal, ethical, and technical competencies to enable a seamless and reasoned deployment of AI technologies. T6.1 work addresses open challenges in the area, with a focus on making the EU's regulatory framework understandable for citizens.

### 3.1.2. Overview

Subsection 4.1 discusses the a summarized version of the report on policy on content content whose aim is to clarify the challenges and opportunities of algorithmic content moderation, and its interplay with funamental rights. Subsection 4.2 presents examples of AI Media Observatory articles related to content moderation. Subsection 4.3 highlights the connection between content moderation and generative AI, and notably (1) the contribution to EU DisinfoLab factsheet on the topic, (2) a white paper on Generative AI and disinformation, (3) the contribution to a handbook on disinformation, and (4) the contribution to AI4Media's WP2 work. Subsection 4.4 explores the challenge of content moderation in the metaverse,in particular through the lens of the Digital Services Act.

## 3.2. Manipulation and synthetic content detection in multimedia (Task 6.2)

### 3.2.1. Introduction

AI advancements in multimedia processing, brought forth with the emergence of deep generative models, has allowed the synthesis and editing of hyperrealistic and high quality digital content. While this is certainly valuable, e.g., for the entertainment and gaming sector, the wide availability of these tools has spread concerns over their misuse. Deepfakes in particular, i.e., tampered videos of humans, have gathered significant criticism over their potential for harmful impersonation, namely, for fraud, propaganda, or disinformation. Due to these fears, the research community has focused intensely on the topic of deepfake detection, acknowledging that the same deep learning methods that made deepfakes possible are the most appropriate to address them. The problem of detection, however, is challenging as it is highly sensitive on the media quality, the preprocessing steps, and the data used for training the detectors.

Task 6.2 takes a holistic view on the topic of synthetic content by investigating AI methods for both generation and detection. Section 5 focuses on the generation methods presenting techniques for 1) content generation in a virtual try-on scenario, 2) generating photo-realistic images conditioned on pixel-level semantic labels, also known as semantic image synthesis, and 3) human reaction generation. Section 6 focuses on the detection methods, presenting a variety of techniques, ranging from novel model architectures, enriched objectives, and augmentation strategies.

#### 3.2.2. Overview

Subsection 5.1 introduces a novel two-stage Cloth Interactive Transformer (CIT) for the challenging virtual try-on task, which can model long-range interactive relations between the clothing and the person representations. Substantial empirical results on a public fashion dataset illustrate that the presented approach attains competitive virtual try-on performance.

Subsection 5.2 proposes a novel edge guided generative adversarial network with contrastive learning for the challenging semantic image synthesis task. This is the first attempt in the literature to explore the edge generation from semantic layouts and then utilize the generated edges to guide the generation of realistic images. Both qualitative and quantitative results show that the proposed approach is able to produce remarkably better results than existing baseline models regarding both visual fidelity and alignment with the input semantic layouts. Moreover, it can generate multi-modal images and edges, which have not been considered by existing state-of-the-art methods.

Subsection 5.3 proposes a novel Interaction Transformer framework for the challenging human reaction generation task. This is the first work that challenges the task of human reaction prediction given the action of the interacting skeleton using a Transformer based architecture. Extensive experiments on several datasets demonstrate the effectiveness of our method which is general and can be used to generate more complex and long-term interactions.

Subsection 6.1 presents the RINE method for synthetic (AI-generated) image detection. RINE leverages image representations extracted by intermediate layers of the large-scale pre-trained CLIP model, which are then processed by a lightweight MLP-based network that predicts the probability of an image being AI generated. The results not only show the superiority of the method compared to state-of-the-art but also increased generalization to unseen image generation models.

Subsection 6.2 presents our work on deepfake manipulation identification (attribution). In adopting manipulation attribution models, it is imperative that performance and reliability differences with the current norm - i.e. binary models - are thoroughly examined. We argue that a good attribution model should be able to correctly identify known manipulations across datasets. Our findings support the two following conclusions. First, deepfake attribution models with vanilla approaches generalize less effectively compared to binary counterparts. Incorporating contrastive methods alongside carefully crafted contrastive set-ups is helpful, especially in larger models. Second, the ability of attribution models to maintain their accuracy across datasets is heavily influenced by data quality. Training on high-quality realistic deepfakes drastically improves the resulting detection performance.

Subsection 6.3 proposes a training method for deepfake detectors that targets their generalizability based on adversarial augmentations. In more detail, a deepfake detector is trained jointly with a generator that creates manipulations that are progressively harder to detect. In this way, the detector becomes more robust and can generalize better to unseen forgeries. The results show that the proposed method outperforms the baseline detectors in standard metrics and robustness, indicating good potential for real-world use.

Subsection 6.4 presents SurFake, an AI model that detects inconsistencies in the geometrical surfaces of a scene, related to the camera acquisition process. This approach exploits the fact that a manipulation could distort the traces left by the camera acquisition, which are not visible to the human eye but are algorithmically detectable. The results show that the proposed surface-based features carry discriminative information for deepfake detection and that can be combined with features at the pixel level to further enhance detection accuracy.

Subsection 6.5 applies an emerging neural network architecture for vision, the MLP Mixer model, to synthetic audio detection. Besides highlighting the cross-modal potential of this architecture, spectral information is incorporated into the model to further enhance its detection performance. This work aspires to achieve a breakthrough in the mitigation of more audio deepfakes that become increasingly harder to detect with the continuous development of more sophisticated generation methods.

Subsection 6.6 focuses on the compression of deepfake detection models so that they can be deployed directly at the user end devices with tolerable decline in accuracy. Towards this goal, it builds and compares four prominent methods for model compression, namely, pruning, knowledge distillation, quantization, and low-rank factorization. In addition, it investigates the impact of transfer learning to the efficiency of model training by reusing the weights of generic pretrained vision models. The results show that knowledge distillation and pruning are the most effective techniques for model compression. In addition, transfer learning can enhance the performance of the compressed model, provided that they are optimized via careful design.

## 3.3. Hybrid, privacy-enhanced recommendation(Task 6.3)

### 3.3.1. Introduction

Recommender systems are easily one of the most used applications of the Internet. We define recommenders in a broad way, as mechanisms that present certain items to a certain user, in a way that it might be of use. This is met in more or less all media platforms and websites but also includes personalized search results.

There are also downsides. Many recommender systems rely heavily on user and usage data, that might compromise a user's privacy. Or they work a little *too good*, starting the effect that is commonly known as filter bubbles.

The goal of this task was to check for certain applications in the scope of AI4Media, if it is possible to counter those effects using techniques of content recommendation and explainable recommenders. We report the research outcomes of Task 6.3 in detail in Section 7

### 3.3.2. Overview

Compared to other tasks within the project, this task is relatively small and also got changed in scope due to the fact that there was no user and usage data available at the use cases side. While in the past deliverable, we focused on explainable recommendation to measure filter bubble effects and counter them, this time in subsection 7.1, we focus on using LLMs to extract content information from journalistic texts for later use in news recommender systems that can assess the (political) stance of a text and the requesting user to present articles from different views instead of enforcing filter bubble effects.

## 3.4. AI for Healthier Political Debate (Task 6.4)

### 3.4.1. Introduction

Task 6.4 focuses on the development of AI models to analyse political debates. The global objective is to 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. We report the research outcomes of Task 6.4 in detail in Section 8.

### 3.4.2. Overview

In Subsection 8.1, CEA develops an analysis of political news in the Dutch-speaking region of Belgium. This analysis complements the one included in D6.3 for France. It highlights common trends (strong representation of governing political tendencies, imbalanced representation of genders) but also differences (lower representation of extreme tendencies in Belgium).

In Subsection 8.2, UvA presents novel AI methodologies for analyzing digital archives of international politics. They focus on new ways of reassembling marginalised, non-canonical entities in these archives. The updated analysis is based on new research and offers insights that are only identified through new questions.

In Subsection 8.3, UCA focuses on detecting and classifying fallacies in political debates. Automating this task is societally important since it contributes to higlighting misinformation spread during debates that have a strong influence of election outcomes. First, they update a dataset including US presidential debates with the 2020 Biden-Trump debates. Second, they propose a transformer-based detection and classification tool that integrates textual and non-textual features.

## 3.5. Perceptions of hyper-local news (Task 6.5)

### 3.5.1. Introduction

Hyper-local news provides information that is directly relevant to the daily lives of community members. It serves as a link between citizens and local institutions, promoting transparency and accountability within democratic processes. By focusing on issues often overlooked by mass media, local news addresses the specific needs and concerns of smaller communities, thereby helping to strengthen social cohesion. It also acts as a platform for local voices, enabling diverse perspectives to be heard and considered. In contrast to mass media, which tends to prioritize broader, more generalized content, local news offers detailed coverage of events and issues that impact residents on a personal level. Researchers are increasingly using AI to analyze local news, offering new insights into patterns, trends, and community concerns that might otherwise go unnoticed. This new angle is important as it enhances our understanding of local dynamics, offers novel insights into the perceptions of journalistic work, and helps to identify gaps in coverage, ultimately contributing to a more informed and engaged audience.

Task 6.5 presents a comprehensive analysis of the dynamics within the ecosystem of local media. Section 9 uses AI methodologies to analyze the various components that contribute to the local media ecosystem, considering journalistic work through framing and the temporal evolution of topic coverage. It then examines several case studies from different regions of Switzerland to extract the essence of what makes this journalism specific. Finally, these analyses are connected to the perceptions of the local population, particularly certain social groups who might not often be considered as target readers.

### 3.5.2. Overview

Subsection 9.1 presents a study on frame analysis of Dutch TV news show transcripts, conducted by IDIAP and NISV. The research examines the use of LLMs for automating the identification of frames in news content, specifically focusing on transcripts from two Dutch public television programs. The methodology involved transcribing spoken content using Kaldi, translating it into English, and annotating it with predefined frame types. The annotated data was then classified using GPT-3.5 with prompt engineering, and the results were compared to human annotations. The study found a moderate level of agreement between human annotators and the LLM, highlighting areas for improvement and suggesting alternative frame types for future research. The work underscores the potential of LLMs in enhancing the efficiency of frame analysis in media studies.

Subsection 9.2 investigates some of the perceptions of members of the migrant community in Lausanne, Switzerland, with respect to the hyper-local media environment. The study employs a mixed-methods approach, combining focus groups with migrant women and men and the application of NLP techniques on articles from a local newspaper. The focus groups discussed news consumption habits, feelings associated with news, and participated in a hands-on annotation activity. The NLP analysis involved named entity recognition, topic modeling, information retrieval, sentiment analysis, and text readability assessment. The findings reveal insights into the information needs and perceptions of the migrant community, as well as whether the local news content meets these needs. The study also highlights the potential of using NLP to support community access to local information.

Subsection 9.3 details research conducted on the dynamics of local news on German news in Switzerland, specifically in the canton of Zürich. The study involved collecting a dataset of approximately 14,500 articles from two local newspapers catering to two cities (Zürich and Winterthur). The research employed NLP techniques such as topic modeling, named entity recognition (NER), and sentiment analysis to examine the content. Key findings include the identification of prominent topics like "Public Transport" and "Gastronomy", or and "Music" and "Art" who clearly cater to local interests. Furthermore, the NER results highlighted the prevalence of local entities, while sentiment analysis revealed a predominance of neutral sentiment in full articles, but a higher occurrence of negative sentiment in headlines. The study underscores the locality of the news content and provides insights into the specific interests and issues pertinent to the Zürich canton community.

Subsection 9.4 examines the unique characteristics and challenges of local media compared to national outlets, with a case of study in Romandy, the French-speaking region of Switzerland. The study employs a comprehensive analysis framework combining quantitative and qualitative methodologies. Quantitative measures include contextual elements, textual features, and topic modeling, revealing distinct patterns in local versus national news coverage. Qualitative analysis, using Grounded Theory, identifies three main axes of differentiation: the proportionality of temporal and geographical scales, the promotion of direct consumption versus advertising campaigns, and the use of pedagogy and popularization versus direct communication. The findings highlight the deeper engagement of local newspapers with regional voices and events, while national media provide broader contexts and integrate advertisements within long-term storytelling. The research underscores the specificity of local journalism and its role in maintaining community ties and addressing societal needs.

Subsection 9.5 focuses on the temporal analysis of media events within the Swiss media ecosystem. The research categorizes media events into three types: broad themes, significant one-time events, and regular, cyclical events, in the model of journalistic "evergreens". Using a dataset of over 1.7 million articles in German, French, and Italian, the study employs advanced techniques such as the BERTopic pipeline for topic modeling and the BM25 algorithm for keyword filtering. The analysis identifies phases in media coverage through changepoint detection and examines the dynamics of stakeholders and sentiment associated with key entities. Results indicate distinct phases in media coverage and variations in sentiment and stakeholder dynamics across different topics. This framework provides insights into the temporal dynamics of media coverage across different linguistic regions in Switzerland, and could be applied to other media ecosystems for similar analyses.

## 3.6. Measuring and Predicting User Perception of Social Media (Task 6.6)

### 3.6.1. Introduction

User perception classification aims at understanding the affective impact that social media data has on the viewers and users of social media platforms. It represents an essential component of media data analysis, as it ties together many aspects of human perception and understanding, spanning but not being limited to: interestingness, aesthetics, emotional content, and memorability. Researchers show increasing interest in studying these concepts, and, while this domain still represents an open research direction, improvements have been made in understanding the way media data influences and affects viewers.

Task 6.6 is reported in detail in Section 10, and presents the advancements made with regards to the study of user perception of social media and multimedia. The three contributions target two critical concepts related to the perception of media data, namely memorability prediction and facial expression recognition.

### 3.6.2. Overview

Subsection 10.1 proposes the study of several methods of augmenting the training dataset for media memorability classification, based on a classification of how hard-to-predict videos are from a memorability standpoint. We base this work on two previous works and look into four categories of videos that would normally be harder to classify, and augment these videos by taking more significant samples from the at training time. This creates an imbalance in the dataset, favoring the presence of more hard-to-classify samples, and showing that two of the methods improve the performance of a ViViT-based approach.

In Subsection 10.2 we present DFER-CLIP, a novel visual-language model based on the CLIP model, designed for in-the-wild Dynamic Facial Expression Recognition (DFER). It features a visual component that uses a temporal model with Transformer encoders to extract temporal facial expression features and a textual component that employs descriptions of facial behavior generated by large language models like ChatGPT to capture the relationships between facial expressions. Additionally, a learnable token is introduced to help the model learn relevant context information for each expression during training.

In subsection 10.3 we present EmoCLIP, a novel visual-language model that aims at zero-shot recognition of facial behavior, that goes beyond the supervised learning paradigm that aims at recognition of the 7 basic emotions. The proposed scheme learns in a contrastive manner from pairs of images and textual descriptions of the facial behaviors image and text encoders that at inference time can be be used to recongize unseen behaviour categories, so long as textual descirptions of the facial behaviours that are associated with it are available.

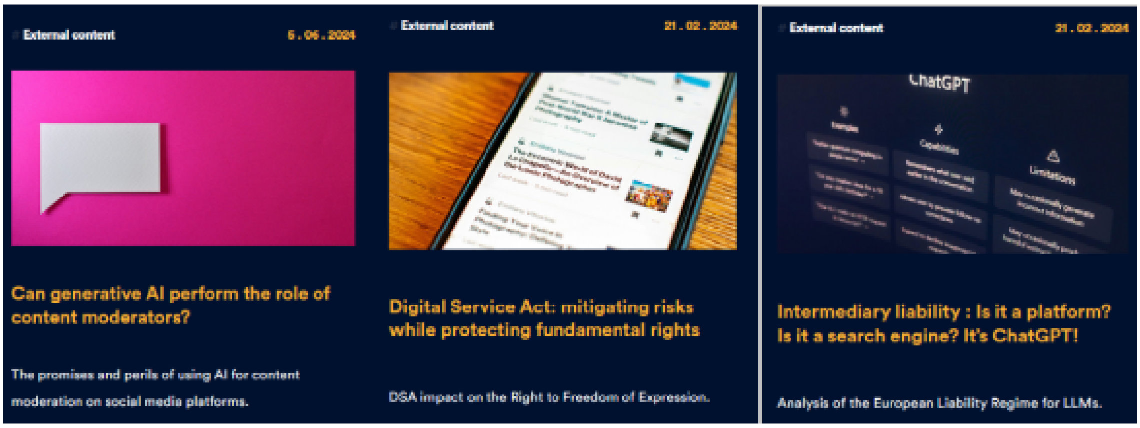



*Figure 1. Illustration of content moderation entries published on the AI Media Observatory website.*

# 4. Policy recommendations on content moderation

## 4.1. Publication of a summary version of D6.2: Report on Policy for Content Moderation

While the report on policy on content moderation (D6.2)[1] was submitted in M30, a summary version under a 'Results in brief' format was published on the AI4Media website in M38. This factsheet/results in brief version[2] is a summary of D6.2. It presents an overview of the EU policy initiatives on content moderation as well as alternative approaches to content moderation by online platforms and civil society. The results in brief version is an appealing and concise format to disseminate the research outputs of the deliverable. It assesses the challenges and advantages of these instruments and diverging approaches to outline policy recommendations for the future of content moderation in the EU landscape. More specifically, it provides an introduction to content moderation, including algorithmic content moderation and its challenges to fundamental rights such as freedom of expression, as well as an analysis of the legal landscape composed of hard law (lex generalis and lex specialis) and other types of regulatory instruments. It also investigates the criticisms addressed to each of these instruments and recommendations for the future. Self-regulatory initiatives as alternative approaches, such as end-user moderation and self-moderation through bodies and new models are analysed. Moreover, it reflects on the AI4Media workshop on AI and content moderation held with media practitioners. Finally, based on the results of the previous analysis, a set of policy recommendations for content moderation is provided. With this format we hoped to provide partners with no legal background a concise and easy-to-read information on content moderation policies which should guide their work when designing any AI content moderation tools (e.g. recommender systems). By doing so we aim to mind the gap between legal and technical sciences towards a common understanding of the regulatory requirements.

## 4.2. AI Media Observatory

The Observatory was officially launched in Autumn 2023 and since its start content moderation has been a regular topic of the curated "Your AI Media Feed". Three such contributions are illustrated

[1] https://www.ai4media.eu/reports/report-on-policy-for-content-moderation-d6-2/
[2] https://www.ai4media.eu/wp-content/uploads/2023/12/Factsheet_AI-and-Content-Moderation.pdf

in Figure 1.

## 4.3. Research and publications on Generative AI and Content Moderation

Generative AI was not explicitly part of the scope of research at the start of the project. However, the outbreak of Generative AI (GenAI) systems tools after the release of Chat GPT by Open AI in November 2022 has triggered additional analysis and research from AI4Media including on content moderation about manipulated or generated content made through GenAI tools. Indeed, many content moderation considerations related to GenAI content range from freedom of expression to copyright and privacy. Between M37 and M48 of the project, several activities based on T6.1's mandate were undertaken.

### 4.3.1. EU DisinfoLab Factsheet on AI Generated content and platform's policies

KUL collaborated with the EU DisinfoLab on their factsheet on Platforms' policies on AI-manipulated and generated misinformation.1 The factsheet delves into how some of these VLOPs – Facebook, Instagram, TikTok, X (formerly Twitter), and YouTube – approach AI-manipulated or AI-generated content in their terms of use, exploring how they address its potential risk of becoming mis- and disinformation. It has been regularly updated following the rapid changes of policies by platforms. It is currently in its version 3. The factsheet contains several key conclusions:

- Evolving Policies: Definitions around AI and disinformation are becoming more aligned across platforms. By 2024, Meta, YouTube, and TikTok began to explicitly reference AI-generated content in their policies, a shift from 2023 when only Facebook and TikTok did so.
- Focus on AI-Generated Media: Although there's a trend towards integrating AI-specific guidelines, the distinction between general misinformation and AI-related issues isn't always clear. Platforms mainly focus on images and videos, with increasing mention of AI-generated audio, but less attention was given so far to AI-generated text.
- Content Moderation Approaches: Platforms like TikTok distinguish between permissible and banned uses of AI-generated content, mostly driven by concerns about misinformation or compliance with copyright and quality standards.
- Regulatory Compliance: All studied platforms qualify as Very Large Online Platforms (VLOPs) under the Digital Services Act (DSA), which mandates compliance with specific obligations. The Code of Practice on Disinformation, will be reinforced by the DSA once endorsed as an official DSA code of practice, requires platforms to establish or confirm their policies for AI-manipulated content and align with transparency obligations just adopted in the AI Act.
- DSA risks assessment framework: The factsheet concludes that the systemic risks assessment framework set up of the DSA should be completed by a framework specific to AI generated content to provide guidance and prevent arbitrariness in the assessment process.
- Collaboration: The factsheet also encourages platforms to update their policies, enhance cooperation with experts, and take responsibility for combating disinformation.
- Responsibility burden: Content moderation is a complex landscape and the responsibility burden between all relevant stakeholders needs to be more strongly and clearly regulated when it comes to GenAI content.

#### 4.3.2. White paper on Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities

AI4Media also collaborated with the EU-funded AI/Disinformation projects cluster on different occasions. In February 2024, projects in collaboration with EDMO joined forces and published a white paper on Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities[3]. One of the focuses of the paper was to discuss the ethical and legal challenges associated with GenAI and disinformation. KUL contributed to the legal and policy aspects of the paper. The challenges and opportunities brought by Generative AI in the context of disinformation production, spread, detection, and debunking were explored in the white paper. It enabled to build AI4Media network and strengthen the collaboration on topics related to AI and content moderation.

#### 4.3.3. Chapter on AI in the handbook: "Disinformation: a multidisciplinary analysis"

Through past collaborations including the White Paper, KUL was invited to contribute to the Handbook on "Disinformation: A multidisciplinary analysis" with a chapter on Generative AI and disinformation. The forthcoming book is edited by Lisa Ginsborg from the European Digital Media Observatory (EDMO) and Paula Gori from the European University Institute (EUI). The book will be published by Springer. KUL's chapter is currently under the review process. This publication was the opportunity to materialise the latest research on AI, disinformation and content moderation challenges related to GenAI.

#### 4.3.4. AI4Media Final Policy recommendations

T6.1 was also strongly linked to WP2 and related tasks led by KUL. The AI4Media final policy recommendation for the use of AI in the media sector and drafted as part of D2.6 due in M48 also includes content moderation considerations. They include the question of researchers' data access, transparency of content moderation, mechanisms and systemic risks mitigation by VLOPS. The role of media in the information ecosystem is also part of the wider analysis including the latest EU policy and regulation initiatives to regulate the moderation of media content. Additionally, the transparency measures contained in the AI Act about GenAI content were also analysed. It includes the discussions around watermarking or labeling and related content moderation implications for platforms hosting GenAI content.

### 4.4. Research and publication on Metaverse, DSA and content moderation

Additional research was conducted about the new EU content moderation lex generalis, namely the DSA and its applicability to the Metaverse for addressing safety considerations. KUL participated in the International Congress: "Towards a Responsible Development of the Metaverse" organised by the University of Alicante on 13-14 June 2024. KUL's paper was selected for the conference. Following the paper's presentation and paper's peer review, the paper was selected as one of the best 12 papers for publication in a forthcoming Special Issue of the Interactive Entertainment Law.[4] Proceedings are already available[5].

[3] https://edmo.eu/edmo-news/new-white-paper-on-generative-ai-and-disinformation-recent-advances-challenges-and-opportunities/
[4] https://www.elgaronline.com/view/journals/ielr/ielr-overview.xml
[5] https://catedrametaverso.ua.es/papers/

The paper focuses on one of the most pressing issues when it comes to the Metaverse which are content moderation considerations and addressing the question of liabilities. The paper explores complex considerations of liability and accountability that Metaverse and content moderation could pose through the prism of the recently adopted Digital Services Act (DSA). The newly enforced legislation aims to create a safer digital space where the fundamental rights of users are protected. This contribution assesses whether the new regulation fulfills its promises in the context of the metaverse. While numerous challenges are associated with metaverse regulation, the paper focused on user safety on metaverse platforms through the lenses of the Digital Services Act exclusively. One of the conclusions of the paper was that there is no clear certainty on whether behavior and content are regulated similarly in the metaverse. Therefore, clarification of what the DSA (and more generally, EU Digital Regulation) defines as information and content appears primordial.

# 5. Manipulation and synthetic content detection in multimedia (T6.2) - Data generation and manipulation

## 5.1. Cloth Interactive Transformer for Virtual Try-On

**Contributing partner:** UNITN

### 5.1.1. Introduction

Virtual try-on (VTON), derived from fashion editing [5, 6], aims to transfer a desired in-shop clothing item onto a customer's body. If properly solved, VTON will provide a time and energy-saving shopping experience in our daily life. Additionally, the applicability of VTON can go beyond shopping, e.g., could be used in a virtual stylist application for editing the costumesn in TV programs or movies. In practice, VTON has already been deployed in some big-brand clothing stores or e-commerce shopping applications based on its convenience. However, most of the existing methods are based on 3D model pipelines [7, 8, 9, 10, 11] and follow the conventions of traditional computer graphics. Despite the detailed results, they require many labor resources, a massive time investment, and complex data acquisition such as multi-view videos or 3D scans [12] that hinder their widespread application. Alternatively, conditional GAN-based methods such as image-to-image translation or other image generation approaches [13, 14], made recently some positive progress. However, there are still obvious artifacts in the generated results.

To make the results of 2D image-based methods more realistic, the classic two-stage pipeline VTON [15] was proposed, utilizing the first stage to warp the in-shop clothing to a desired deformation style and in the second stage, the warped cloth is aligned to the body shape of a given consumer. Although the result is improved, the visual performance is still far from the plausible generation. Many approaches following this pipeline, e.g., CP-VTON [16], ACGPN [17], and CP-VTON+ [18] were proposed with better performance. However, their results are good only for plain texture or simple style clothes. When dealing with hard cases like rich textures or complex patterns, their performance becomes far from satisfactory. To ease this problem Xu *et al.* [19] proposed an intermediate operation which considers the transformation of the target person image. But the improvement in visual performance is largely built on a complicated network architecture, which is time-consuming for training such a model. This is mainly because they rarely pay sufficient attention to the correlation between two crucial factors, i.e., the person and the in-shop clothing representations. Undoubtedly, some inevitable mismatch phenomena occur in the warped in-shop clothing items, and consequently, this will harm the final try-on results. Moreover, most of the aforementioned methods use convolution neural networks (CNN) to model the relations of the input data. However, it is not easy for the fix-size convolutional kernels to capture the global long-range dependence because of their limited receptive fields.

### 5.1.2. Methodology

Based on these observations, we argue that it is beneficial to guide the person feature learning with the guidance of the in-shop cloth feature, and vice versa. To this end, we also adopt the two-stage pipeline like VTON's [15], but to address the aforementioned limitations, we propose a novel Cloth Interactive Transformer (CIT) for the virtual try-on task. The overall architecture of the proposed CIT is shown in Figure 2.

In the geometric matching stage, we design the CIT matching block which can model long-range relations between the person and clothing representations interactively. Also, a valuable correlation map is produced to boost the performance of the thin-plate spline (TPS) transform [20]. Unlike the

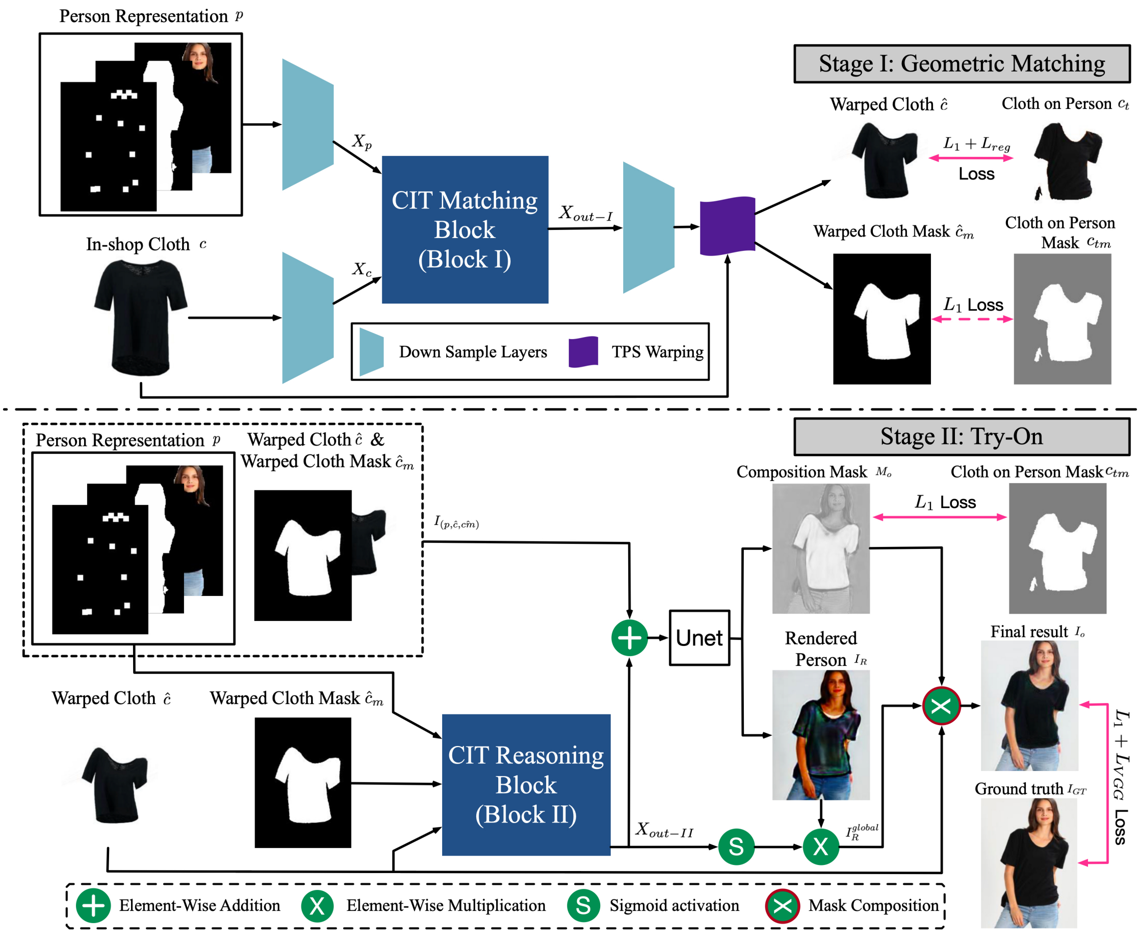



*Figure 2. The overall architecture of the proposed CIT for virtual try-on. The upper part is the Geometric Matching stage warping the in-shop clothing items while the bottom part is the Try-On stage synthesizing the final try-on person image.*

traditional hand-crafted shape context matching strategies [21, 22, 23] which are only suitable for a certain feature type, the proposed CIT matching block has learnable features and can model global relationships via the cross-modal transformer encoder. Consequently, the warped clothing becomes more natural and can fit a wearer's pose and shape more appropriately. In the second stage, unlike previous methods [16, 18] that treat the warped in-shop clothing and its corresponding mask as only one modality, we proposed a novel CIT reasoning block that takes the person representation, the warped clothing, and the warped clothing mask separately as inputs. Then, the activated global consistency correlation is established and it is used to strengthen the significant region in the input data, leading to more natural intermediate results from the UNet. This also serves as an attention map to activate the rendered person image, making the final results sharper and more realistic.

In the CIT matching block, our main idea is to improve the modeling of the person feature by explicitly encoding the target in-shop clothing feature. Since the in-shop clothing items are non-rigid, it is hard to learn the relationship directly from the clothing items. Hence, we resort to the correlation between the person and the target in-shop clothing item. With the help of this learned correlation, the person-related feature can be refined indirectly by the in-shop clothing

| Method | Q1 | Q2 |
|---|---|---|
| CP-VTON [16] | 19.5 | 14.6 |
| CP-VTON+ [18] | 24.8 | 25.2 |
| ACGPN [17] | 23.6 | 24.8 |
| CIT (Ours) | **32.1** | **35.4** |

*Table 2. User study comparison on two questions. Q1 denotes 'Which image is the most photo-realistic?', and Q2 denotes 'Which image preserves the details of the target clothing the most?' in the user study*

features, and vice versa. Such an analysis is also considered in the CIT reasoning block. With the CIT matching and the reasoning blocks, our method directly explores the interaction and the interplay between the person and in-shop clothing features in a global way. Both kinds of features are updated synchronously and affect each other.

### 5.1.3. Experiments

**5.1.3.1. Datasets.** We conduct all our experiments on the dataset collected by Han *et al.* [15] and used in VITON, CP-VTON, and CP-VTON+. Because of the copyright issues, we could only utilize the reorganized version. It contains around 19,000 front-view women and top clothing image pairs. Specifically, there are 16,253 cleaned pairs which are split into a training set and a validation set with 14,221 and 2,032 pairs, respectively. In the training set, the target clothing and the clothing worn by the wearer are the same. However, in the test stage, there are two kinds of test settings. The first one is the same as the training settings, where the target clothing and the clothing item worn by the wearer are the same (we refer to this kind as a retry-on case because it is just like the wearer takes off the cloth first then retries this cloth, hence we have the ground truth for this case). Another setting means that the target clothing item is different from the one worn by the wearer (we refer to this kind as the try-on case). Due to space limits, we will present qualitative results only on the latter setting.

**5.1.3.2. Evaluation Metrics.** To evaluate the performance of our method, we follow [18, 17, 24] to use the Structural Similarity (SSIM) [25], and Learned Perceptual Image Patch Similarity (LPIPS) [26] metrics in the try-on stage. Note that we adopt the original human image with the original clothing item as the reference image for SSIM and LPIPS (lower score means better), and utilize the parsed segmentation area for the current upper clothing as the JS reference. For the different clothing try-on case (where no ground truth is available), we evaluate the performance of our method and other state-of-the-art methods by the Inception Score (IS) [27].

**5.1.3.3. Qualitative Comparisons.** To validate the performance of the proposed CIT for virtual try-on, we firstly present the visualization comparison results for try-on person images (see Figure 3). There is no doubt that our CIT outperforms the other state-of-the-art methods. The proposed CIT can keep the original clothing texture and its pattern as much as possible, and the final resulting images are more realistic and natural. Compared to our method, the other approaches display many artifacts, for example, the irregular logo pattern (the first row), the over-warped cloth texture (the third row), the ridiculous results for unique or complicated-style cloth (the last row), etc. These try-on results strengthen the evidence that our proposed CIT is superior to others.

**User Study.** We also evaluate the proposed CIT and other methods via a user study. We randomly select 120 sets of reference and target clothing images from the test dataset. Given the reference


*Figure 3. Qualitative comparisons of different state-of-the-art methods.*

images and the target in-shop clothing items, 30 users are asked to choose the best outputs of our model and baselines (i.e., CP-VTON, CP-VTON+, and ACGPN) according to the two questions: (Q1) Which image is the most photo-realistic? (Q2) Which image preserves better the details of the target clothing? As shown in Table 2, we can see that the proposed CIT achieves significantly better results than the other methods, which further demonstrates that our model generates more realistic images. Additionally, CIT also preserves the details of the clothing items compared to the other methods.

**5.1.3.4. Quantitative Evaluation.** To further evaluate our method CIT, we also adopt five metrics, i.e., JS, SSIM, LPIPS, PSNR, and IS for numerical result comparison. JS is to evaluate the quality of the warped mask in the first geometric matching stage with same-pair test samples, which is similar to the IoU metric used in CP-VTON+ but is more convenient for implementation.

| Method | JS ↑ | SSIM↑ | LPIPS↓ | IS↑ | PSNR↑ |
|---|---|---|---|---|---|
| CP-VTON [16] | 0.759 | 0.800 | 0.126 | 2.832 | 14.544 |
| CP-VTON+ [18] | **0.812** | 0.817 | 0.117 | **3.074** | 21.789 |
| CIT (Ours) | 0.800 | **0.827** | **0.115** | 3.060 | **23.464** |

*Table 3. Quantitative comparison in terms of JS, SSIM, LPIPS, PSNR, and IS metric comparisons. For JS, SSIM, PSNR, and IS, the higher is better, while for LPIPS, the lower is better. What's more, IS is used to evaluate the unpaired try-on test settings while others are all for paired retry-on settings. We also provide detailed explanation and evidence to show the conclusion that higher JS is not always indicate higher visual result in the following ablation study part.*

| Method | Parameters | GMACs | FPS |
|---|---|---|---|
| CP-VTON [16] | 40.40 M | 13.14 | 6.34 |
| CP-VTON+ [18] | 40.41 M | 13.45 | 6.27 |
| CIT (Ours) | 78.26 M | 26.70 | 5.70 |

*Table 4. Complexity Analysis in terms of the overall parameters, GMACs, and running time (FPS) for each method.*

Note that we take cloth masks on person as the reference images. Other metrics are designed to evaluate the performance of the second try-on stage.

The results of JS are shown in Table 3. We think that with the help of the proposed Interactive Transformer in the CIT matching block, our method can learn more reasonable texture transformation patterns. This learned strong texture-focused transformation pattern may affect the shape alignment. Hence, we want to argue that the JS score only focuses on the shape alignment aspect between the ground truth mask and the warped clothing mask. As a result, it cannot reveal the overall quality of the final generated human images. For instance, though Table 3 shows that CP-VTON+ has the best JS score of 0.812, which is higher than ours of 0.800, the qualitative results in Figure 3 show that our method is superior to CP-VTON+, which means that the higher JS or IoU scores do not always mean a better visual result.

For the same-pair settings (retry-on case), we adopt SSIM, PSNR, and LPIPS to evaluate the performance. The numerical results are shown in Table 3. It can be seen that the proposed CIT achieves the best numerical evaluation results for both these three metrics, the highest SSIM and PSNR scores and the lowest LPIPS score among other state-of-the-art methods such as CP-VTON [16], and CP-VTON+ [18]. For the unpaired test settings (try-on case), we utilize IS to evaluate it. The results are shown in Table 3. Our CIT achieves just a slightly lower IS score 3.060 compared to 3.074 of CP-VTON+ but with more realistic visual results shown in Figure 3. We think the most possible reason for this problem is that IS is an objective metric that is usually used to measure the quality of the generated images at the feature level based on the image diversity and clarity. Hence, it may ignore some pixel-level properties. Note that all numerical results of CP-VTON were trained by ourselves with its officially released code, while for CP-VTON+, the results were obtained based on the officially provided checkpoints. Overall, though we do not obtain the best quantitative scores on JS and IS metrics, our proposed CIT can generate sharper and more realistic try-on images compared to CP-VTON+ (see qualitative results). Hence, we want to clarify that better numerical results do not always mean better visual results.

We also conduct the experiments for each method in terms of the overall parameters, the GMACs (Multiply Accumulate), and the running time by FPS (frames per second). The results are shown in Table 4. Compared to CP-VTON and CP-VTON+, though our method gets more realistic visual results, the proposed CIT achieves the worst performance in all the efficiency metrics. Hence, reducing the complexity of the proposed CIT will be an interesting exploration in our future work.

#### 5.1.4. Conclusion

The main contributions of this work are as follows:

- We design a novel two-stage Cloth Interactive Transformer (CIT) for the challenging virtual try-on task, which can model long-range interactive relations between the clothing and the person representations. To the best of our knowledge, the proposed CIT is the first transformer-based framework for virtual try-on.
- We propose a new two-modality CIT matching block in the geometric matching stage, making the in-shop clothing to be better warped to the desired direction more.
- We introduce a new three-modality CIT reasoning block; based on this block, the latent global long-range correlation can be strengthened to make the final try-on results more realistic.

#### 5.1.5. Relevant publications

- B. Ren, H. Tang, F. Meng, R. Ding, P. Torr, and N. Sebe, Cloth Interactive Transformer for Virtual Try-On, ACM Transactions on Multimedia Computing, Communications, and Applications, 20(4), article 92, December 2023 [28]
  Zenodo record: https://zenodo.org/records/11303375

#### 5.1.6. Relevant software/datasets/other outcomes

- The Pytorch implementation can be found in https://github.com/Amazingren/CIT

#### 5.1.7. Relevance to AI4Media use cases and media industry applications

Our virtual try-on tool is generic and can be applied to person image manipulation and generation. 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".It could also have an impact on other media applications, e.g., virtual try-on of costumes for film/TV or generation of images for advertisement and content creators (e.g., useful for fashion influencers).

### 5.2. Edge Guided GANs for Semantic Image Synthesis

**Contributing partner:** UNITN

#### 5.2.1. Introduction

Semantic image synthesis refers to generating photo-realistic images conditioned on pixel-level semantic labels. This task has a wide range of applications such as image editing and content generation [29, 30, 31, 32, 33]. Although existing methods conducted interesting explorations, we still observe unsatisfactory aspects, mainly in the generated local structures and details, as well as small-scale objects, which we believe are mainly due to three reasons: 1) Conventional methods [34, 35, 32] generally take the semantic label map as input directly. However, the input label map provides only structural information between different semantic-class regions and does

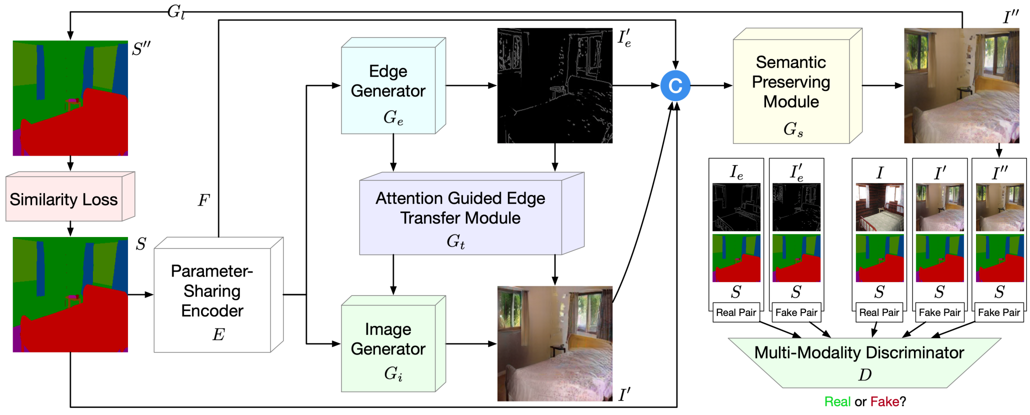



*Figure 4. Overview of the proposed ECGAN. It consists of a parameter-sharing encoder $E$, an edge generator $G_e$, an image generator $G_i$, an attention guided edge transfer module $G_t$, a label generator $G_l$, a similarity loss module, a contrastive learning module $G_c$ (not shown for brevity), and a multi-modality discriminator $D$. Both $G_e$ and $G_i$ are connected by $G_t$ from two levels, i.e., edge feature-level and content-level, to generate realistic images. $G_s$ is proposed to preserve the semantic information of the input semantic labels. $G_l$ aims to transfer the generated image back to the label for calculating the similarity loss. $G_c$ tries to capture more semantic relations by explicitly exploring the structures of labeled pixels from multiple input semantic layouts. $D$ aims to distinguish the outputs from two modalities, i.e., edge and image. The whole framework can be end-to-end trained so that each component can benefit from each other. The symbol Ⓒ denotes channel-wise concatenation.*

not contain any structural information within each semantic-class region, making it difficult to synthesize rich local structures within each class. Taking the label map $S$ in Figure 4 as an example, the generator does not have enough structural guidance to produce a realistic bed, window, and curtain from only the input label ($S$). 2) The classic deep network architectures are constructed by stacking convolutional, down-sampling, normalization, non-linearity, and up-sampling layers, which will cause the problem of spatial resolution losses of the input semantic labels. 3) Existing methods for this task are typically based on global image-level generation. In other words, they accept a semantic layout containing several object classes and aim to generate the appearance of each one using the same network. In this way, all the classes are treated equally. However, because different semantic classes have distinct properties, using specified network learning for each would intuitively facilitate the complex generation of multiple classes.

### 5.2.2. Methodology

To address these three issues, in this work, we propose a novel edge guided generative adversarial network with contrastive learning (ECGAN) for semantic image synthesis. The overall framework of ECGAN is shown in Figure 4. To tackle 1), we first propose an edge generator to produce the edge features and edge maps. Then the generated edge features and edge maps are selectively transferred to the image generator and improve the quality of the synthesized image by using our attention guided edge transfer module. To tackle 2), we propose an effective semantic preserving module, which aims at selectively highlighting class-dependent feature maps according to the original semantic layout. We also propose a new similarity loss to model the relationship between semantic categories. Specifically, given a generated label $S''$ and the corresponding ground truth $S$, the similarity loss constructs a similarity map to supervise the learning. To tackle 3), a straightforward solution would be to model the generation of different image classes individually. By doing this, each class could have its own generation network structure or parameters, thus greatly avoiding the learning of a biased generation space. However, there is a fatal disadvantage to this. That is, the number of parameters of the network will increase linearly with the number of semantic classes $N$, which will cause memory overflow and will make it impossible to train the model. If we use $p_e$ and $p_d$ to denote the number of parameters of the encoder and decoder, respectively, then the total number of the network parameter should be $p_e+N\times p_d$ since we need a new decoder for each class. To further address this limitation, we introduce a pixel-wise contrastive learning approach that elevates the current image-wise training method to a pixel-wise method. By leveraging the global semantic similarities present in labeled training layouts, this method leads to the development of a well-structured feature space. In this case, the total number of the network parameters only is $p_e+p_d$. Moreover, we explore image generation from a class-specific context, which is beneficial for generating richer details compared to the existing image-level generation methods. A new class-specific pixel generation strategy is proposed for this purpose. It can effectively handle the generation of small objects and details, which are common difficulties encountered by the global-based generation.

With the proposed ECGAN, we achieve new state-of-the-art results on Cityscapes [36], ADE20K [37], and COCO-Stuff [38] datasets, demonstrating the effectiveness of our approach in generating images with complex scenes and showing significantly better results compared with existing methods.

The initial results were presented in [39] and the further extension was three-fold: (1) We considered a more detailed analysis of related works by including recently published works dealing with semantic image synthesis and contrastive learning. (2) We proposed a novel module, i.e., multi-scale contrastive learning, to push the same-class features from different scales to be similar by using the proposed multi-scale and cross-scale contrastive learning losses. Equipped with this new module, our ECGAN proposed in [39] has been upgraded to ECGAN++. (3) We extended the quantitative and qualitative experiments by comparing our ECGAN and ECGAN++ with the very recent works on three public datasets. Extensive experiments showed that the proposed ECGAN++ achieves the best results compared with existing methods.

### 5.2.3. Experiments

#### 5.2.3.1. Experimental Setup

**Datasets.** We follow [34] and conduct experiments on three datasets, i.e., Cityscapes [36], ADE20K [37], and COCO-Stuff [38]. For more details about these datasets, please refer to GauGAN [34].

**Evaluation Metrics.** We employ the mean Intersection-over-Union (mIoU), Pixel Accuracy (Acc), and Fréchet Inception Distance (FID) [40] as the evaluation metrics. For more details about these evaluation metrics, please refer to GauGAN [34].

#### 5.2.3.2. Experimental Results

**Qualitative Comparisons.** We adopt GauGAN as the encoder $E$ to validate the effectiveness of the proposed method. Visual comparison results on all three datasets with the state-of-the-art method (i.e., OASIS [41]) are shown in Figure 5. We can see that ECGAN and ECGAN++ achieve visually better results with fewer visual artifacts than the existing state-of-the-art method. Examining Figure 5, it is evident that the SOTA method produces numerous visual artifacts across varied categories like vegetation, cars, buses, roads, buildings, fences, beds, cabinets, curtains, elephants, etc. In contrast, our approach generates significantly more realistic content, as can be observed on both sides of the figure. Moreover, the proposed methods generate more local structures and details than the SOTA method.

**User Study.** We follow the same evaluation protocol as GauGAN and conduct a user study. Specifically, we provide the participants with an input layout and two generated images from


*Figure 5. Existing state-of-the-art method (i.e., OASIS) vs. our proposed ECGAN on three datasets. Cityscapes: left; ADE20K: top right four; COCO-Stuff: bottom right four.*

| AMT ↑ | Cityscapes | ADE20K | COCO-Stuff |
|---|---|---|---|
| Our ECGAN vs. CRN [29] | 88.8 ± 3.4 | 94.8 ± 2.7 | 95.3 ± 2.1 |
| Our ECGAN vs. Pix2pixHD [35] | 87.2 ± 2.9 | 93.6 ± 3.1 | 93.9 ± 2.4 |
| Our ECGAN vs. SIMS [33] | 85.3 ± 3.8 | - | - |
| Our ECGAN vs. GauGAN [34] | 84.7 ± 4.3 | 88.4 ± 3.7 | 90.8 ± 2.5 |
| Our ECGAN vs. DAGAN [42] | 81.8 ± 3.9 | 86.2 ± 3.6 | - |
| Our ECGAN vs. CC-FPSE [32] | 79.5 ± 4.2 | 85.1 ± 3.9 | 86.7 ± 2.8 |
| Our ECGAN vs. LGGAN [43] | 78.4 ± 4.7 | 82.7 ± 4.5 | - |
| Our ECGAN vs. OASIS [41] | 76.7 ± 4.8 | 80.6 ± 4.5 | 82.5 ± 3.1 |

*Table 5. User study on Cityscapes, ADE20K, and COCO-Stuff. The numbers indicate the percentage of users who favor the results of the proposed ECGAN over the competing methods.*

different models and ask them to choose the generated image that looks more like a corresponding image of the layout. The users are given unlimited time to make the decision. For each comparison,

| AMT ↑ | Cityscapes | ADE20K | COCO-Stuff |
|---|---|---|---|
| Our ECGAN++ vs. Our ECGAN [39] | 64.3 ± 3.2 | 67.5 ± 3.8 | 70.4 ± 2.6 |

*Table 6. User study on Cityscapes, ADE20K, and COCO-Stuff. The numbers indicate the percentage of users who favor the results of the proposed ECGAN++ over the proposed ECGAN.*

| Method | Cityscapes | | | ADE20K | | | COCO-Stuff | | |
|---|---|---|---|---|---|---|---|---|---|
| | mIoU ↑ | Acc ↑ | FID ↓ | mIoU ↑ | Acc ↑ | FID ↓ | mIoU ↑ | Acc ↑ | FID ↓ |
| CRN [29] | 52.4 | 77.1 | 104.7 | 22.4 | 68.8 | 73.3 | 23.7 | 40.4 | 70.4 |
| SIMS [33] | 47.2 | 75.5 | 49.7 | - | - | - | - | - | - |
| Pix2pixHD [35] | 58.3 | 81.4 | 95.0 | 20.3 | 69.2 | 81.8 | 14.6 | 45.8 | 111.5 |
| GauGAN [34] | 62.3 | 81.9 | 71.8 | 38.5 | 79.9 | 33.9 | 37.4 | 67.9 | 22.6 |
| DPGAN [44] | 65.2 | 82.6 | 53.0 | 39.2 | 80.4 | 31.7 | - | - | - |
| DAGAN [42] | 66.1 | 82.6 | 60.3 | 40.5 | 81.6 | 31.9 | - | - | - |
| SelectionGAN [45] | 83.8 | 82.4 | 65.2 | 40.1 | 81.2 | 33.1 | - | - | - |
| SelectionGAN++ [46] | 64.5 | 82.7 | 63.4 | 41.7 | 81.5 | 32.2 | - | - | - |
| LGGAN [43] | 68.4 | 83.0 | 57.7 | 41.6 | 81.8 | 31.6 | - | - | - |
| LGGAN++ [47] | 67.7 | 82.9 | 48.1 | 41.4 | 81.5 | 30.5 | - | - | - |
| CC-FPSE [32] | 65.5 | 82.3 | 54.3 | 43.7 | 82.9 | 31.7 | 41.6 | 70.7 | 19.2 |
| SCG [48] | 66.9 | 82.5 | 49.5 | 45.2 | 83.8 | 29.3 | 42.0 | 72.0 | 18.1 |
| OASIS [41] | 69.3 | - | 47.7 | 48.8 | - | 28.3 | 44.1 | - | 17.0 |
| RESAIL [49] | 69.7 | 83.2 | 45.5 | 49.3 | 84.8 | 30.2 | 44.7 | 73.1 | 18.3 |
| SAFM [50] | 70.4 | 83.1 | 49.5 | 50.1 | **86.6** | 32.8 | 43.3 | **73.4** | 24.6 |
| PITI [51] | - | - | - | - | - | - | - | - | 19.36 |
| T2I-Adapter [52] | - | - | - | - | - | - | - | - | 16.78 |
| SDM [53] | - | - | **42.1** | - | - | 27.5 | - | - | 15.9 |
| ECGAN (Ours) | 72.2 | 83.1 | 44.5 | 50.6 | 83.1 | 25.8 | 46.3 | 70.5 | 15.7 |
| ECGAN++ (Ours) | **73.3** (+1.1) | **83.9** (+0.8) | 42.2 (-2.3) | **52.7** (+2.1) | 85.9 (+2.8) | **24.7** (-1.1) | **47.9** (+1.6) | 72.3 (+1.8) | **14.9** (-0.8) |

*Table 7. Quantitative comparison of different methods on Cityscapes, ADE20K, and COCO-Stuff.*

we randomly generate 400 questions for each dataset, and each question is answered by 10 different participants. For other methods, we use the public code and pretrained models provided by the authors to generate images. As shown in Table 5, users favor our synthesized results on all three datasets compared with other competing methods, further validating that the generated images by ECGAN are more natural. Moreover, we can see in Table 6 that users favor our synthesized results by the proposed ECGAN++ compared with the proposed ECGAN, validating the effectiveness of the proposed multi-scale contrastive learning method.

**Quantitative Comparisons.** Although the user study is more suitable for evaluating the quality of the generated images, we also follow previous works and use mIoU, Acc, and FID for quantitative evaluation. The results of the three datasets are shown in Table 7. The proposed ECGAN and ECGAN++ outperform other leading methods by a large margin on all three datasets, validating the effectiveness of the proposed methods.

**Memory Usage.** The proposed method is memory-efficient compared to those methods which model the generation of different image classes individuals such as LGGAN [43]. Thus, we compare the memory usage during training/testing when the batch size is set to 1. The memory (GB) of LGGAN on CityScapes (30 categories), ADE20K (150 categories), and COCO-Stuff (182 categories) datasets are about 17.8, 23.9, and 28.1, respectively. The memory (GB) of our proposed method on the Cityscapes, ADE20K, and COCO-Stuff datasets is about 6.3, 5.6, and 5.9 respectively. It is clear that LGGAN's memory requirement significantly escalates as category numbers increase, whereas our method maintains comparable memory demands. This advantage becomes even more prominent when using larger batch sizes, implying we can train/test the model with larger batches on the same GPU devices.

### 5.2.4. Conclusion

The main contributions of this work are as follows:

- We propose a novel ECGAN for the challenging semantic image synthesis task. To the best of our knowledge, we are the first to explore the edge generation from semantic layouts and then utilize the generated edges to guide the generation of realistic images.
- We propose an effective attention guided edge transfer module to selectively transfer useful edge structure information from the edge generation branch to the image generation branch.
- We design a new semantic preserving module to highlight class-dependent feature maps based on the input semantic label map for generating semantically consistent results.
- We propose a new similarity loss to capture the intra-class and inter-class semantic dependencies, leading to robust training.
- We propose a novel contrastive learning method, which learns a well-structured pixel semantic embedding space by utilizing global semantic similarities among labeled layouts. Moreover, we propose a multi-scale contrastive learning method with two novel multi-scale and cross-scale losses that enforces local-global feature consistency between low-resolution global and high-resolution local features extracted from different scales.
- We conduct extensive experiments on three challenging datasets under diverse scenarios, i.e., Cityscapes [36], ADE20K [37], and COCO-Stuff [38]. Both qualitative and quantitative results show that the proposed methods are able to produce remarkably better results than existing baseline models regarding both visual fidelity and alignment with the input semantic layouts. Moreover, our methods can generate multi-modal images and edges, which have not been considered by existing state-of-the-art methods.

### 5.2.5. Relevant publications

- H. Tang, G. Sun, N. Sebe, and L. Van Gool, Edge Guided GANs with Multi-Scale Contrastive Learning for Semantic Image Synthesis, IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(12): 14435-14452, December 2023 [54]
  Zenodo record: https://zenodo.org/records/11303045
- H. Tang, X. Qi, G. Sun, D. Xu, N. Sebe, R. Timofte, and L. Van Gool, Edge Guided GANs with Contrastive Learning for Semantic Image Synthesis, International Conference on Learning Representations (ICLR) 2023.[39]
  Zenodo record: https://zenodo.org/records/10037488

### 5.2.6. Relevant software/datasets/other outcomes

- The Pytorch implementation can be found in https://github.com/Ha0Tang/ECGAN

### 5.2.7. Relevance to AI4Media use cases and media industry applications

Our framework for semantic image synthesis is rather generic and is able to achieve significantly better results than existing models. Concretely, our approach could be useful for any media related application that requires high-quality realistic image synthesis.

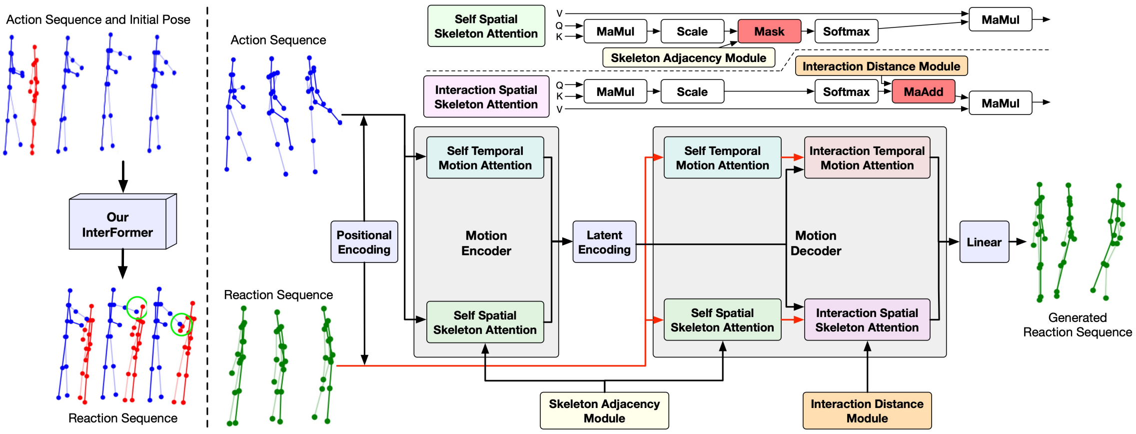



*Figure 6. **Left:** InterFormer during testing: given an action sequence (blue) and the first frame of a reaction sequence (red), we generate the full reaction sequence. We predict one frame at a time based on the previously generated frames. **Right:** Overview of InterFormer during training. The motion encoder takes an action sequence as the input and outputs a latent encoding. The motion decoder takes as inputs the reaction sequence corresponding to the action sequence and the latent encoding from the encoder. The motion decoder outputs a generated reaction sequence. Both the encoder and decoder contain several attention modules. **Top Right:** The skeleton adjacency and interaction distance modules interact directly with spatial attention.*

## 5.3. Interaction Transformer for Human Reaction Generation

**Contributing partner:** UNITN

### 5.3.1. Introduction

Modeling the dynamics of human motion is at the core of many applications in computer vision and robotics. Most works on human motion generation ignore human interactions and focus rather on the generation of actions of a single person. In addition, only a few works investigating human interaction generation [55] look at the reaction generation problem. What makes human reaction generation a challenging problem is the non-linearity in the temporal evolution of human motion and the two sources that condition the motion: the action and its corresponding reaction. The first issue arises because human motion is generally performed at varying evolution rates. In other words, a person performing the same activity will go roughly through the same stages but at slightly different rates every time. In addition, as stated by [56], unlike simple actions such as walking or running, complex human interactions such as duet dancing generate highly complex pose sequences operating close to the limit of human kinematics with very low periodicity. The second issue arises because the same action can have a different reaction depending on the interaction context, e.g., when reacting to a punch depending on the position, one can react more or less strongly. These two issues make the problem of reaction generation and evaluation challenging. Several questions arise as we try to tackle this challenge. How to translate action to reaction? How to model the long-term sequence? How to represent a complex action-reaction sequence?

### 5.3.2. Methodology

Our goal is to learn the reaction from a training sequence of actions and reactions by using Transformer architectures. The breakthroughs from Transformer networks in the Natural Language Processing (NLP) domain have sparked great interest in computer vision. Transformer architectures are based on a self-attention mechanism that learns the relationships between elements of a sequence. Unlike recurrent networks that process the elements of the sequence recursively, Transformers can attend to complete sequences and thereby are able to learn spatial and temporal relationships making them a good candidate for modeling human motion. In this research, we propose InterFormer, which with its spatial and temporal attention modules, is able not only to model the spatial and temporal dependencies in the action and in the reaction but also in the interaction between the two humans providing a solution to the two previously mentioned issues. Figure 6 (Left) shows how our InterFormer can generate a proper reaction sequence (red skeleton) by taking as input an action sequence (blue skeletons) and the initial position of the reaction sequence. Green circles highlight the reaction parts of the motion: the head goes backward in reaction to the punch; the hand is raised as the body continues to move backward to keep its balance.

#### 5.3.3. Experiments

##### 5.3.3.1. Datasets

**SBU Dataset** [57] contains 8 classes of simple interaction motions: walking toward, walking away, kicking, pushing, shaking hands, hugging, exchanging, and punching. The data which are too noisy, and in particular the class “hugging”, have been removed from this dataset. The “walking away” and “walking toward” classes have the same reactions (standing still), so we decided to fuse those two classes into a single “walking” class. This leaves us with 6 classes, 195 training and 30 test samples.
**K3HI Dataset** [58] contains the same 8 classes as SBU aside from the “hugging” class which is replaced by “pointing”. Also, unlike SBU, “approaching” and “departing” have reactions that are different, so we do not fuse the two classes. We also removed the noisy samples from the dataset but this time, we normalize the data in the same way as SBU was normalized by the authors. This leaves us with 236 training samples and 28 test samples.
**DuetDance Dataset** [56] contains 5 classes of dance motions: cha-cha, jive, rumba, salsa, and samba. Given the nature of the dataset, the motions are more complex than those in SBU and K3HI, and there are a lot of intra-class variabilities. We do not perform normalization, but since most samples are very long sequences (up to 160s), we decided to cut each sequence into smaller sequences of 50 frames (2s), leading to 273 training samples and 3,991 test samples.

For all three datasets, the poses are represented by their absolute 3D coordinates, furthermore, training and testing splits are selected randomly for fair comparisons. Duet-Dance was provided with neither train/test split nor subject information, and we used a random split. For the two others, the evaluation proposed by their respective authors is made using k-fold validation so we decided to split the dataset between train and test, randomly for K3HI and by selecting all the samples from a random subject for SBU.

**5.3.3.2. Evaluation Metrics** We use metrics commonly used in motion generation. Metrics used for motion prediction based on the distance between the generated sample and the ground truth are not fit for reaction generation as several different motions can be considered good reactions to the same action. While this choice of metric can seem contradictory with our losses that use direct comparison with the ground truth, it is important to understand that our evaluation metrics do not contain direct information about the skeleton that our network is supposed to generate and could not be efficiently used as losses.
**Classification Accuracy** measures how well our generated samples are classified by a motion classifier. We use the DeepGRU classifier [59]. We only train and test the classifier on the reaction part of the interaction, so the results are not influenced by the action, which is always the ground

truth. We report the percentage of correctly classified samples for each class and the average over the entire test set.

**Fréchet Video Distance** (FVD) is an adaptation of the Fréchet Inception distance (FID) [60] for video sequences [61]. FVD computes the distance between the ground truth and the generated data distribution.

$$\mathrm{FVD} = |\mu_{gt} - \mu_{gen}|^2 + \mathrm{tr}\left[\mathbf{C}_{gt}+\mathbf{C}_{gen}-2\left(\mathbf{C}_{gt} * \mathbf{C}_{gen}\right)^{1/2}\right], \tag{1}$$

where $\mu_{gt}$, $\mu_{gen}$ and $\mathbf{C}_{gt}$ and $\mathbf{C}_{gen}$ are the means and covariance matrices of the deep features from ground truth and the generated samples respectively, $\mathrm{tr}(\cdot)$ is the trace. The deep features are obtained from the classifier used for the classification accuracy.

**Diversity Score.** Following the metric defined by [62, 63], we compute the average deep feature distance between all the samples generated by each method and then compare it to the average deep feature distance of the ground truth. A low diversity score means that the generated samples have a diversity close to that of the ground truth and a high score means that the diversity is either lower (all motions are more similar) or higher (more noise in the generation). The average deep feature distance is calculated as follows:

$$div = \frac{1}{b(b-1)} \sum_{i=1}^{b} \sum_{j=1}^{b} ||F_i - F_j||_2, \tag{2}$$

where $b$ is the number of samples considered, $F_i$ and $F_j$ are deep features of the samples $i$ and $j$, respectively. The score is obtained using $div_{gt}$, the diversity distance of the ground truth and $div_{gen}$, the diversity of the generated samples.

$$score = 100 \times \frac{|div_{gt} - div_{gen}|}{div_{gt}}. \tag{3}$$

**5.3.3.3. Quantitative Evaluation.** Table 8 (left) shows the classification accuracy for SBU, DuetDance, and K3HI. Our method outperforms the five others on all the datasets. For SBU, we obtain results very close to the ground truth, and we outperform the other methods on all classes but "exchanging" where [55] get better results and vastly outperform the simple ZeroV baseline [64]. InterFormer is able to generate simple motions that are realistic enough to be correctly classified. We can see however, that on "kicking" we score less than ZeroV; this is due to the small size of the SBU dataset. A few misclassifications will cause a sharp drop in classification accuracy, and as we can see, "Kicking" is the class that has the lowest accuracy on the ground truth as the reaction can be similar to those of punching and pushing. The good performance of ZeroV in some classes can be explained by the fact that the overall accuracy is below chance (16.7%). This means that the classifier is unable to properly classify the motion from ZeroV as it only shows unmoving skeletons and for some classes, the two skeletons start in a neutral position that carries no information about the action. All these cause the classifier to fail at classifying the sample and likely classify many samples as "kicking", including some that are from the "kicking" class leading to the high score in this class.

For K3HI, we can see that the results are worse than for SBU for all methods and even for the ground truth. This is due to the very noisy nature of the K3HI dataset; even after removing the worse samples (that showed extreme deformation and no recognizable motion), the exchanging class has a 0% recognition rate even for the ground truth. However, our method provides better results than the two others in all classes except "approaching", which may be due to the noisy nature of the

| Method | GT | ZeroV [64] | VRNN [55] | MixMatch [65] | STT [66] | PGBIG[67] | InterFormer | GT | VRNN [55] | MixMatch [65] | InterFormer |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Classification Accuracy ↑ | | | | | | | User Preference ↑ | | | |
| | | | | | SBU | | | | | | |
| Walking | 100.0 | 0.0 | 58.3 | **100.0** | 91.7 | 58.3 | **100.0** | 34.2% | 21.4% | 15.5% | **28.9%** |
| Kicking | 66.7 | 66.7 | 0.0 | 0.0 | 0.0 | 0.0 | **33.3** | 38.8% | 23.8% | 5.6% | **31.8%** |
| Pushing | 80.0 | 0.0 | **60.0** | 0.0 | 0.0 | 0.0 | **60.0** | 35.6% | 19.7% | 15.4% | **29.3%** |
| Shaking Hands | 100.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | **100.0** | 37.5% | 21.8% | 7.8% | **32.9%** |
| Exchanging | 80.0 | 0.0 | **80.0** | 0.0 | 50.0 | 60.0 | 60.0 | 41.9% | 19.4% | 13.0% | **25.7%** |
| Punching | 100.0 | 33.3 | 0.0 | 33.3 | 0.0 | 0.0 | **100.0** | 43.1% | 19.3% | 11.3% | **26.3%** |
| Average | 90.0 | 10.0 | 46.7 | 43.3 | 40.0 | 33.3 | **80.0** | 38.5% | 20.9% | 11.4% | **29.2%** |
| | | | | | DuetDance | | | | | | |
| Cha-Cha | 28.0 | 1.8 | 26.4 | 19.2 | **37.1** | 28.6 | 26.7 | 45.9% | 17.8% | 5.5% | **30.8%** |
| Jive | 24.6 | 0.4 | 13.8 | **25.8** | 16.7 | 19.7 | 22.8 | 48.4% | 13.2% | 6.7% | **31.7%** |
| Rumba | 34.8 | 0.7 | **36.4** | 30.0 | 30.0 | 34.5 | 32.0 | 40.7% | 16.9% | 8.2% | **34.2%** |
| Salsa | 27.8 | 93.1 | **29.5** | 28.9 | 10.0 | 20.2 | 28.1 | 49.3% | 12.8% | 7.1% | **30.8%** |
| Samba | 22.2 | 18.6 | 21.0 | 17.2 | 18.2 | **24.4** | **24.4** | 44.5% | 15.8% | 6.3% | **33.6%** |
| Average | 28.0 | 24.6 | 26.2 | 24.9 | 24.4 | 25.7 | **27.1** | 45.8% | 15.3% | 6.7% | **32.2%** |
| | | | | | K3HI | | | | | | |
| Approaching | 100.0 | 25.0 | **75.0** | 50.0 | 50.0 | 0.0 | 0.0 | 34.9% | 23.1% | 14.1% | **27.9%** |
| Departing | 33.3 | 33.3 | 0.0 | **33.3** | **33.3** | **33.3** | **33.3** | 34.2% | 24.2% | 13.2% | **28.4%** |
| Kicking | 40.0 | 80.0 | 0.0 | 40.0 | 40.0 | 0.0 | **60.0** | 31.8% | 21.7% | 18.8% | **27.7%** |
| Pushing | 100.0 | 33.3 | 33.3 | 33.3 | 33.3 | 33.3 | **66.7** | 33.1% | 24.8% | 13.5% | **28.6%** |
| Shaking | 50.0 | 0.0 | 50.0 | 50.0 | **100.0** | **100.0** | 50.0 | 36.9% | 21.4% | 10.8% | **30.9%** |
| Exchanging | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 36.3% | 20.5% | 13.9% | **29.3%** |
| Punching | 100.0 | 25.0 | **50.0** | 25.0 | 0.0 | **50.0** | **50.0** | 33.9% | 21.7% | 16.9% | **27.5%** |
| Pointing | 100.0 | 50.0 | **100.0** | 0.0 | 25.0 | 25.0 | **100.0** | 37.3% | 20.2% | 16.1% | **26.4%** |
| Average | 67.9 | 35.7 | 39.3 | 28.6 | 32.1 | 25.0 | **46.4** | 34.8% | 22.2% | 14.7% | **28.3%** |

*Table 8. **Left**: Classification accuracy for each class of the SBU, DuetDance, and K3HI datasets. **Right**: User study for each class of the SBU, DuetDance, and K3HI datasets.*

data for this class. VRNN obtaining very high results in this class might be a consequence of the wrong classification present in many of the classes (a lot of samples are classified as approaching). For shaking, PGBIG [67] and STT [66] obtain better results but since the results are worse overall quantitatively and qualitatively this can be explained by the classifier putting many samples in that class as we have explained for ZeroV.

For DuetDance, the classification accuracy for all methods and the GT is much closer than for the other datasets. This is due to the complex motions contained in the dataset with a lot of intra-class variabilities. Furthermore, we use sequences of 50 frames which are short enough that some sequences from two different classes can be very similar. We can still notice that our method provides results that are the closest to the ground truth and that, unlike the five other methods no class has a score below chance (i.e., 20%) which means that our results are more consistent and closer to the ground truth, despite being beaten on some individual class e.g., STT score 37.1% on "cha-cha" but only 10.0% on "salsa" while we score 26.7% and 28.1%, respectively.

In Table 9 we show the FVD and diversity score for all methods on all datasets. We outperform VRNN, MixMatch [65], STT and PGBIG on the FVD measure, often by a large margin meaning that the features extracted by the classifier are closer to the features of the ground truth than for [55] and [65]. For the diversity score, we also outperform the two other methods and provide diversity that is close to that of the ground truth. We can see a significant increase in K3HI. This is due to the noisy nature of the dataset, which means that the diversity distance of the ground truth takes into account the noise of the sample, we, however, manage to score the closest to the diversity of the ground truth when compared to the other methods, without generating noisy samples. This can also explain why PGBIG diversity is better than ours despite performing much worse in terms of classification and qualitative results.

**User Study.** To evaluate the quality of the generated videos, we also conduct a user study. Specifically, the users are given four videos (two generated by existing methods VRNN and MixMatch,

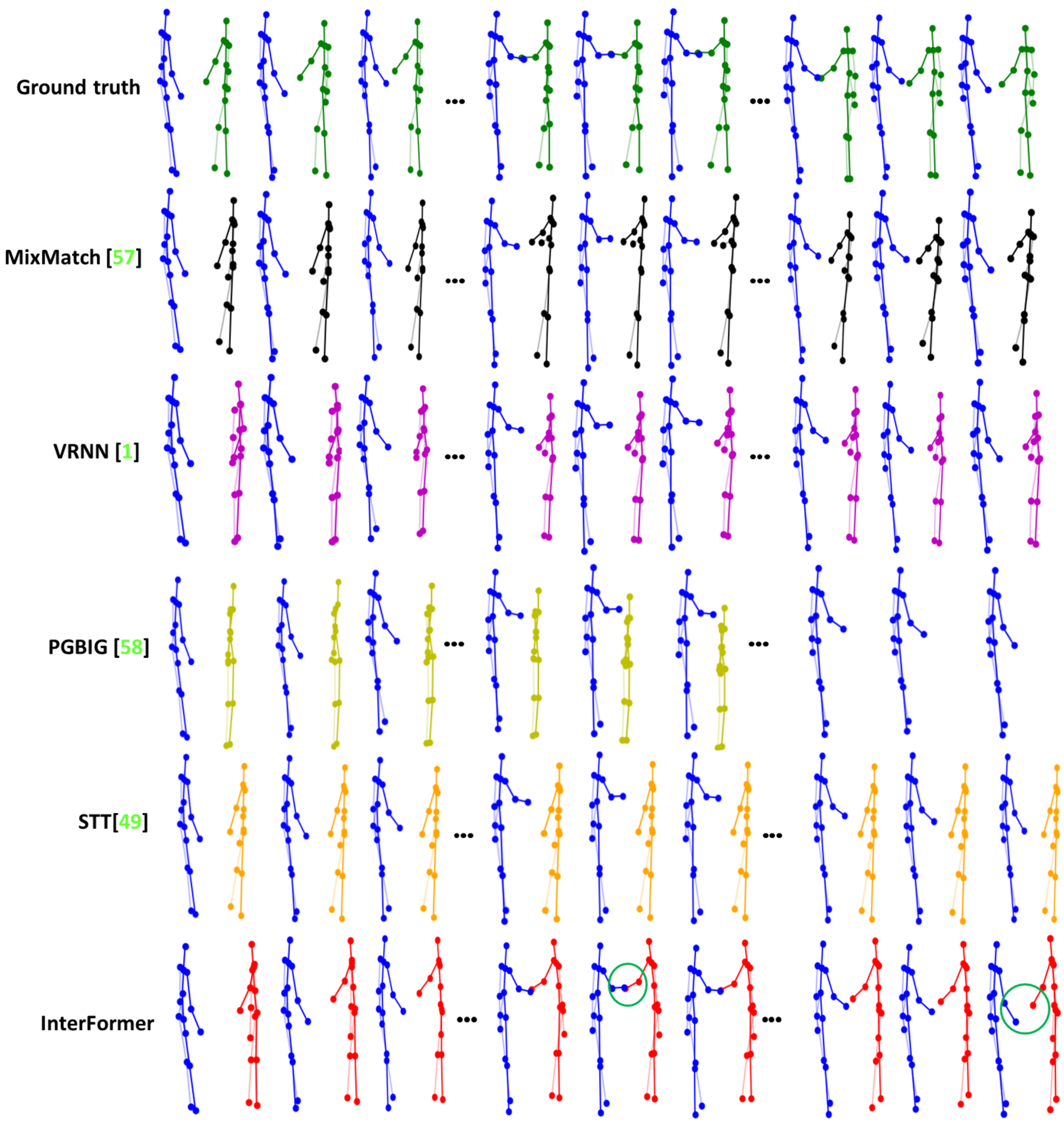



*Figure 7.* ***Qualitative results.*** *In blue the action motion is used as a condition. In other colors, the reaction is either from the ground truth or generated by the different models. Shaking hands class from the SBU dataset.*

one generated by our proposed InterFormer, and one real video) with the corresponding class label. Each participant needs to answer one question: 'Which video is more realistic regardless of the input label?'. 20 users have unlimited time to select their choices. PGBIG and STT are not represented in this study due to the extremely low quality of the results . The results are shown in Table 8 (right). We can see that the users show more preference for our method than the other two methods, which indicates the results generated by our method are more realistic.

| Method | FVD ↓ | | | Diversity ↓ | | |
|---|---|---|---|---|---|---|
| | SBU | DuetDance | K3HI | SBU | DuetDance | K3HI |
| ZeroV [64] | 493.3 | 41058.1 | 392.1 | 65.1 | 47.2 | 19.3 |
| VRNN [55] | 113.61 | 789.23 | 195.47 | 11.5 | 6.1 | 16.8 |
| MixMatch [65] | 314.38 | 1460.44 | 406.63 | 45.3 | 0.9 | 32.2 |
| STT[66] | 321.04 | 2610.95 | 7579.87 | 47.8 | 3.9 | 27.6 |
| PGBIG[67] | 267.27 | 317.0 | 379.4 | 35.7 | 1.5 | **10.1** |
| InterFormer (Ours) | **48.78** | **31.81** | **125.40** | **0.9** | **0.4** | 13.7 |

*Table 9. FVD and diversity on all datasets.*

**5.3.3.4. Qualitative Evaluation.** In Figure 7, we show an interaction from the "shaking hands" class of SBU. It shows that our method is able to generate the motion better than the two other methods. For [65], the character raises its hand to shake but never comes really close to the other character's hand and also shifts its entire body backward toward the end of the sequence. [55] generates a motion that raises slightly the hand but is then stuck in this position. [67] does not generate a shaking hand motion and fails to generate poses for the entire length of the action. STT [66] also fails to generate a shaking hand motion. Our method generates motion that is very close to the ground truth and contains the three main steps of the motion: raising the hand, shaking, and going back to starting position.

The very poor performances of PGBIG [67] and STT [66], our two baselines with unmodified code, can be explained by the fact that they were designed for human motion prediction. With human motion prediction, we seek to reduce as much as possible the discontinuities between the input and the output while we want to generate a different skeleton to the one used as input which implies a very strong discontinuity. Also, methods for human motion prediction are typically trained to always take the motion of the same duration as input and predict sequences that always have the same length e.g., the input of $500ms$ to predict $1s$ of motion. With reaction generation, the length of the sequences can vary (greatly in the case of K3HI) and the unmodified motion prediction method might struggle with the varying lengths. This is illustrated by the early stop in the generation of [67] in Figure 7 but also by the fact that [66] is unable to stop generating until it reaches the maximum sequence length of the dataset (not pictured in our figures).

### 5.3.4. Limitations

InterFormer presents two main limitations: (i) Due to the huge variability of complex motions, it is hard to stay true to the ground truth, making it difficult to evaluate the results in these cases; (ii) We are able to generate realistic motion for long sequences (tested up to 40 seconds). To do this, we cut the action sequence into smaller sub-sequences that we use for generation. We then generate all these sequences the same way as we do for shorter sequences. Only for the second sub-sequence onward the first frame used to give the initial position does not come from ground truth but instead is the last generated frame from the previous sub-sequence. In this way, InterFormer is able to generate reaction sequences for long motion. However, due to the accumulation of errors over time, the generation diverges more and more from the ground truth up to the point where it is hard to know how much action is taken into account in the generation. It is even more true that very long motions are usually complex ones, which means we also face the first limitation.

#### 5.3.5. Conclusion

The main contributions of this work are as follows:

- We propose a novel Interaction Transformer framework for the challenging human reaction generation task. To the best of our knowledge, this is the first work that challenges the task of human reaction prediction given the action of the interacting skeleton using a Transformer based architecture.
- We formulate the reaction generation problem as a translation problem, where we translate a given action of a skeleton to its corresponding reaction such that the entire interaction looks coherent and natural.
- We adopt a graph representation for self-attention to better exploit the skeleton structure while we ignore this representation for computing the attention between the two interacting skeletons. In this case, instead of a graph representation, we exploit the distance between the interacting joints assuming that closer joints involve stronger interaction. By introducing this distance, we provide the prior knowledge that helps to model the interaction.
- While the previous methods for interaction generation address limited and simple short-term interactions, we evaluate our method on the DuetDance dataset that provides more complex and long-term interactions.

#### 5.3.6. Relevant publications

- B. Chopin, H. Tang, N. Otberdout, M. Daoudi, and N. Sebe, Interaction Transformer for Human Reaction Generation, IEEE Transactions on Multimedia, 25:8842-8854, December 2023. [68]
  Zenodo record: https://zenodo.org/records/11302818

#### 5.3.7. Relevant software/datasets/other outcomes

- The Pytorch implementation can be found in https://github.com/CRISTAL-3DSAM/InterFormer

#### 5.3.8. Relevance to AI4Media use cases and media industry applications

InterFormer is a novel human reaction generation Transformer and as such is useful when action generation is needed. Concretely, our approach could be useful for any media related application that involves action generation, e.g. in gaming or animation.

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

## 6.1. AI-generated image detection based on CLIP's intermediate layers

**Contributing partner:** CERTH

### 6.1.1. Introduction

Generative Adversarial Networks (GANs), their many improvements [2, 69], and the latest breed of diffusion models [70] can synthesize highly realistic images that are mostly indistinguishable from real ones to the human eye [71]. Unfortunately, this carries risks ranging from fake pornography to hoaxes, identity theft, and financial fraud [72]. As a result, several machine learning solutions for detection have been recently proposed, based on spatial [73, 74], frequency-level [75, 76, 77], and deep learning features derived from large-scale foundation models [78]. Despite the sophistication of these techniques, especially the ones using deep learning, generalization to unseen image synthesis methods is still a big challenge, as elaborated in Section 4 of deliverable D6.3.

In this context, we propose RINE (Representations from Intermediate Encoder-blocks), which fuses the image representations provided by CLIP's intermediate Transformer blocks carrying low-level visual information, and projects them with learnable linear mappings to a synthetic image-aware vector space, capable of generalizing exceptionally well. This approach can generalize well to a wide range of unseen generators and even different families of generators (i.e., trained on GAN images but succeeding in detecting Diffusion Model images).

### 6.1.2. Methodology

Figure 8 presents the proposed architecture. A batch of $b$ input images $\mathbf{X} \in \mathbb{R}^{b\times 3\times w\times h}$, with $w$ width and $h$ height, is first reshaped into a sequence of $p$ flattened image patches $\mathbf{X}_p \in \mathbb{R}^{b\times p\times(P^2\cdot 3)}$, where $P$ denotes patch side length and $p = w\cdot h/P^2$. Then, a $d$-dimensional linear mapping is considered on top of $\mathbf{X}_p$, the learnable $d$-dimensional CLS token is concatenated to the projected sequence, and positional embeddings are added to all $p+1$ tokens in order to construct the CLIP's input, $\mathbf{Z}_0 \in \mathbb{R}^{b\times(p+1)\times d}$. Finally, the tensor $\mathbf{K}$ containing the representations from intermediate encoder-blocks of CLIP is the concatenation of the $n$ CLS tokens stemming from each of the $n$ Transformer blocks:

$$\mathbf{K} = \oplus\{\mathbf{Z}_l^{[0]}\}_{l=1}^{n} \in \mathbb{R}^{b\times n\times d} \tag{4}$$

where $\oplus$ denotes concatenation, and $\mathbf{Z}_l^{[0]} \in \mathbb{R}^{b\times 1\times d}$ denotes the CLS token from the output of Transformer block $l$. We keep CLIP, frozen during training and we use $\mathbf{K}$ to construct discriminative features.

The extracted representations $\mathbf{K}$ are processed by a projection network ($\mathfrak{Q}_1$ in Figure 8):

$$\mathbf{K}_m = \mathrm{ReLU}(\mathbf{K}_{m-1}\mathbf{W}_m + \mathbf{b}_m) \in \mathbb{R}^{b\times n\times d'} \tag{5}$$

where $m$=1,...,q denotes the index of the network's layer (with $\mathbf{K}_0 = \mathbf{K}$), $\mathbf{W}_m \in \mathbb{R}^{d'\times d'}$ (except $\mathbf{W}_1 \in \mathbb{R}^{d\times d'}$) and $\mathbf{b}_m \in \mathbb{R}^{d'}$ define the linear mapping, and ReLU denotes the Rectified Linear Unit [79] activation function. After each layer, dropout [80] with rate 0.5 is applied.

Additionally, each of the $d'$ learned features can be more relevant for the task at various processing stages, thus we employ a Trainable Importance Estimator (TIE) module to adjust their impact to the final decision. More precisely, we consider a randomly-initialized learnable variable $\mathbf{A} = \{\alpha_{lk}\} \in \mathbb{R}^{n\times d'}$ the elements of which will estimate the importance of feature $k$ at processing

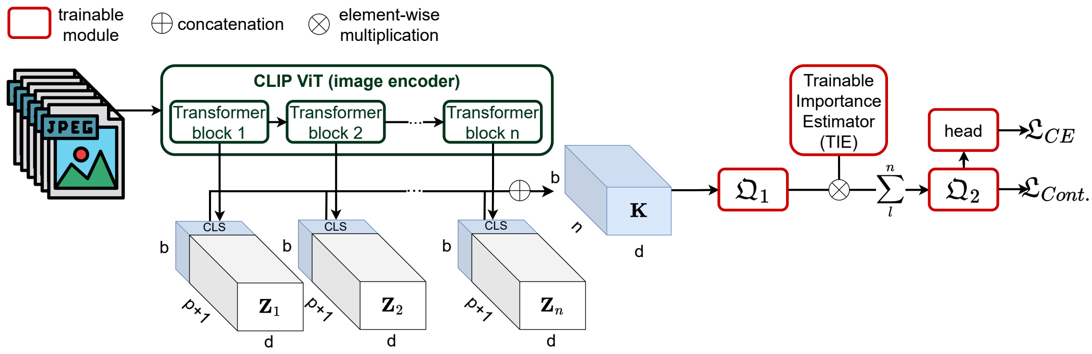



*Figure 8. The RINE architecture. A batch of b images is processed by CLIP's image encoder. The concatenation of the n d-dimensional CLS tokens (one from each intermediate stage, i.e., Transformer block) is first projected and then multiplied with the trainable importance scores of the intermediate stages, estimated by the Trainable Importance Estimator (TIE) module. Summation across the second dimension results in one feature vector per image. Finally, after the second projection and the consequent classification head modules, two loss functions are computed. The binary cross-entropy $\mathfrak{L}_{CE}$ directly optimizes the real vs. fake objective, while the contrastive loss $\mathfrak{L}_{Cont.}$ assists the training by forming a dense feature vector cluster per class.*

stage (i.e., Transformer block) $l$. This is then used to construct one feature vector per image, as a weighted average of features across processing stages:

$$\tilde{\mathbf{K}} = \sum_{l}^{n} \mathcal{S}(\mathbf{A}) \otimes \mathbf{K}_q \tag{6}$$

where $\mathcal{S}$ denotes the Softmax activation function acting across the first dimension of $\mathbf{A}$, $\otimes$ denotes broadcasted Hadamard product, and summation is conducted across the dimension of Transformer blocks. Finally, a second projection network ($\mathfrak{Q}_2$ in Figure 8) with the same architecture as the first one (cf. Equation 5) takes as input $\tilde{\mathbf{K}}$ and outputs $\tilde{\mathbf{K}}_q$, which is consequently processed by the classification head that predicts the final output (probability to be fake). The classification head consists of two $d' \times d'$ ReLU-activated dense layers and one $d' \times 1$ dense layer that produces the logits.

We consider the combination of two objective functions to optimize the parameters of the proposed model. The first is the binary cross-entropy loss [81], here denoted as $\mathfrak{L}_{CE}$, which measures the classification error and directly optimizes the main real vs. fake objective. The second is a contrastive loss function, specifically the Supervised Contrastive Learning loss [82], which we here denote as $\mathfrak{L}_{Cont.}$, and is considered in order to assist the training process by bringing closer the feature vectors inside $\tilde{\mathbf{K}}_q$ that share targets and move apart the rest. We combine the two objective functions by a tunable factor $\xi$, as shown in Equation 7:

$$\mathfrak{L} = \mathfrak{L}_{CE} + \xi \cdot \mathfrak{L}_{Cont.} \tag{7}$$

### 6.1.3. Experiments

Following the training protocol of related work [78, 73, 84], we use ProGAN [87] generated images and the corresponding real images of the provided dataset for training. Similarly with previous works [84, 83], we consider three settings with ProGAN-generated training data from 1 (horse), 2 (chair, horse) and 4 (car, cat, chair, horse) object classes. For testing, we consider 20 datasets with generated and real images combining the evaluation datasets used in [78, 73, 84]. Specifically,

| | | Generative Adversarial Networks | | | | | | | | Low level vision | | Perceptual loss | | | Latent Diffusion | | | Glide | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| method | # cl. | Pro-GAN | Style-GAN | Style-GAN2 | Big-GAN | Cycle-GAN | Star-GAN | Gau-GAN | Deep-fake | SITD | SAN | CRN | IMLE | Guided | 200 steps | 200 CFG | 100 steps | 100 27 | 50 27 | 100 10 | DALL-E | AVG |
| Wang [73] (prob. 0.5) | 20 | **100.0** | 66.8 | 64.4 | 59.0 | 80.7 | 80.9 | 79.2 | 51.3 | 55.8 | 50.0 | 85.6 | 92.3 | 52.1 | 51.1 | 51.4 | 51.3 | 53.3 | 55.6 | 54.2 | 52.5 | 64.4 |
| Wang [73] (prob. 0.1) | 20 | **100.0** | 84.3 | 82.8 | 70.2 | 85.2 | 91.7 | 78.9 | 53.0 | 63.1 | 50.0 | 90.4 | 90.3 | 60.4 | 53.8 | 55.2 | 55.1 | 60.3 | 62.7 | 61.0 | 56.0 | 70.2 |
| Patch-Forensics [74] | † | 66.2 | 58.8 | 52.7 | 52.1 | 50.2 | 96.9 | 50.1 | 58.0 | 54.4 | 50.0 | 52.9 | 52.3 | 50.5 | 51.9 | 53.8 | 52.0 | 51.8 | 52.1 | 51.4 | 57.2 | 55.8 |
| FrePGAN [83] | 1 | 95.5 | 80.6 | 77.4 | 63.5 | 59.4 | 99.6 | 53.0 | 70.4 | -* | - | - | - | - | - | - | - | - | - | - | - | - |
| FrePGAN [83] | 2 | 99.0 | 80.8 | 72.2 | 66.0 | 69.1 | 98.5 | 53.1 | 62.2 | - | - | - | - | - | - | - | - | - | - | - | - | - |
| FrePGAN [83] | 4 | 99.0 | 80.7 | 84.1 | 69.2 | 71.1 | **99.9** | 60.3 | 70.9 | - | - | - | - | - | - | - | - | - | - | - | - | - |
| LGrad [84] | 1 | 99.4 | 96.1 | 94.0 | 79.6 | 84.6 | 99.5 | 71.1 | 63.4 | 50.0 | 44.5 | 52.0 | 52.0 | 67.4 | 90.5 | 93.2 | 90.6 | 80.2 | 85.2 | 83.5 | 89.5 | 78.3 |
| LGrad [84] | 2 | 99.8 | 94.5 | 92.1 | 82.5 | 85.5 | 99.8 | 73.7 | 61.5 | 46.9 | 45.7 | 52.0 | 52.1 | 72.1 | 91.1 | 93.0 | 91.2 | 87.1 | 90.5 | 89.4 | 88.7 | 79.4 |
| LGrad [84] | 4 | 99.9 | 94.8 | **96.1** | 83.0 | 85.1 | 99.6 | 72.5 | 56.4 | 47.8 | 41.1 | 50.6 | 50.7 | 74.2 | 94.2 | **95.9** | 95.0 | 87.2 | 90.8 | 89.8 | 88.4 | 79.7 |
| DMID [85] | 20 | **100.0** | **99.4** | 92.9 | 96.9 | 92.0 | 99.5 | 94.8 | 54.1 | 90.6** | 55.5 | **100.0** | **100.0** | 53.9 | 58.0 | 61.1 | 57.5 | 56.9 | 59.6 | 58.8 | 71.7 | 77.6 |
| UFD [78] | 20 | 99.8 | 79.9 | 70.9 | 95.1 | 98.3 | 95.7 | 99.5 | 71.7 | 71.4 | 51.4 | 57.5 | 70.0 | 70.2 | 94.4 | 74.0 | 95.0 | 78.5 | 79.0 | 77.9 | 87.3 | 80.9 |
| | 1 | 99.8 | 88.7 | 86.9 | 99.1 | **99.4** | 98.8 | 99.7 | **82.7** | 84.7 | **72.4** | 93.4 | 96.9 | **77.9** | 96.9 | 83.5 | 97.0 | 83.8 | 87.4 | 85.4 | 91.9 | 90.3 |
| RINE (Ours) | 2 | 99.8 | 84.9 | 76.7 | 98.3 | **99.4** | 99.6 | **99.9** | 66.7 | **91.9** | 67.8 | 83.5 | 96.8 | 69.6 | 96.8 | 80.0 | 97.3 | 83.6 | 86.0 | 84.1 | 92.3 | 87.7 |
| | 4 | **100.0** | 88.9 | 94.5 | **99.6** | 99.3 | 99.5 | 99.8 | 80.6 | 90.6 | 68.3 | 89.2 | 90.6 | 76.1 | **98.3** | 88.2 | **98.6** | **88.9** | **92.6** | **90.7** | **95.0** | **91.5** |

* Hyphens denote scores that are neither reported in the corresponding paper nor the code and models are available in order to compute them.
** We applied cropping at 2000x1000 on SITD [86] for DMID [85] due to GPU memory limitations.
† Patch-Forensics has been trained on ProGAN data but not on the same dataset as the rest models. For more details please refer to [74].

*Table 10. Accuracy (ACC) scores of baselines and our model across 20 test datasets. The second column (# cl.) presents the number of used training classes. Best performance is denoted with **bold** and second to best with underline. Our method yields +10.6% average accuracy compared to the state-of-the-art.*

| | | Generative Adversarial Networks | | | | | | | | Low level vision | | Perceptual loss | | | Latent Diffusion | | | Glide | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| method | # cl. | Pro-GAN | Style-GAN | Style-GAN2 | Big-GAN | Cycle-GAN | Star-GAN | Gau-GAN | Deep-fake | SITD | SAN | CRN | IMLE | Guided | 200 steps | 200 CFG | 100 steps | 100 27 | 50 27 | 100 10 | DALL-E | AVG |
| Wang [73] (prob. 0.5) | 20 | **100.0** | 98.0 | 97.8 | 88.2 | 96.8 | 95.4 | 98.1 | 64.8 | 82.2 | 56.0 | 99.4 | 99.7 | 69.9 | 65.9 | 66.7 | 66.0 | 72.0 | 76.5 | 73.2 | 66.3 | 81.7 |
| Wang [73] (prob. 0.1) | 20 | **100.0** | 99.5 | 99.0 | 84.5 | 93.5 | 98.2 | 89.5 | 87.0 | 68.1 | 53.0 | 99.5 | 99.5 | 73.2 | 71.2 | 73.0 | 72.5 | 80.5 | 84.6 | 82.1 | 71.3 | 84.0 |
| Patch-Forensics [74] | † | 94.6 | 79.3 | 77.6 | 83.3 | 74.7 | 99.5 | 83.2 | 71.3 | 91.6 | 39.7 | 99.9 | 98.9 | 58.7 | 68.9 | 73.7 | 68.7 | 50.6 | 52.8 | 48.4 | 66.9 | 74.1 |
| FrePGAN [83] | 1 | 99.4 | 90.6 | 93.0 | 60.5 | 59.9 | **100.0** | 49.1 | 81.5 | -* | - | - | - | - | - | - | - | - | - | - | - | - |
| FrePGAN [83] | 2 | 99.9 | 92.0 | 94.0 | 61.8 | 70.3 | **100.0** | 51.0 | 80.6 | - | - | - | - | - | - | - | - | - | - | - | - | - |
| FrePGAN [83] | 4 | 99.9 | 89.6 | 98.6 | 71.1 | 74.4 | **100.0** | 71.7 | 91.9 | - | - | - | - | - | - | - | - | - | - | - | - | - |
| LGrad [84] | 1 | 99.9 | 99.6 | 99.5 | 88.9 | 94.4 | **100.0** | 82.0 | 79.7 | 44.1 | 45.7 | 82.3 | 82.5 | 71.1 | 97.3 | 98.0 | 97.2 | 90.1 | 93.6 | 92.0 | 96.9 | 86.7 |
| LGrad [84] | 2 | **100.0** | 99.6 | 99.6 | 92.6 | 94.7 | 99.9 | 83.2 | 71.6 | 42.4 | 45.3 | 66.1 | 80.9 | 75.6 | 97.2 | 98.1 | 97.2 | 94.6 | 96.5 | 95.8 | 96.5 | 86.4 |
| LGrad [84] | 4 | **100.0** | 99.8 | 99.9 | 90.8 | 94.0 | **100.0** | 79.5 | 72.4 | 39.4 | 42.2 | 63.9 | 69.7 | 79.5 | 99.1 | **99.1** | 99.2 | 93.3 | 95.2 | 95.0 | 97.3 | 85.5 |
| DMID [85] | 20 | **100.0** | **100.0** | **100.0** | 99.8 | 98.6 | **100.0** | 99.8 | 94.7 | **99.8**** | 87.7 | **100.0** | **100.0** | 73.0 | 86.8 | 89.4 | 87.3 | 86.5 | 89.9 | 89.0 | 96.1 | 93.9 |
| UFD [78] | 20 | **100.0** | 97.3 | 97.5 | 99.3 | 99.8 | 99.4 | **100.0** | 84.4 | 89.9 | 62.6 | 94.5 | 98.3 | 89.5 | 99.3 | 92.5 | 99.3 | 95.3 | 95.6 | 95.0 | 97.5 | 94.3 |
| | 1 | **100.0** | 99.1 | 99.7 | **99.9** | **100.0** | **100.0** | **100.0** | 97.4 | 95.8 | 91.9 | 98.5 | 99.9 | 95.7 | 99.8 | 98.0 | **99.9** | **98.9** | **99.3** | **99.1** | 99.3 | 98.6 |
| RINE (Ours) | 2 | **100.0** | 99.5 | 99.6 | **99.9** | **100.0** | **100.0** | **100.0** | 96.4 | 97.5 | 93.1 | 98.2 | 99.8 | 95.7 | **99.9** | 98.0 | **99.9** | **98.9** | 99.0 | 98.8 | **99.6** | 98.7 |
| | 4 | **100.0** | 99.4 | **100.0** | **99.9** | **100.0** | **100.0** | **100.0** | **97.9** | 97.2 | **94.9** | 97.3 | 99.7 | **96.4** | 99.8 | 98.3 | **99.9** | 98.8 | **99.3** | 98.9 | 99.3 | **98.8** |

* Hyphens denote scores that are neither reported in the corresponding paper nor the code and models are available in order to compute them.
** We applied cropping at 2000x1000 on SITD [86] for DMID [85] due to GPU memory limitations.
† Patch-Forensics has been trained on ProGAN data but not on the same dataset as the rest models. For more details please refer to [74].

*Table 11. Average precision (AP) scores of baselines and our model across 20 test datasets. The second column (# cl.) presents the number of used training classes. Best performance is denoted with **bold** and second to best with underline. Our method yields +4.5% mean average precision (mAP) compared to the state-of-the-art.*

the evaluation sets are from ProGAN [87], StyleGAN [88], StyleGAN2 [89], BigGAN [90], CycleGAN [91], StarGAN [13], GauGAN [34], DeepFake [3], SITD [86], SAN [92], CRN [29], IMLE [93], Guided [94], LDM [95] (3 variants), Glide [96] (3 variants), and DALL-E [97]. The training of RINE is conducted with batch size 128 and learning rate 1e-3 for only 1 epoch using the Adam optimizer. We consider a hyperparameter grid, namely $\xi \in \{0.1, 0.2, 0.4, 0.8\}$, $q \in \{1, 2, 4\}$, and $d' \in \{128, 256, 512, 1024\}$, and two CLIP variants (for the extraction of representations), namely ViT-B/32 and L/14, to obtain the best performance. For the training images, following the best practice in related work, we apply Gaussian blurring and `JPEG` compression with probability 0.5, then random cropping to 224×224, and finally random horizontal flip with probability 0.5. The validation and test images are only center-cropped at 224×224. State-of-the-art methods are also evaluated in our analysis for comparison purposes, and ablation experiments are conducted to support our implementation choices. Performance is measured with accuracy (ACC) and average precision (AP) metrics.

Tables 10 and 11 present the performance scores (ACC & AP respectively) of our method versus the competing ones. Our 1-class model outperforms all state-of-the-art methods irrespective of training class number. On average, we surpass the state-of-the-art by +9.4% ACC & +4.3% AP

| | 1-class | | 2-class | | 4-class | | AVG | |
|---|---|---|---|---|---|---|---|---|
| | ACC | AP | ACC | AP | ACC | AP | ACC | AP |
| w/o contr. loss | 87.3 | 98.3 | **87.9** | 98.5 | 90.0 | 98.8 | 88.4 | 98.5 |
| w/o TIE | 85.0 | 97.8 | 86.4 | 98.5 | 90.5 | 98.8 | 87.3 | 98.3 |
| w/o intermediate | 78.9 | 93.1 | 81.1 | 94.7 | 82.5 | 94.8 | 80.8 | 94.2 |
| full | **90.3** | **98.6** | 87.7 | **98.7** | **91.5** | **98.8** | **89.8** | **98.7** |

*Table 12. Ablation analysis compares the full architecture with different instances of the architecture after removing the contrastive loss, the TIE module, and the intermediate representations.*

with the 1-class model, by +6.8% ACC & +4.4% AP with the 2-class model, and by +10.6% ACC & +4.5% AP with the 4-class model. In terms of ACC, we obtain the best score in 14 out of 20 test datasets, and simultaneously the first and second best performance in 10 of them. In terms of AP, we obtain the best score in 15 out of 20 test datasets, and simultaneously the first and second best performance in 14 of them. The biggest performance increase achieved by our method is on the SAN dataset [92] (+16.9% ACC). In Table 12, we present an ablation study, where we remove three of RINE's main components, namely the intermediate representations, the TIE, and the contrastive loss, one-by-one. To be more precise, "w/o intermediate" means that we use only the last layer's features (equivalent to the SotA method [78]). We measure the RINE's performance on 20 test datasets, after removing each of the three components, and present the average ACC and AP for the 1-, 2-, and 4-class models, as well as the average across the three models. The results demonstrate the positive impact of all proposed components, as the full architecture achieves the best performance. Also, ablating intermediate representations yields the biggest performance loss reducing the metrics to the previous state-of-the-art levels (i.e., [78]), as expected.

#### 6.1.4. Conclusion

The main contributions of this work are:

- The architecture RINE that leverages representations from intermediate encoder-blocks of CLIP in order to address the task of AI-generated image detection.
- A comprehensive experimental study demonstrating the superiority of RINE compared to the state of the art.
- The efficiency of the proposed method, requiring only one epoch, translating to ∼8 minutes of training time, and limited training data to achieve maximum performance.

#### 6.1.5. Relevant publications

- Koutlis, C., & Papadopoulos, S. (2024). Leveraging Representations from Intermediate Encoder-blocks for Synthetic Image Detection. arXiv preprint arXiv:2402.19091. (Accepted in ECCV 2024; Currently available at https://arxiv.org/abs/2402.19091)

#### 6.1.6. Relevant software

- Code and model checkpoints available at https://github.com/mever-team/rine

### 6.1.7. Relevance to AI4Media use cases and media industry applications

The method developed here contributes to UC1 (AI for Social Media and Against Disinformation), and specifically, Feature 1A (Detection/Verification of Synthetic Media). AI-generated image detection systems could be used by social media platforms to assess the authenticity of uploaded content, and by journalists to verify real-world content. The proposed method exhibits generalization ability to deepfakes and Diffusion models beyond GAN generators.

## 6.2. Manipulation identification of deepfake content

**Contributing partner:** CERTH

### 6.2.1. Introduction

Recent advancements in deepfake synthesis, along with the emergence and popularity of free open-source generators, have led to an unprecedented proliferation of false online information. This is especially alarming considering that fabricated content can be weaponised by malicious actors to disseminate disinformation and erode trust in public institutions. As a counterweight, the research community has invested considerable effort towards deepfake detection [98, 99]. In particular, a wide variety of AI approaches have been deployed, based on the understanding that they are superior to traditional forensics methods relying on statistical analysis and handcrafted features. Such methods however tend to perform well on the manipulation types and data used during training while generalising poorly in different settings. For example, [100] has benchformed the performance of 51 deepfake detectors and found that all of them exhibit less than 70% AUC when tested in the wild, leaving a big room for improvement.

While the majority of the literature has focused on deepfake detection, i.e., asserting that a media (image or video) is fake or authentic, there has been little attention to the problem of *attribution*, i.e., associating a media with a known manipulation technique. Attribution can be trivially used for detection, offering an aspect of explainability that is missing from current detectors, with potentially superior performance according to promising preliminary evidence. This is the case in the intra-dataset evaluation, where attribution has been successfully applied to disentangle face-swaps based on different auto-encoder settings [101] and the artifacts of different manipulation methods of the FF++ dataset [102]. [103] has also demonstrated that CNN attribution models can generalize better to cross-dataset manipulations and their discriminative ability can be enhanced with contrastive techniques that spread apart the features of different manipulations in the embedding space. Contrastive techniques are gaining traction in computer vision [104, 105] and have been successfully applied to (binary) deepfake detection [106, 107, 108] yet, despite their merits, both contrastive techniques and attribution have been under-explored in the deepfake literature.

Based on this research gap, we have performed a comprehensive investigation of attribution and contrastive techniques for deepfake detection. To this end, we have compared the performance of 4 state-of-the-art backbone models, trained for deepfake attribution and binary detection, across 6 deepfake datasets. Special attention has been paid to the cross-dataset performance, distinguishing the cases of seen and unseen manipulations, intending to gauge the applicability of state-of-the-art models "in the wild". Finally, we have investigated the benefits of contrastive techniques on generalization.

### 6.2.2. Methodology

Our approach is structured around 3 research questions (RQs):

| Dataset | #Real | #Fake | Manipulations |
|---|---|---|---|
| FF++ [112] | 1,000 | 1,000 | FaceSwap, DeepFakes, Face2Face, NeuralTextures |
| CelebDF [113] | 500 | 5,639 | DeepFakes |
| FakeAVCeleb [114] | 500 | 19,500 | FaceSwap, FSGAN, Wav2Lip, SV2TTS |
| DFDC [115] | ∼24K | ∼105K | faceswaps (undisclosed methods) |
| ForgeryNet [116] | ∼100K | ∼120K | FaceShifter, FSGAN, DeepFakes, BlendFace MMR, DSS, TalkingHeads, ATVG-Net |
| DF-Platter [117] | 764 | ∼132K | FaceSwap, FSGAN, FaceShifter |

*Table 13. Selected datasets*

RQ1: How do attribution and binary models compare in terms of cross-dataset generalization?

RQ2: Do attribution models maintain intra-manipulation performance in cross-dataset settings?

RQ3: Do contrastive methods in training improve attribution performance?

To address these RQs we consider deepfake detection and attribution with neural networks of different backbones operating at the image/frame level. Detection models output binary decisions if an image is fake or real, while attribution models output multi-class labels corresponding to a list of known manipulations and the real class. For training and evaluation, we consider multiple datasets with known manipulations as labels, some of which are common andothers unique to each dataset. We describe the approach of each RQ in more detail below.

**RQ1**: Since the manipulations of the training and the test datasets may not coincide, we convert multi-class labels and predictions to binary, allowing for a straightforward comparison of attribution and detection models. Given a dataset with $N$ manipulation methods, a multi-class label $l_{\text{mc}}$ takes values in $\{0, 1, \ldots, N\}$, where 0 corresponds to "real" images and 1 to $N$ to manipulation methods. It can be easily converted to a binary label $l_{\text{bin}}$ by setting $l_{\text{bin}} = 1$ if $l_{\text{mc}} > 0$. Similarly, the predictions of multi-class attribution models are probability distributions $\mathbf{p}_{\text{mc}}$ over $N+1$ labels. Denoting by $p_{\max}$ the maximum probability at $j = \underset{i}{arg\,max}\ \mathbf{p}_{\text{mc}}[i]$, we can infer a binary prediction $p_{\text{bin}}$ as:

$$p_{\text{bin}} = \begin{cases} p_{\max} & \text{if } j = 0 \text{ and } p_{\max} < 0.5, \\ 1 - p_{\max} & \text{if } j = 0 \text{ and } p_{\max} \geq 0.5, \\ p_{\max} & \text{if } j \neq 0. \end{cases}$$

**RQ2**: We first identify all pairs of datasets with common manipulation methods. Then, for each ordered pair $(d_i, d_j)$ we train a backbone network on the whole $d_i$ and evaluate performance on the data of $d_j$ with each common manipulation separately. This systematic procedure guarantees the investigation of all possible data combinations.

**RQ3**: We retrain the models of RQ1 with contrastive loss functions and compare the vanilla and contrastive implementations. The considered loss functions are Triplet loss [109], NT-Xent loss [105, 110, 111], and SupCon loss [104].

### 6.2.3. Experiments

**Datasets:** We considered FaceForensics++ (FF++), CelebDF-V2 (CelebDF), FakeAVCeleb, DFDC, ForgeryNet, and DF-Platter (see Table 13). The common manipulations are DeepFakes, FaceSwap, FaceShifter, and FSGAN. We note that we used the low compression (c23) videos of FF++, the video part of FakeAVCeleb and ForgeryNet, and the set A of DF-Platter containing only single subjects. For pre-processing, we sampled 1 frame per second from each video with *FFMPEG*[118], and extracted faces with the RetinaFace [119] face detector.

**Models:** We considered the CNN models EfficientNetV2 and ConvNextV2, and the transformer models Pyramid Vision Transfomer V2 (PVT-V2) and SwinV2 (see Table 14). Both architectures are well-established in computer vision, in particular, EfficientNetV2 has been extensively utilized in the deepfake literature [120, 121, 122, 123, 124], ConvNextV2 and SwinV2 have shown exceptional performance in downstream vision tasks [125, 125], and PVT-V2 is a computationally efficient version of the standard vision transformer. For implementation, we used the PyTorch and PyTorch Lightning libraries, and trained the models on NVIDIA GPUs with random cropping, AugMix [126], and horizontal flipping for augmentations.

**Evaluation Metrics:** We used AUC, EER, and Balanced Accuracy for a comprehensive view of each model's effectiveness across different conditions. These metrics are chosen for their ability to balance type I and type II errors, imbalanced classes, and for compatibility with the deepfake detection literature.

**Results:** Table 15 shows the results for RQ1. We see that binary models exhibit consistently better cross-dataset generalization than multi-class detectors, which can be attributed to the increased diversity of samples per class in binary models.

Table 16 shows the results for RQ2. We see that all models struggle to detect seen manipulations in unseen datasets. Larger models, such as ConvNext and Swin, experience significant drops in accuracy, particularly in FaceSwap manipulations where all models fail (accuracy $0-1\%$). However, generalization is improved in models trained on newer or higher-quality datasets like DF-Platter and CelebDF, notably, in the case of FSGAN manipulation. Similarly, models trained on DF-Platter achieve an accuracy of $43-45\%$ in detecting FaceShifter deepfakes on ForgeryNet, while those trained on FaceForensics++ are only $5-15\%$ accurate. Aggregating the results for CCN and transformer architectures actually reveals that CNNs outperform ViTs in terms of accuracy across all scenarios and manipulations, except for the intra-dataset evaluation of FaceShifter and FaceSwap where ViTs are slightly better.

Finally, Table 17 shows the results for RQ3. Contrastive methods offer minimal gains for smaller networks like EfficientNet and PVT across all metrics, with a slight improvement for PVT when tested on CelebDF. In contrast, the Swin model outperforms the baseline in terms of generalization to CelebDF and DFDC. Conversely, while ConvNext performs below the baseline on CelebDF, it shows significant improvement on DFDC. This indicates that a well-designed contrastive formulation could yield superior results.

| Category | Model | Input Size | Parameters |
|---|---|---|---|
| CNN | EfficientNetV2 B0 [127] | $(192 \times 192 \times 3)$ | 5.87M |
| | ConvNextV2 Tiny [128] | $(384 \times 384 \times 3)$ | 27.9M |
| Transformer | PyramidNetV2 B0 [129] | $(224 \times 224 \times 3)$ | 3.41M |
| | SwinV2 Tiny [130] | $(256 \times 256 \times 3)$ | 27.58M |

*Table 14. Selected models*

| Setting\Dataset | FaceForensics++ | | | Celeb-DF | | | DFDC | | |
|---|---|---|---|---|---|---|---|---|---|
| | AUC (%) | BA (%) | EER ↓ (%) | AUC (%) | BA (%) | EER ↓ (%) | AUC (%) | BA (%) | EER ↓ (%) |
| | | | | EfficientNetV2 | | | | | |
| Binary | 98.81 | 95.24 | 4.02 | **69.19** | **63.30** | **36.70** | **69.31** | **62.91** | **36.84** |
| Multiclass | **99.20** | **95.66** | **3.64** | 60.56 | 57.95 | 42.08 | 65.47 | 61.26 | 38.52 |
| | | | | PVT-V2 | | | | | |
| Binary | 98.95 | 95.50 | 3.91 | **69.51** | **63.81** | **36.07** | **68.67** | **63.25** | **36.66** |
| Multiclass | **99.47** | **96.69** | **2.93** | 59.53 | 54.59 | 43.27 | 63.02 | 59.33 | 40.36 |
| | | | | ConvNextV2 | | | | | |
| Binary | 99.60 | 98.08 | **1.62** | **62.13** | **56.96** | **38.98** | **63.05** | **59.49** | **38.70** |
| Multiclass | **99.83** | **98.17** | 1.71 | 54.70 | 55.56 | 46.26 | 57.77 | 57.03 | 44.41 |
| | | | | SwinV2 | | | | | |
| Binary | 98.83 | 97.06 | 2.18 | **70.71** | **63.91** | **34.01** | **67.23** | **62.75** | **36.76** |
| Multiclass | **99.75** | **97.70** | **1.94** | 58.99 | 57.36 | 45.05 | 59.25 | 58.32 | 43.80 |

*Table 15. Comparison of binary and multi-class models in intra- and cross-dataset evaluations for RQ1. All models are trained on FF++.*

#### 6.2.4. Conclusions

Overall our findings support two conclusions. First, deepfake attribution models with vanilla approaches generalize less compared to binary counterparts, although carefully incorporating is helpful, especially in larger models. Second, the ability of attribution models to maintain their accuracy across datasets is heavily influenced by data quality as training on high-quality deepfakes drastically improves their performance.

#### 6.2.5. Relevance to AI4Media use cases and media industry applications

The developed method contributes to UC1 (AI for Social Media and Against Disinformation), specifically, Feature 1A (Detection/Verification of Synthetic Media). Deepfake detection systems could be used by social media platforms to assess the authenticity of uploaded content, and by journalists to verify real-world content. The proposed study contributes to better understanding the behaviour of deepfake detection and attribution models, ultimately leading to more robust detection systems.

### 6.3. Improving generalization in Deepfake detection with adversarial augmentation

**Contributing partner:** UNSTPB

| Train\Test | | DeepFakes | | | FaceShifter | | |
|---|---|---|---|---|---|---|---|
| | | FF | CDF | FN | FF | FN | DFP |
| EfficientNetV2 | FF | **96.99** | 10.20 | 11.28 | **97.04** | 14.72 | 4.39 |
| | CDF | 35.82 | **99.00** | 33.04 | - | - | - |
| | DFP | - | - | - | 33.26 | 50.40 | **99.88** |
| PVT-V2 | FF | **97.09** | 26.44 | 15.06 | **98.11** | 5.52 | 1.05 |
| | CDF | 18.44 | **98.15** | 20.69 | - | - | - |
| | DFP | - | - | - | 18.70 | 43.71 | **99.76** |
| ConvNextV2 | FF | **97.82** | 2.90 | 9.15 | **98.36** | 9.46 | 1.32 |
| | CDF | 20.95 | **99.78** | 20.76 | - | - | - |
| | DFP | - | - | - | 19.87 | 41.61 | **99.82** |
| SwinV2 | FF | **97.82** | 7.56 | 5.32 | **98.08** | 5.04 | 0.86 |
| | CDF | 25.51 | **98.95** | 14.60 | - | - | - |
| | DFP | - | - | - | 15.14 | 46.41 | **99.93** |
| | | FaceSwap | | | FSGAN | | |
| | | FF | FAV | | DFP | FAV | FN |
| EfficientNetV2 | FF | **95.47** | 0.00 | | - | - | - |
| | FAV | 0.86 | **95.64** | | 38.52 | **98.20** | 38.16 |
| | DFP | - | - | | **99.59** | 98.12 | 69.42 |
| PVT-V2 | FF | **97.32** | 0.00 | | - | - | - |
| | FAV | 0.54 | **93.67** | | 17.80 | **97.11** | 17.02 |
| | DFP | - | - | | **98.77** | 99.46 | 62.66 |
| ConvNextV2 | FF | **96.58** | 0.0 | | - | - | - |
| | FAV | 0.08 | **96.48** | | str 22.81 | **99.72** | 12.28 |
| | DFP | - | - | | 99.63 | **99.77** | 66.39 |
| SwinV2 | FF | **96.46** | 0.0 | | - | - | - |
| | FAV | 0.08 | **96.91** | | 12.91 | **99.79** | 10.93 |
| | DFP | - | - | | 99.90 | **100.00** | 69.24 |

*Table 16. Results for RQ2: Investigating the generalization performance of multiclass attribution models on common deepfake manipulations across datasets in terms of accuracy. Dashes indicate that the specific manipuataltion method in not a part of the training dataset, e.g. DeepFakes and FaceSwap are not part of the DFP dataset. Legend: FF: FaceForesnics++; CDF: Celeb-DF; FN: ForgeryNet; FAV: FakeAVCeleb; DFP: DF-Platter.*

| Setting\Dataset | FaceForensics++ | | | Celeb-DF | | | DFDC | | |
|---|---|---|---|---|---|---|---|---|---|
| | AUC (%) | BA (%) | EER ↓ (%) | AUC (%) | BA (%) | EER ↓ (%) | AUC (%) | BA (%) | EER ↓ (%) |
| | | | | EfficientNetV2 | | | | | |
| B | **99.20** | **95.66** | **3.64** | 60.56 | **57.95** | 42.08 | **65.47** | **61.26** | **38.52** |
| T-H | 98.08 | 92.43 | 6.86 | **61.55** | 57.40 | **41.07** | 64.29 | 59.16 | 39.91 |
| T-HS | 84.75 | 68.70 | 22.15 | 58.74 | 55.37 | 44.01 | 62.41 | 57.32 | 41.10 |
| NT | 98.73 | 95.57 | 4.41 | 60.04 | 54.84 | 42.25 | 64.02 | 58.89 | 40.29 |
| SC | 97.70 | 91.68 | 7.93 | 58.31 | 54.02 | 43.05 | 61.85 | 58.48 | 41.73 |
| | | | | PVT-V2 | | | | | |
| B | **99.47** | **96.69** | **2.93** | 59.53 | 54.59 | 43.27 | **63.02** | **59.33** | **40.36** |
| T-H | 50.80 | 50.00 | 50.00 | 51.20 | 50.00 | 50.00 | 51.20 | 50.00 | 50.00 |
| T-HS | 96.99 | 90.27 | 8.09 | 57.33 | 53.14 | 46.49 | 63.63 | 59.10 | 40.58 |
| NT | 99.23 | 96.16 | 3.51 | **59.56** | 54.49 | **42.28** | 62.47 | 58.19 | 40.77 |
| SC | 97.65 | 91.74 | 7.69 | 56.61 | **54.76** | 46.11 | 56.61 | 54.76 | 46.11 |
| | | | | ConvNextV2 | | | | | |
| B | **99.83** | **98.17** | 1.71 | **54.70** | **55.56** | 46.26 | 57.77 | 57.03 | 44.41 |
| T-H | 99.62 | 98.04 | **1.55** | 52.55 | 53.20 | 49.08 | 58.89 | 58.54 | 43.53 |
| T-HS | 99.82 | 97.91 | 1.91 | 50.55 | 53.28 | 48.37 | 60.14 | 58.74 | 42.70 |
| NT | 99.56 | 97.58 | 2.20 | 46.44 | 51.61 | 48.28 | 60.36 | 58.06 | 42.39 |
| SC | 99.44 | 95.91 | 3.30 | 50.36 | 50.94 | **49.82** | **61.21** | **58.93** | **41.72** |
| | | | | SwinV2 | | | | | |
| B | **99.75** | **97.70** | **1.94** | 58.99 | 57.36 | 45.05 | 59.25 | 58.32 | 43.80 |
| T-H | 99.29 | 96.59 | 3.40 | 57.60 | 54.96 | 44.79 | 60.96 | 56.67 | 42.21 |
| T-HS | 99.49 | 97.56 | 2.28 | **61.51** | **58.40** | **42.54** | **62.58** | **60.04** | **40.63** |
| NT | 99.63 | 97.45 | 1.96 | 59.67 | 57.16 | 43.64 | 60.18 | 58.32 | 42.50 |
| SC | 99.61 | 97.03 | 2.36 | 58.77 | 54.91 | 44.44 | 58.43 | 57.45 | 43.77 |

*Table 17. Comparison between vanilla and contrastive attribution models for RQ3. All models are trained on FF++. Legend: B: Baseline, vanilla multiclass training; T-H: Triplet with hard mining; T-HS: Triplet with hard positive and semihard negative mining; SC: Supervised Contrastive loss with 2 views and projection head; NT: NT-Xent loss with 2 views and projection head.*

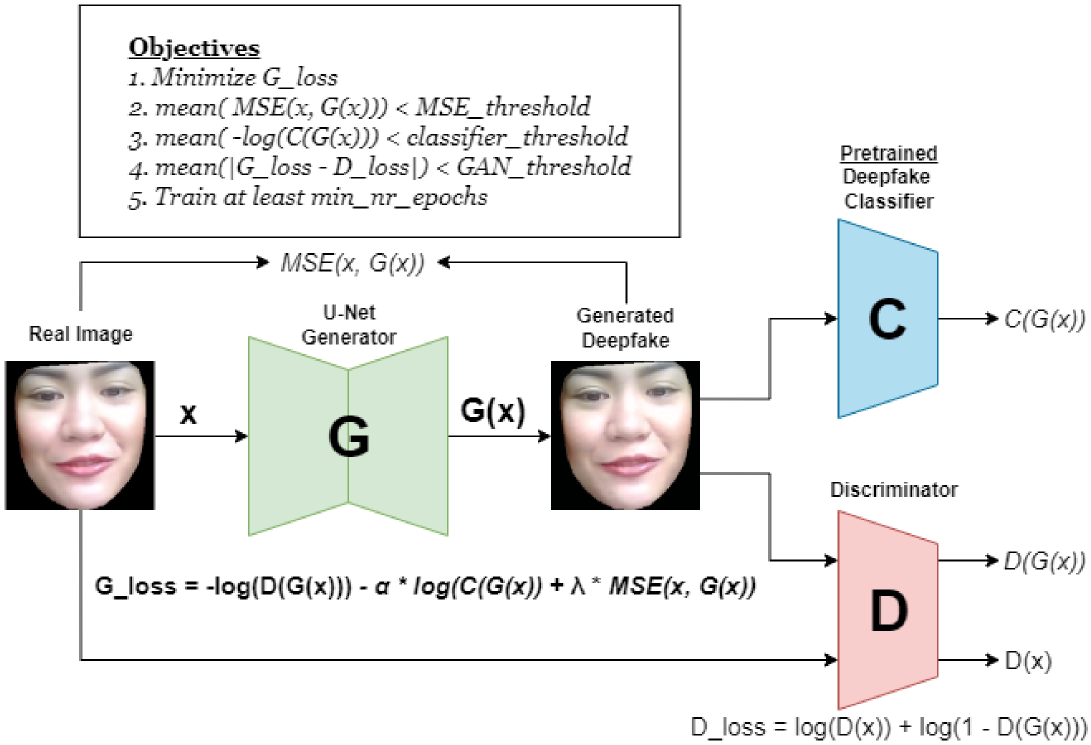



*Figure 9. The training and augmentation framework uses real images to generate deepfakes. A U-Net model is used to generate an augmentation deepfake, starting from a real image. To assess how good the deepfake is, it must meet 2 criteria: (1) the deepfake must be able to fool a pretrained deepfake classifier, (2) the deepfake must not be detected by a Discriminator. The generator and discriminator are trained similar to a GAN [2], until the generator can produce deepfakes that are realistic and are detected as 'real' by the deepfake classifier. Afterwards, augmentation data is generated and the classifier is generated using it. The process is repeated to improve generalization.*

### 6.3.1. Introduction

In the last years, machine learning and AI have gained significant attention due to the introduction of many innovative AI products. However, the authenticity of online information has become increasingly questionable, particularly with the rise of deepfakes — images or videos of people generated by deep neural networks. Most current detection methods for deepfakes suffer from a lack of generalization, as new models and algorithms for generation are constantly emerging. This work aims to improve generalization by drastically increasing the available training data through augmentations. In particular, we propose a novel augmentation algorithm that transforms real images and videos into deepfakes to create a virtually limitless amount of training data that trains robust classifiers in an adversarial manner.

### 6.3.2. Methodology

Our proposed method aims to improve the generalization capabilities of deepfake detectors by utilizing a dynamic data augmentation strategy based on adversarial attacks. The core of this approach lies in generating new deepfake samples during the training process, thereby continuously challenging the detector and enhancing its robustness. The method, presented in Figure 9, consists of several key components and steps:

- **The Generator.** This module is responsible for creating perturbations that are adversarial in nature. These perturbations are designed to fool the deepfake detector by subtly altering real video frames. To make sure it converges fast, it is a U-Net trained in an autoencoder configuration.

- **The Deepfake Detector.** A pretrained deepfake detector is trained to improve its generalization. For the training to be effective, it should be trained with images that cannot already be classified by it. Therefore, the Generator tries to create images that would fool it.

- **The Discriminator.** It is used to determine whether the generated images are realistic. In this case, if an image would be generated to fool the deepfake classifier, it is not necessarily realistic. In fact, it may present local artifacts that are clearly easy to spot with the naked eye. Therefore, a discriminator is needed to ensure that the newly generated images are realistic enough to be indistinguishable from the training real images.

- **Training Pipeline.** The training pipeline integrates the newly generated deepfakes into the training dataset. The deepfake detector is trained in a cyclical manner, alternating between standard training epochs and adversarial training epochs. During standard epochs, the detector learns from both real and previously generated deepfakes, while during adversarial epochs, it focuses on the new adversarially-generated samples. This cyclical training ensures that the detector is continually exposed to novel and challenging examples. To train the generator and discriminator, a GAN loss is used but with some additions: a term that determines how well the generator fools the deepfake classifier and a term that determines how far the generated image is from a real one.

- **Hyperparameter Tuning.** The training process involves careful tuning of several hyperparameters to maintain the balance between the generator and the detector. The learning rates for the generator and deepfake classifier are set to $10^{-3}$ and $10^{-5}$ respectively, while the discriminator is trained for every 5th iteration of the generator. The stopping criteria for training involve specific thresholds for generator and detector loss, ensuring that neither component dominates the training process.

- **Loss Functions.** The loss function for the generator combines multiple objectives, balancing the need to create convincing deepfakes and the need to mislead the detector. The loss function uses a weighted sum of adversarial loss and perceptual loss, with $\alpha = 300$ and $\lambda = 2$. This formulation encourages the generator to produce high-quality perturbations that are effective in fooling the detector.

- **Multi-Generator Approach.** To further enhance the diversity of the generated deepfakes, multiple generators can be employed. Each generator operates independently, creating unique perturbations. The outputs of these generators are combined, providing a richer and more varied set of training samples. This multi-generator approach ensures that the detector is not overfitting to a specific type of adversarial perturbation.

- **Objectives.** The training for the generator and discriminator lasts until certain objectives are met. For example, the generated image ought not to be too different from the input image (minimize MSE). Also, the deepfake classifier should be fooled to a certain level (classifier_threshold). Neither the generator nor the discriminator must dominate the loss function.

Once the generator is fully trained, the deepfake generator is trained with the initial training data and with the generated training data. Afterwards, the process repeats.

### 6.3.3. Experimental results

The experiments were conducted on several benchmark datasets, including Celeb-DF [131], FaceForensics++ [3], and DFDC [132]. The main evaluation metric was AUC, as this is a binary classification.

| Iterations | AUC ClebDF | AUC DFDC |
|---|---|---|
| 0 | 0.689 | 0.633 |
| 2 | 0.704 | 0.674 |
| 5 | 0.732 | 0.687 |
| 10 | 0.756 | 0.718 |
| 20 | 0.760 | 0.722 |
| 30 | 0.783 | 0.721 |

*Table 18. Number of iterations vs AUC on 2 datasets. 0 iterations symbolizes initial pretrained model*

Table 18 presents the performance increase while using this method. The model was trained on FaceForensics++ and evaluated an CelebDF and DFDC Preview. The number of iterations represents the number of times a new generator was trained, new deepfakes were generated and the deepfake generator was retrained.

The results show a clear improvement by using the presented method. The first few iterations improve performance the most, while adding more iterations only works up to a point. Overall, the performance was increased by almost 9% in both cases. More than that, by using different generator architectures, the boost in performance increased even more. By using 3 generator architectures instead of one, this approach achieved 81.32 % AUC for CelebDF and 73.8% AUC on DFDC preview.

### 6.3.4. Conclusion

We have presented a novel approach to improve deepfake detector generalization through continuous adversarial attack-based augmentation. The proposed method significantly outperforms baseline detectors in standard metrics and robustness tests, showcasing its potential for real-world applications. Future work could explore the integration of more sophisticated adversarial techniques and the application of this framework to other types of generative adversarial networks.

### 6.3.5. Relevant publications

- D. C. Stanciu, B. Ionescu, "Improving generalization in deepfake detection via augmentation with recurrent adversarial attacks", 3rd ACM International Workshop on Multimedia AI against Disinformation, 2024, https://dl.acm.org/doi/10.1145/3643491.3660291

### 6.3.6. Relevant software, datasets and other resources

- A draft for the proposed algorithm is published at https://github.com/StanciuC12/deepfake-augmentation-attack-gan.

## 6.4. Deepfake Detection by Exploiting Surface Anomalies

**Contributing partner:** UNIFI

### 6.4.1. Introduction

Nowadays, the ease of manipulating digital content and circulating it online has eroded public trust in the media. This has been accentuated by the recent advancements of AI for image and video processing, which have enabled a wide range of realistic manipulations. Deepfakes are one such class of manipulations which can tamper with parts of an image or a video scene, frequently, the human faces. They include reenacting or transferring facial expressions from a source to a target face, imitating the voice and inflection of a speaker, and lip syncing to reconstruct the mouth of a face according to the speech. From a technical perspective, the realism of these manipulations has been achieved with the emergence of deep generative models, such as GAN-style architectures [2, 13] and diffusion probabilistic models (DPMs) [133]. While these models are welcome in Hollywood for movie production, they can be easily exploited to create fake content for malicious purposes like propaganda, urgently raising the need for tools for deepfake detection.

This work focuses on detecting face manipulations in images. The fundamental idea behind deepfake detection is that a neural network that generates fake content always leaves traces that are embedded as a fingerprint over the manipulated image or video. Most existing approaches for detection try to uncover this hidden fingerprint by analyzing the pixels of images and video frames (RGB raw data). Beyond RGB-level inconsistencies however, the forgery can create more anomalies in the visual space without compromising the visual perception of the image. For example, the camera acquisition process is a signature that is incorporated into the image. This signature can be distorted by a manipulation by creating inconsistencies in various characteristics of the scene, for instance, face poses or the external illumination source, lighting, shadows and reflections.In addition, even the camera parameters like lens distortions and intrinsic noises can be embedded into the image and become a useful signal for detection.

### 6.4.2. Methodology

The proposed approach leverages on the analysis of the features of the framed scene that are determined by the overall geometry of the scene itself (e.g. surfaces) and by the original image acquisition process (e.g. illumination, camera orientation). Different from other research works focusing on individuating fakes by detecting specific patterns in terms of depth map or face motion, we focus on the surfaces present within the acquired scene by exploiting the modifications induced to the surface normals and determined by the deepfake alteration.

The proposed method, named *SurFake*, detects inconsistencies in the features of surfaces belonging to the acquired scene to perform deepfake detection. Our pipeline is organized into three steps as depicted in Figure 10: first, we perform face detection on each video frame using dlib [134] obtaining a face crop with a fixed resolution of $224 \times 224$. Secondly, we run a pretrained UpRightNet[135] on face crops to extract features (see Section 6.4.2.1 for details) in order to get the Global Surface Descriptor (GSD). Finally, we concatenate the RGB face crop and the GSD feature which constitutes a 6-channel tensor to train a deep convolutional neural network for binary detection of deepfakes.

**6.4.2.1. The Global Surface Descriptor (GSD)** In order to extract subtle scene details, we face deepfake detection from a new perspective. In contrast to looking for inconsistencies in the visual perception domain, we highlight anomalies by looking into the geometrical aspects related

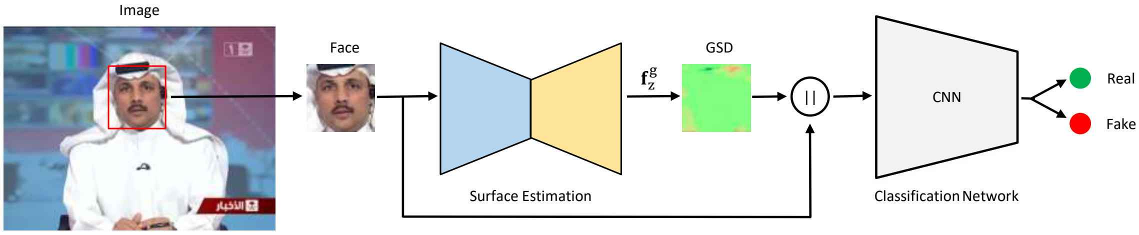



*Figure 10. Pipeline of SurFake for deepfake detection.*

to the camera acquisition process. Such aspects permanently mark low-level peculiarities of the image without visually affecting the content. Therefore, tampered images may contain fine-grained distortions which do not stand in the visible space.

Typically, deepfakes depict a person in the foreground whose entire face (or some of its parts) has been tampered with. The idea behind our approach is to address forgery detection by focusing on the surface geometry of the face. To do that, we employ a deep learning model named UpRightNet [135] to estimate such geometrical characteristics presented by the oval surface of the face but also by other different surfaces such as the chin, the nose, the eye sockets or any headgear. UpRightNet is a neural network that learns to estimate the 2DoF camera orientation, i.e. roll and pitch, from a single RGB image using intermediate representations, called surface frames, estimated from both the local camera and the global up-right coordinate systems. Let us suppose we want to predict the per-pixel surface normals of an indoor image in the camera perspective. Surface normals on the ground and other horizontal surfaces point in the same direction as the camera up vector, instead, walls and other vertical surfaces are perpendicular to the up vector. Thus, camera orientation can be estimated as finding the vector which is most parallel to the ground normals and most perpendicular to the wall normals. UpRightNet solved the camera orientation problem by computing the rotation that best aligns the two estimated representations of the surface frames.

A surface geometry frame $\mathbf{F}(i)$ is estimated from each pixel $i$, as a $3 \times 3$ matrix of mutually orthogonal unit vectors, that is normals, tangents and bitangents, respectively: $\mathbf{F}(i) = [\mathbf{n}(i), \mathbf{t}(i), \mathbf{b}(i)]$ with $\mathbf{n}(i), \mathbf{t}(i), \mathbf{b}(i) \in \mathbb{R}^3$. UpRightNet estimates two surface frames, one in the local camera coordinate system, $\mathbf{F}^c(i)$, and one in the global up-right coordinate system, $\mathbf{F}^g(i)$. In order to predict roll and pitch of the camera, UpRightNet aligns the up-vector in the two representations, by using the z-component of $\mathbf{F}^g(i)$, i.e. $\mathbf{f}_z^g(i) \in \mathbb{R}^3$. Such alignment is computed by learning weights to solve a constrained least squared problem using ground-truth camera orientations. Due to the fact that the feature $\mathbf{f}_z^g(i)$ substantially provides a 3-channel global description of the surfaces belonging to the acquisition scene, we have considered that it could be a good candidate to possibly give evidence of a manipulation. Such a feature, denominated Global Surface Descriptor (GSD), will be analysed more in depth within the next sub-section.

**6.4.2.2. Analysis of Deep Geometric Representations** This section presents how UpRightNet features can be useful in the context of deepfake detection; in Figure 11, we depict an example of a pristine face crop along with various face manipulation methods implemented in FaceForensics++[3] (first row). We also show in the second row the corresponding GSD feature extracted by UpRightNet after passing the face image as input. Finally, in the last row, we enhance the colorization of GSD in order to visually highlight how the GSD feature could be useful to detect anomalies by depicting the logarithm of the feature. In fact, upon a visual inspection, the GSD features may appear similar to each other for different faces, even appearing uniform regardless of the content. This is due to the fact that the faces are typically framed frontally in the global

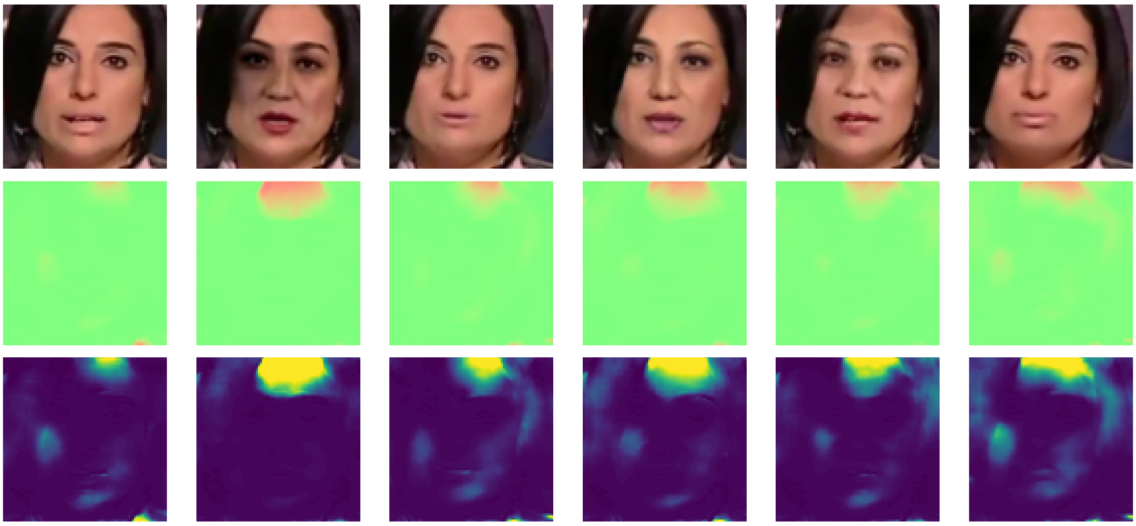


*Figure 11. Sample frames (first row) and the corresponding Global Surface Descriptors (second row) and* $\log(GSD)$ *(third row) for each of the 5 different forgeries in FF++, from left to right: Real, DF, F2F, FSH, FS, NT [3].*

upright coordinate system. Similarly to global representations estimated for indoor images, the light green pixels stand for surfaces whose normals are perpendicular to the up-ward vector, e.g. the walls in a room, just like most of the face pixels too. There are also other parts of the image which are sometimes colored with shades of red and dark green, that encode surfaces parallel to the ground of a room. In our face domain, pixels with normals parallel to the ground are located at the top and at the bottom of the image. Therefore, the geometry estimated for face images is quite consistent with indoor environments. However, we would like to demonstrate how this geometry, i.e. our proposed GSD feature, is useful in our task. To this aim, we here highlight anomalies, by calculating the logarithm of each image pixel on the first of the 3 channels and we plot the results in the third row in Figure 11, for each face manipulation.

#### 6.4.3. Experiments

In this section, we will present the experimental results carried out in order to verify the effectiveness of the presented approach and, in particular, of the GSD feature.

**6.4.3.1. Dataset** We conduct experiments on FaceForensics++ (FF++) [3], one of the most widely used datasets for deepfake detection. It has collected 1,000 original real videos from the internet and for each video 5 different forged versions are generated. This dataset comprises two types of face manipulation techniques: face swapping, in which the face identity in the source image is replaced with the target one, and face reenactment, in which the facial expression in the source image is altered from the one in a target image, while maintaining the identity. FaceForensics++ includes three swapping methods, DeepFakes (DF), FaceShifter (FSH) and FaceSwap (FS), and two reenactment methods, i.e. Face2Face (F2F) and NeuralTextures (NT).

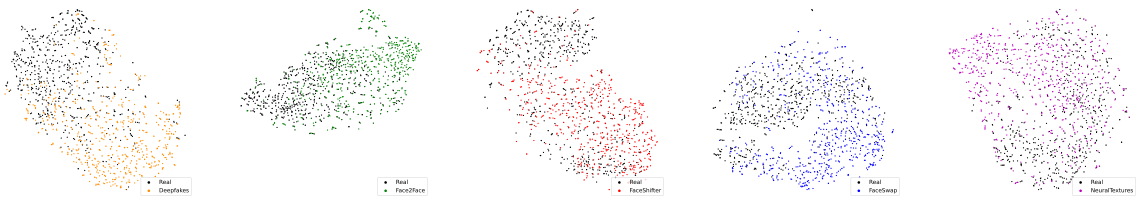


*Figure 12. T-SNE [4] plots of the GSD feature activations for real and fake samples of the test set for each of the different forgeries (MobilNetV2 architecture). Only a reduced number of samples is plotted for the sake of visibility.*

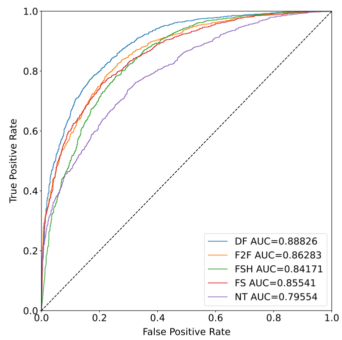



*Figure 13. ROC Curve of GSD features for real and fake using MobileNetV2 as classifier. We also reported the Area Under Curve (AUC) for each forgery.*

**6.4.3.2. Data preparation** To reduce the computational burden and exclude redundancies, we sample one frame out of ten in each video sequence. For all our experiments we follow the 72:14:14 data split, respectively for train, validation and test sets, as indicated in [3], i.e. 720, 140 and 140 videos. Following [3], 224 × 224 face crops are obtained by first extracting a 1.3-factor enlarged crop centered at the detected face in the input image and then scaling it to the fixed resolution. We consider face crops of such a resolution as it is a standard image dimension that can be processed into most of the existing architectures. Since UpRightNet gets input images of 288 × 384 and generates outputs at the same resolution, we adapt face crops to such dimension in order to extract the Global Surface Descriptor (GSD) and then we rescale to 224 × 224, as done in [135]. Among the UpRightNet pretrained weights on InteriorNet and ScanNet (i.e. two indoor image datasets), we chose the former one since this model estimates a more coherent representation of GSD than the latter, in accordance with the true geometry in the face domain in terms of normals perpendicular or parallel to the global upright coordinate system (see Section 6.4.2.1).

**6.4.3.3. Architectures** In order to classify images as real or fake, we train 4 different well-known and standard architectures: ResNet50, MobileNetV2, EfficientNet-B0 and Xception with pretrained weights on ImageNet. Since we are using a neural network pretrained on classical RGB images from ImageNet, we modify the first convolutional layer, which gets the input, to handle a different type of data, i.e. the 3-band GSD feature and also a different number of input channels, as our approach is employing a total of 6, i.e. the first 3 channels for the RGB concatenated to the second 3 additional ones from GSD. Besides, to make the training more stable and to allow the model to converge faster, we make a proper weight initialization. We calculate the average of the three original input channels from the pretrained model and we replace this initialization for

| Architectures | DF | F2F | FSH | FS | NT | Avg |
|---|---|---|---|---|---|---|
| | **FF++ forgeries** | | | | | |
| ResNet50 | 0.766 | 0.725 | 0.735 | 0.690 | 0.674 | 0.718 |
| MobileNetV2 | 0.800 | 0.773 | 0.756 | 0.764 | 0.713 | 0.761 |
| EfficientNet-B0 | 0.802 | 0.773 | 0.761 | 0.754 | 0.707 | 0.759 |
| Xception | 0.796 | 0.759 | 0.759 | 0.747 | 0.726 | 0.757 |

*Table 19. Performance in terms of accuracy for the GSD feature on the test set with respect to the different network architectures.*

| Architec. | DF | F2F | FSH | FS | NT | Average |
|---|---|---|---|---|---|---|
| | **FF++ forgeries** | | | | | |
| | RGB/RGB+G | RGB/RGB+G | RGB/RGB+G | RGB/RGB+G | RGB/RGB+G | RGB/RGB+G |
| ResNet50 | 0.981/**0.984** | 0.988/**0.989** | **0.980**/0.971 | 0.985/**0.986** | 0.938/**0.947** | 0.974/**0.975** |
| MobileNetV2 | 0.987/**0.992** | **0.990**/0.989 | 0.985/**0.989** | **0.990/0.990** | 0.958/**0.966** | 0.982/**0.985** |
| Effic.-B0 | 0.989/**0.992** | 0.983/**0.986** | **0.982/0.982** | 0.984/**0.985** | 0.955/**0.958** | 0.978/**0.981** |
| Xception | **0.976**/0.975 | **0.979/0.979** | **0.976/0.976** | 0.977/**0.978** | **0.939/0.939** | **0.969/0.969** |

*Table 20. Performance in terms of accuracy on the test set for the different architectures with respect to the FF++ forgeries for RGB and RGB+GSD cases.*

each and every channel of GSD. We chose this initialization for all our experiments.

**6.4.3.4. Training setting** We implement SurFake in Pytorch. We model the deepfake detection as a binary classification problem and we train each classification network on an NVIDIA TITAN RTX. Specifically, we use a standard cross entropy loss with two classes, real and fake, for 30 epochs and batch size 32. We utilize SGD as optimizer with momentum 0.9, weight decay 0.0001 and learning rate 0.001.

**6.4.3.5. Analysis of the proposed GSD feature performance** In this section, experimental results to evaluate the effectiveness of the proposed GSD feature will be presented. To better understand the capacity to provide distinctiveness between real and deepfake images, we have considered the activations obtained at the final layer of SurFake, just before getting the output decision step. Therefore, we train MobileNetV2 to detect each face manipulation by using only our GSD feature as input without RGB. MobileNetV2 contains the initial fully convolution layer with 32 filters, followed by 19 residual bottleneck layers and ends with a linear layer with 1,280 dimensional feature. We then plot these activations, by resorting to T-SNE [4]. As it can be seen in Figure 12 for all the 5 manipulation techniques of the FF++ dataset, it is possible to appreciate a certain separation between real samples (black dots) and fake ones (colored dots) which is coherent for all the different cases. Even though our proposed GSD features between real and fake are often uniform (see Figure 11), subtle differences can be perceived using a neural network, while, instead, being almost invisible to the human eye. The depicted T-SNE in Figure 12 clearly demonstrates how relevant these GSD patterns are (similar representations can be obtained with other network architectures). Although we are processing patches that describe surfaces of faces rather than canonical RGB face images, a significant amount of test samples have been correctly separated in the projection space. We also plot the ROC curves in the test set for all the face manipulations detected using MobileNetV2, in Figure 13. We deduce that in all the forgery techniques the AUC (Area Under Curve) is around 0.85, which is quite interesting considering the scarce visible information carried on this feature. Similarly, we have tried to make a quantitative evaluation of such a phenomenon and we have computed the accuracy values for all the five different distortions. To do that, we evaluate our approach, still using GSD features as unique input to the classification network, and we report our results for all the 4 architectures. As listed in Table 19, we can notice that in each case and, above all, coherently for different kinds of network architectures, an average accuracy around 0.75 can be globally achieved. We observe that GSD exhibits well distinctiveness in most of the manipulations for all the architectures with high accuracy, e.g. DeepFakes (DF), Face2Face (F2F) and FaceShifter (FSH) are well-detected. Performance on NeuralTextures (NT) is lower than the other manipulations and possibly this is due to the fact that NT deals with facial reenactment performed just around the mouth region [3].

**6.4.3.6. Composing GSD with RGB frames** Hereafter, we will present the results obtained by composing the 3-channel GSD with the RGB frames that are usually adopted as primary source of information in most deepfake detection methods. This has been done in order to understand if the proposed GSD feature is able to provide an improvement in deepfake detection, thanks to the fact that it takes into account geometrical components related to the acquisition scene. In this case the diverse network architectures have been trained by receiving as input a 6-channel tensor composed of 3 RGB bands concatenated with the 3 GSD channels.

The achieved performances in terms of detection accuracy are listed in Table 20. As it can be seen, by looking at the last column of the table, a general increment is registered on average.

Overall, we notice that our approach employing either EfficientNet-B0 or MobileNetV2 gets more performance improvement. These two architectures show a superior trend of the ROC curves with respect to others, and with the corresponding Area Under Curve values highest among the manipulations. Such behavior demonstrates that our proposed GSD feature introduced in a classification network can benefit the detection of deepfakes.

### 6.4.4. Conclusion

In summary, the main contributions of this work are listed hereafter:

- to utilize surface geometry features of the acquired scene to highlight inconsistent patterns revealing fake images;
- to study and evaluate, in which extent, such features can constitute by themselves an effective mean to discriminate between pristine and fake contents;
- to conduct experiments on different kinds of forgeries and network architectures to verify that the new proposed surface-based feature can be advantageously combined with RGB frames to get an improvement in terms of accuracy performance.

### 6.4.5. Relevant publications

- A. Ciamarra, R. Caldelli, F. Becattini, L. Seidenari and A. Del Bimbo, "Deepfake Detection by Exploiting Surface Anomalies: The Surfake Approach," 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 2024, pp. 1024-1033, doi: 10.1109/WACVW60836.2024.00112 [136]

### 6.4.6. Relevance to AI4Media use cases and media industry applications

The present 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 platforms 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.

## 6.5. Synthetic Audio Detection with Frequency MLP-Mixer

**Contributing partner:** CERTH

### 6.5.1. Introduction

The integrity of speaker verification systems is increasingly challenged by the rapid advancement of synthetic speech technology, necessitating the development of robust countermeasures. Our study contributes to this critical field of research by exploring novel architectures for synthetic speech detection derived from the MLP Mixer model [137].

In more detail, we first implement the MLP Mixer as a baseline model for synthetic speech detection, and then introduce a novel variant, enhanced with frequency-domain processing techniques, the Frequency Mixer. The latter is inspired by recent advancements in adaptive frequency filtering [138] which aim to effectively capture spatial and spectral relationships in audio signals. For evaluation, both models are evaluated on the Logical Access (LA) partition of the ASVspoof 2019 dataset [139] using the Equal Error Rate (EER) and F1-score metrics.

By leveraging the computational efficiency of the MLP-based model and enhancing it with rich spectral information captured by frequency-domain processing, we aspire to contribute to more effective countermeasures against sophisticated audio spoofing attacks. This work not only addresses the need for advanced synthetic speech detection methods but also explores the adaptability of emerging machine learning architectures to novel domains. In particular, the adaptability of vision-oriented architectures to audio processing tasks could open new avenues for cross-modal learning in synthetic speech detection.

### 6.5.2. Methodology

**MLP-Mixer:** The MLP-Mixer, introduced by Tolstikhin et al [137], represents a departure from convolutional and attention-based architectures in computer vision. This model consists entirely of multi-layer perceptrons (MLPs) applied independently to image patches and feature channels. The architecture processes an input image by splitting it into patches, linearly embedding these patches and then passing the resulting tokens through alternating token-mixing and channel-mixing MLP layers.

The token-mixing MLPs operate on each channel independently across spatial locations, while channel-mixing MLPs act on each spatial location independently across channels. This approach allows the model to capture both spatial and channel dependencies without relying on convolutions or attention mechanisms.

A key advantage of the MLP-Mixer is its computational efficiency. Its complexity scales linearly with the number of patches, contrasting favorably with the quadratic scaling of self-attention in Vision Transformers [140].

**Frequency Token Mixer:** The concept of a Frequency Token Mixer builds upon the MLP-Mixer framework by incorporating frequency domain processing. This approach is inspired by recent work in adaptive frequency filtering for vision transformers [138]. In the Frequency Token Mixer, input data is transformed input data into the frequency domain using the Fast Fourier Transform (FFT). Learned transformations are then applied in the frequency domain before converting the data back to the spatial domain via Inverse Fast Fourier Transform (IFFT).

This method leverages global information within the input data, potentially enhancing the model's ability to extract features across image patches. By operating in the frequency domain, the Frequency Token Mixer can capture a broader spectrum of frequency components, potentially improving robustness against noise and variations.

**Frequency MLP-Mixer:** The Frequency MLP Mixer represents a novel fusion of the MLP-Mixer architecture with frequency domain processing. This approach introduces a FrequencyMixer module that replaces the standard mixer block in the original MLP Mixer design. The FrequencyMixer module operates as follows:

1. Patch-embedded tokens are transformed into the frequency domain using FFT.
2. Learnable weights are applied to mix these frequency-domain representations.
3. The mixed tokens are returned to the spatial domain using IFFT.

This process is similar to the Adaptive Frequency Filter, but adapted to the MLP-Mixer architecture. The FrequencyMixer is integrated using a residual connection, allowing the model to preserve both spatial and frequency-domain information.

### 6.5.3. Experiments

**Feature Extraction:** For audio signal processing, we utilized Mel spectrograms, which are time-frequency representations derived from the Short-Time Fourier Transform (STFT). The Mel scale, approximates the human auditory system's perception of pitch. The feature extraction process generated a $256 \times 256$ matrix for each audio sample, representing 256 time frames and 256 Mel frequency bins. This approach is similar to that used in recent audio classification tasks [141].

**Metrics:** The models' performance was assessed using two principal metrics:

1. Equal Error Rate (EER): This metric determines the threshold where the false acceptance rate and the false rejection rate are equivalent, effectively measuring the trade-off between these two error types [142].
2. F1-score: Representing the harmonic mean of precision and recall, this score offers a balanced assessment of the model's effectiveness, especially valuable for datasets with class imbalances [143].

**Results:** The results from the experiments, as shown in Table 21, demonstrate the effectiveness of the proposed models in detecting synthetic speech. The baseline MLP Mixer model achieved an Equal Error Rate (EER) of 13.60% and an F1-score of 88.73%, which serves as a competent starting point but lags behind more specialized synthetic speech detection models.

In contrast, the novel Frequency Mixer model significantly outperforms the baseline, achieving an EER of 7.75% and an F1-score of 92.82%. This improvement underscores the benefits of integrating frequency-domain processing into the MLP Mixer architecture, which enhances its ability to analyze the spectral characteristics essential for distinguishing between genuine and synthetic speech.

Comparatively, while state-of-the-art systems like AASIST and RawNet2 report lower EERs, at 1.13% and 1.12% respectively, the Frequency Mixer shows a competitive edge over the CQCC-ResNet model, which reports an EER of 7.69%. These results highlight the potential of our approach in bridging the gap with specialized architectures, suggesting that further enhancements in frequency-domain processing could yield even more robust detection capabilities.

| Model | EER (%) | F1-Score (%) | Source |
|---|---|---|---|
| MLP-Mixer | 13.60 | 88.73 | Our experiment |
| Frequency MLP-Mixer | 7.75 | 92.82 | Our experiment |
| AASIST | 1.13 | - | [144] |
| RawNet2 | 1.12 | - | [145] |
| CQCC-Resnet | 7.69 | - | [146] |

*Table 21. Comparison of models' performance on EER and F1-Score metrics.*

#### 6.5.4. Conclusions

In this study, we have explored the application of MLP-Mixer architecture and its frequency-domain variant to synthetic speech detection. The experiments, conducted on the ASVspoof 2019 LA dataset, have demonstrated the potential of these architectures in audio security applications. The Frequency Mixer model has shown particularly promising results, achieving an F1-score of 0.9198 and an EER of 7.75%, outperforming the baseline MLP Mixer. These findings suggest that integrating frequency-domain processing into the MLP-Mixer architecture can enhance its synthetic speech detection capabilities. The improved performance can be attributed to the models' ability to capture both spatial and frequency-domain relationships in the input data, similar to the benefits observed in vision tasks [138]. Future research directions could include further refinements to this architecture, exploration of its performance on additional audio security tasks, and more comprehensive comparisons with state-of-the-art models. Additionally, investigation into the interpretability of these models and their robustness against adversarial attacks could provide valuable insights for the field of audio security.

#### 6.5.5. Relevant publications

- No relevant publications published yet.

#### 6.5.6. Relevant software, datasets and other resources

- No additional resources published yet.

#### 6.5.7. Relevance to AI4Media use cases and media industry applications

This study centers on creating a sophisticated deep-learning framework designed for identifying synthetic audio, tailored for integration with UC1's current synthetic audio detection application. This enhanced tool is intended for a diverse group of media professionals in journalism, film, and gaming. It seeks to go beyond simple binary classification, providing a comprehensive solution that categorizes audio samples as 'real' or 'synthetic'. Additionally, it aims to improve the decision-making process by incorporating transparent and explainable AI techniques.

### 6.6. Compression and Transfer Learning of DeepFake Detection Models

**Contributing partner:** CERTH

### 6.6.1. Introduction

Model compression aims at reducing the size and computational complexity of models, thereby enhancing their efficiency in terms of storage, memory usage, and computational resources. An important application of this is to enable the deployment of capable models on resource-constrained devices. This is critical for achieving fast response times and to conserve battery. The four major approaches to model compression are: Pruning, Knowledge Distillation, Quantization and Low Rank Factorization.

In addition, transfer learning can significantly enhance the efficiency of model training by leveraging pretrained models on related tasks, thereby reducing the computational resources and time required for training. Instead of training a model from scratch, transfer learning allows the use of a pretrained model's features, which have already captured relevant patterns and representations.

In the following, we evaluate the impact of the above techniques on the task of deepfake detection. In particular, Section 6.6.2 describes the compression techniques, Section 6.6.3 the combination with transfer learning, and Section 6.6.4 the experimental setup and the results.

### 6.6.2. Compression of Deepfake Models

We consider an initial CNN model that has been trained on appropriate datasets for the deepfake detection task in order to achieve baseline performance. The CNN architecture is selected because it is a standard and efficient choice for computer vision tasks. In the following, we describe our methodology for each compression method.

**Pruning:** We apply L1-norm unstructured local pruning to the CNN, wherein parameters are pruned at specified percentages based on their L1-norm values. This selective pruning method reduces model complexity while retaining critical features. After pruning, a fine-tuning process is conducted to make a performance recovery following the parameter reduction. With the post fine-tuning, a pruning mask is used to indicate which parameters in a neural network should be pruned and which should remain active, resulting in a compact model.

**Knowledge Distillation:**

We consider the trained CNN as a teacher model and created smaller, computationally efficient student models to replicate its performance using response-based offline distillation. In this method, the teacher's outputs guide the student models' training. The distillation loss function combines cross-entropy loss from the labels and Kullback-Leibler (KL) divergence loss from the teacher's soft targets, which are softened using a temperature parameter set to 2. The total loss is scaled by the square ($T^2$) of this temperature as suggested by the authors of [147]. The weights assigned to the cross-entropy and KL divergence losses determine the emphasis on each component.

**Quantization:** We quantized the linear layers of a trained CNN by reducing their precision from float32 to int8, effectively decreasing model size, storage requirements, and inference time to enhance efficiency. However, practical challenges arose, particularly with GPU operability and the inability to quantize convolutional layers due to limited support in frameworks like PyTorch. This limitation prevented full model compression, as only the linear layers benefited from quantization [148]. Despite these challenges, the method remains valuable for exploration, though the overall compression was less than expected.

**Low Rank Factorization:**

Low-rank factorization represents an original weight matrix $W$ as the product of two smaller matrices, $U$ and $V$, with $V$ being $m \times k$ and $V$ being $k \times n$ with $k$ much smaller than $m$ and $n$. We apply this method to the first linear layer of a trained CNN, using Singular Value Decomposition (SVD) to decompose it into two smaller layers, thereby reducing the parameter count. This new model architecture replaces the original large linear layer with these smaller layers, transferring the weights accordingly. While this approach decreases parameter count, its effectiveness is limited by the fact that most CNN layers are convolutional and not suited for factorization, making overall compression less significant. However, it may still be beneficial in cases where linear layers are dominant and the performance impact is manageable.

### 6.6.3. Transfer Learning in Deepfake Models

Transfer learning is a technique in neural network training that utilizes pre-trained models on large datasets to enhance performance on new, smaller tasks. By leveraging the knowledge gained from the initial broad training, transfer learning accelerates learning, reduces computational resources, and improves the effectiveness of models for specialized applications. This approach often involves fine-tuning the pre-trained model on a new dataset, which is especially advantageous for deep learning tasks like DeepFake detection, where accuracy and efficiency are crucial.

**Pruning with Fine Tuning:** Our approach to adapting models for different datasets involves using pruned DeepFake models for transfer learning to enhance their efficiency on new datasets. By retraining only specific layers while freezing others, we improve detection performance and significantly reduce training time and computational resources. It's crucial to retain pruning masks during this process to preserve the pruned status of the layers, which helps maintain a compact model. After transfer learning, removing the pruning mask results in an efficient model with improved performance on the new dataset, showcasing the effectiveness of combining pruning with transfer learning for optimization.

**Pruning with Fine Tuning and Adapters:** Furthermore we use a transfer learning approach that enhances pruned CNN models' performance by incorporating adapters, inspired by their successful use in transformers. Adapters are added as lightweight layers within the DeepFake model architecture, strategically placed to optimize feature detection and boost performance without significantly increasing model size or training resources. The adapters consist of depthwise separable convolutions, batch normalization, and ReLU activation, and are trained exclusively while the original pruned layers remain frozen. This method effectively improves model performance with minimal complexity, aligning with recent advancements in parameter-efficient transfer learning [149].

**Knowledge Distillation using New Data:**

In this approach, we utilize a pre-trained large detection model, trained on dataset A, as a teacher. This teacher progressively distills its knowledge in a smaller student model on a different dataset, dataset B. Since the teacher model may perform suboptimally on dataset B, we adjust the loss weighting during distillation to emphasize actual labels more heavily. This adaptation helps create compact models that maintain high performance on the new dataset, effectively balancing model size and accuracy through knowledge distillation tailored for transfer learning.

**Knowledge Distillation with Adapter:**

Building on the previous use of knowledge distillation for transfer learning, this approach incorporates adapters into the distilled detection models to further boost performance. These adapters, similar in architecture to those used in pruned models, are strategically placed at different points to optimize feature detection in smaller models. The process involves creating new model instances with adapters, transferring original weights, and freezing the original layers to focus training on the adapters. This results in slightly larger yet still compact models that exhibit enhanced performance in DeepFake detection without altering the pre-existing knowledge.

**Quantization with Fine Tuning:** In this approach, we combined knowledge transfer with quantization to enhance both the performance and efficiency of a detection model. By selectively training only the last convolutional and first linear layers on a new dataset, we improved the model's performance on this data. Subsequently, we applied quantization to compress the model, reducing the precision of the linear layers from float32 to int8, resulting in a smaller and more efficient model.

**Low Rank Factorization with Fine Tuning.**

The final approach combined knowledge transfer with low-rank factorization to improve detection performance while reducing model size. By selectively training only the last convolutional and first linear layers on a new dataset, we enhanced performance. Then, low-rank factorization was applied to compress the first linear layer, reducing the model's parameter count and improving efficiency. However, this compression led to a significant performance drop, indicating that the effectiveness of this method may be context-dependent.

### 6.6.4. Experiments

This section provides an overview of the experimental design, detailing the datasets employed, the evaluation metrics utilized, and the obtained results.

**Datasets.** Three different datasets were used. The first one is the '140k Real and Fake Faces'[6]. This comprises a total of 140,000 images. Of these, 100,000 images are allocated for training purposes, 20,000 images for validation, and the remaining 20,000 for testing. Each image in the dataset has a resolution of 256x256 pixels. The second one is the 'deepfake and real images', which is also pertaining to real and DeepFake human photos [7]. This consists of 190,335 images in total, with 140,002 images designated for training, 39,428 for validation, and 10,905 for testing. Each image has a resolution of 256×256 pixels. The third dataset used for the transfer learning experimental evaluation is called 'dogs vs. cats'. This consists of images of cats and dogs for the task of binary classification [8]. This contains a total of 25,000 images of varying resolution, 20,000 of which are used for training and 5,000 for testing. This dataset does not include a validation set.

**Implementation.** The main Python libraries used were PyTorch for neural network modeling, PIL (Python Imaging Library) for image processing, the time module for tracking experiments, psutil for monitoring system resources, and scikit-learn for additional machine learning tasks. Experiments were conducted on Kaggle, utilizing its free access to GPUs and TPUs. The hardware setup included two NVIDIA T4 GPUs with the NVIDIA Turing architecture, each featuring 2560 CUDA cores, 16 GB of GDDR6 memory, with a maximum memory clock speed of 5001 MHz and delivering peak FP32 performance of 8.1 TFLOPS and INT8 performance of 130 TOPS.

**Outcomes.** Based on the methods evaluated in our study, we present some of our experimental outcomes in Tables 22 and 23, we see that knowledge distillation and pruning emerged as the most effective techniques in terms of both compression and performance for our tasks. In the context of pruning, we observed that increasing the number of pruned parameters proportionally heightens the challenge for the model to maintain its performance. We observed with more complex DeepFake datasets, higher pruning percentages would likely result in more pronounced performance degradation.

Regarding knowledge distillation, the smaller models consistently demonstrated strong performance on our datasets, similar to the results observed with pruning. Importantly, the efficacy of knowledge distillation is heavily influenced by the teacher model's performance. In instances where the teacher model's performance is suboptimal, it is more advantageous to emphasize the loss from the ground truth over the loss from the teacher for better results during the distillation process. Something worth mentioning is that the distinct architecture of the smaller student models will likely make them less prone to performance degradation on more challenging datasets, compared to larger pruned models based on our empirical results.

[6] https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces
[7] https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images
[8] https://www.kaggle.com/datasets/salader/dogs-vs-cats

| Method | Full Model | 40% parameters | 30% parameters | 20% parameters | 10% parameters |
|---|---|---|---|---|---|
| **Pruning** | Precision: 0.9878<br>Recall: 0.9914<br>F-1 score: 0.9896<br>Accuracy: 0.9896 | Precision: 0.9671<br>Recall: 0.9507<br>F-1 score: 0.9588<br>Accuracy: 0.9592 | Precision: 0.9296<br>Recall: 0.9141<br>F-1 score: 0.9217<br>Accuracy: 0.9224 | Precision: 0.8577<br>Recall: 0.8371<br>F-1 score: 0.8473<br>Accuracy: 0.8491 | Precision: 0.9418<br>Recall: 0.1134<br>F-1 score: 0.2024<br>Accuracy: 0.5532 |
| **Knowledge Distillation** | Precision: 0.9878<br>Recall: 0.9914<br>F-1 score: 0.9896<br>Accuracy: 0.9896 | Precision: 0.9899<br>Recall: 0.9721<br>F-1 score: 0.9809<br>Accuracy: 0.9811<br>Train Time: 1862 seconds | Precision: 0.9845<br>Recall: 0.981<br>F-1 score: 0.9827<br>Accuracy = 0.9828<br>Train Time: 1825 seconds | Precision: 0.9928<br>Recall: 0.974<br>F-1 score: 0.9833<br>Accuracy: 0.9835<br>Train Time: 1913 seconds | Precision: 0.9856<br>Recall: 0.969<br>F-1 score: 0.9772<br>Accuracy: 0.9774<br>Train Time: 3195 seconds |
| | **Original Model** | | **Compressed Model** | | |
| **Quantization** | Size: 17.6 Mb<br>Precision: 0.9878<br>Recall: 0.9914<br>F-1 score: 0.9896<br>Accuracy: 0.9896<br>CPU inference on 20000 images: 5378 seconds | | Size: 12.8 Mb<br>Precision: 0.9875<br>Recall: 0.9914<br>F-1 score: 0.9894<br>Accuracy: 0.9894<br>CPU inference on 20000 images: 3918 seconds | | |

*Table 22. Results on "Real versus Fake" Dataset*

The quantization method maintained performance effectively; however, the overall compression achieved is not on a par with knowledge distillation and pruning, making direct comparisons less meaningful. Additionally, the inability to execute quantized detection models on processing accelerators such as GPUs restricts the applicability of this method in general scenarios. Yet, in our target scenarios of deployment on edge devices, where no GPU is available, quantization is still valuable.

Low-rank factorization reduced the size of the original models to a certain extent, but it frequently resulted in significant performance degradation. The substantial performance degradation limits the utility of this method, making it highly context-dependent. Furthermore, the limited compression achieved necessitates its use in conjunction with other compression methods to attain substantial compression that is commonly needed in DeepFake edge computing scenarios.

In terms of transfer learning, our findings indicate that the position of the adapter significantly influences the performance of the method. For deeper models with more layers, the adapter typically needs to be placed closer to the middle layers to maximize its effect and consequently enhance model performance. This suggests that recognizing moderately complex features is more effective than very simple or highly complex ones.

Therefore, in deeper pruned detection models, we positioned the adapter after the fourth convolutional layer, which is approximately in the middle of the model. Conversely, in smaller models, such as most of the student models used in the knowledge distillation method, placing the adapter at the end of the convolutional layers, the deepest position possible, yielded the greatest benefit. Finally, our experiments revealed that adapters are only effective for feature extraction. Adapters consisting of linear layers did not demonstrate any noticeable performance gains in any of our experiments.

### 6.6.5. Conclusion

Our study has evaluated several compression and transfer learning methods for deepfake detection models, highlighting pruning and knowledge distillation as valuable techniques for enhancing both performance and efficiency. Through experimentation across various datasets, we confirmed that these methods not only effectively reduce model complexity and resource utilization but also maintain detection accuracy. This advancement is crucial for deploying effective deepfake detection systems in resource-constrained environments.

### 6.6.6. Relevant publications

An article, titled "Compression and Transfer Learning of DeepFake Detection Models" by A. Karathanasis, J. Violos, S. Papadopoulos and I. Kompatsiaris, is currently under preparation.

### 6.6.7. Relevant software/datasets/other outcomes

The code will be made publicly available on GitHub once the paper is accepted: `https://github.com/andreaskarathanasis/Compression-Transfer-of-DeepFake-Models`

### 6.6.8. Relevance to AI4Media use cases and media industry applications

The present solution contributes to UC1 (AI for Social Media and Against Disinformation). Journalists can use it to verify content they encounter on the Web. The work is particularly useful to make the process more efficient and fit the detection tools in compute-constrained environments.

| Method | 40% parameters | 30% parameters | 20% parameters | 10% parameters |
|---|---|---|---|---|
| **Pruning + transfer** | Precision: 0.8535<br>Recall: 0.8860<br>F-1 score: 0.8694<br>Accuracy: 0.8660<br>Train Time: 1881 seconds | Precision: 0.8477<br>Recall: 0.8800<br>F-1 score: 0.8635<br>Accuracy: 0.8599<br>Train Time: 1789 seconds | Precision: 0.8303<br>Recall: 0.8687<br>F-1 score: 0.8490<br>Accuracy: 0.8444<br>Train Time: 1790 seconds | Precision: 0.8116<br>Recall: 0.8443<br>F-1 score: 0.8276<br>Accuracy: 0.8229<br>Train Time: 1795 seconds |
| **Knowledge Distillation** | Precision: 0.9467<br>Recall: 0.9198<br>F-1 score: 0.9331<br>Accuracy: 0.9336<br>Train Time: 2715 seconds | Precision: 0.9537<br>Recall: 0.8867<br>F-1 score: 0.9190<br>Accuracy: 0.9213<br>Train Time: 2675 seconds | Precision: 0.9392<br>Recall: 0.9146<br>F-1 score: 0.9267<br>Accuracy: 0.9271<br>Train Time: 2583 seconds | Precision: 0.9534<br>Recall: 0.7605<br>F-1 score: 0.8461<br>Accuracy: 0.8607<br>Train Time: 2559 seconds |
| **KD + adapter after last Conv2d layer** | Precision: 0.9179<br>Recall: 0.9431<br>F-1 score: 0.9303<br>Accuracy: 0.9289<br>Train time: 1332 seconds | Precision: 0.9329<br>Recall: 0.9304<br>F-1 score: 0.9317<br>Accuracy: 0.9313<br>Train time: 1187 seconds | Precision: 0.9105<br>Recall: 0.9439<br>F-1 recall: 0.9269<br>Accuracy: 0.9250<br>Train time: 1183 seconds | Precision: 0.8845<br>Recall: 0.8872<br>F-1 score: 0.8859<br>Accuracy: 0.8849<br>Train time: 1188 seconds |
| **Pruning + adapter after last Conv2d layer** | Precision: 0.8633<br>Recall: 0.8801<br>F-1 score: 0.8716<br>Accuracy: 0.8695<br>Train time: 1865 second | Precision: 0.8582<br>Recall: 0.8831<br>F-1 score: 0.8705<br>Accuracy: 0.8676<br>Train time: 1918 seconds | Precision: 0.8540<br>Recall: 0.8461<br>F-1 score: 0.8500<br>Accuracy: 0.8497<br>Train time: 1912 seconds | Precision: 0.8424<br>Recall: 0.8102<br>F-1 score: 0.8260<br>Accuracy: 0.8281<br>Train time: 1890 seconds |
| **Pruning + adapter after fourth Conv2d layer** | Precision: 0.8748<br>Recall: 0.8845<br>F-1 score: 0.8796<br>Accuracy: 0.8781<br>Train time: 2415 seconds | Precision: 0.8720<br>Recall: 0.9011<br>F-1 score: 0.8863<br>Accuracy: 0.8836<br>Train time: 2462 seconds | Precision: 0.8861<br>Recall: 0.8617<br>F-1 score: 0.8738<br>Accuracy: 0.8746<br>Train time: 2458 seconds | Precision: 8745<br>Recall: 0.8343<br>F-1 score: 0.8539<br>Accuracy: 0.8563<br>Train time: 2415 seconds |
| | **Transfered Model** | | **Compressed Model** | |
| **Quantization** | Size: 17.6 Mb<br>Precision: 0.8526<br>Recall: 0.8996<br>F-1 score: 0.8755<br>Accuracy: 0.8711<br>Transfer time: 1959 seconds<br>CPU inference on 10905 images: 2949 seconds | | Size: 12.8 Mb<br>Precision: 0.8523<br>Recall: 0.8998<br>F-1 score: 0.8754<br>Accuracy: 0.8710<br>CPU inference on 10905 images: 3036 seconds | |

*Table 23. Results on "Real versus Fake" transfer to "deepfake and real images"*

# 7. Hybrid, privacy-enhanced recommendation (T6.3)

## 7.1. LLM news articles annotation

**Contributing partner:** FhG

### 7.1.1. Introduction

"Large Language Models (LLMs) like GPT-3 have revolutionized the fields of text analysis and text generation, significantly enhancing the capabilities of natural language processing (NLP). Prior to LLMs, text analysis and generation relied heavily on rule-based systems and traditional machine learning algorithms, which often required extensive feature engineering and domain-specific knowledge. LLMs, however, leverage deep learning techniques and massive datasets to understand and generate human-like text, making them far more versatile and powerful.

LLMs can perform a variety of tasks such as sentiment analysis, summarization, translation, and question-answering with unprecedented accuracy. Their ability to understand context and generate coherent and contextually relevant text has set a new benchmark for NLP. This advancement has significant implications for the media industry. Content creation can now be partially automated, reducing the time and cost associated with producing articles, reports, and other media. LLMs can assist journalists by providing drafts, suggesting headlines, and even conducting preliminary research, thereby streamlining the editorial process.

Moreover, LLMs enable more personalized content delivery. By analyzing user preferences and behavior, media companies can generate tailored content that engages readers more effectively. This personalization extends to marketing and advertisements, where LLMs can craft messages that resonate more with target audiences, ultimately driving higher engagement and conversion rates.

In summary, the advent of LLMs has not only elevated the state of the art in text analysis and generation but also promises to transform the media industry by enhancing content creation, personalization, and audience engagement."

At least, this is what ChatGPT is telling about itself[9]. For AI4Media, we decided to leverage this technology, for analyzing the capabilities of LLMs for news-recommendation, as a joint work between this task and Task 4.5 "Methods for detection and mitigation of bias affecting fairness in recommender systems" (see D4.7).

### 7.1.2. Methodology

For recommending items with the goal of minimizing bias effects, it is important to know the framing of a news article or to what extent a given news article is an opinion piece. With that information, it is possible to recommend items conforming or going against the bias of the user asking for the recommendation, or in other words: it is possible to select whether the recommended items are from within or from outside the filter bubble of the user.

At the start of AI4Media, building such a system would have been desirable, but very hard to build using the state of the art of text analysis of that time. To cater those relatively new developments, we decided to close this task with building a demonstrator that showcases how state of the art LLMs can be used to analyze news from various sources and languages – and how to use the metadata gained there to train a more traditional (and explainable) graph based news recommender, following up on the work done in this task, previously.

[9] Prompt to a Chat GPT 4o model: "Please write a 0.5 page introduction of how LLMs changed the state of the art of text analysis and text generation; and how this will have big influences on the media industry."

The idea is to use the text analysis capabilities of the LLMs to analyze and also compare articles and store this information as a graph, e.g. if multiple articles are written from a certain political standpoint, they will be connected indirectly over a node, representing that stance. Then we will use a graph based recommender system implementation based on "Pixie" [150], formerly deployed by pinterest, to combine the analysis capabilities of LLMs and the explainability features of this graph based recommender system.

This work will go on after the release of the deliverable, so what we present here is progress until M48.

#### 7.1.3. Experiments

To give a glimpse of the application, Figures 14 and 15 show screenshots of the backend view, analyzing two articles on the upcoming US elections. The important choices we made are:

- We use pre-trained LLMs (llama3.1 and ChatGPT 4o) that are easily accessible, and do not fine tune LLMs on news data;
- We extract metadata from the news article by "asking" the LLM, not using fine tuned classifiers;
- Likewise, the article comparison is done by the LLM, according to a custom "receipt", i.e., prompts for the LLM.

The background of these decisions is, that we want it to make it very easy to switch between LLMs. First, to avoid a vendor lock-in and to preserve a certain degree of freedom regarding the project's dependencies, but second, to profit in the expected progress in the state of the art. While a fine tuned LLM might be better *today*, compared to a generic LLM, it is not clear whether a generic LLM *tomorrow* will outperform it again.

#### 7.1.4. Conclusion

The main contributions of this work are to:

- showcase the use of LLMs as a source of labelled data for hybrid recommender systems;
- showcase the use of LLMs for cross language text comparison;
- build an analysis platform aggregating different content with different (political) stances for a similar topic or event.

#### 7.1.5. Relevant software/datasets/other outcomes

- For an update on the current status of the demo, please contact us via https://www.idmt.fraunhofer.de/en/contact.html.

#### 7.1.6. Relevance to AI4Media use cases and media industry applications

This work is relevant for all partners involved with news recommendation, especially UC1 and UC2. In general, this work has relevance for all news publishing media.

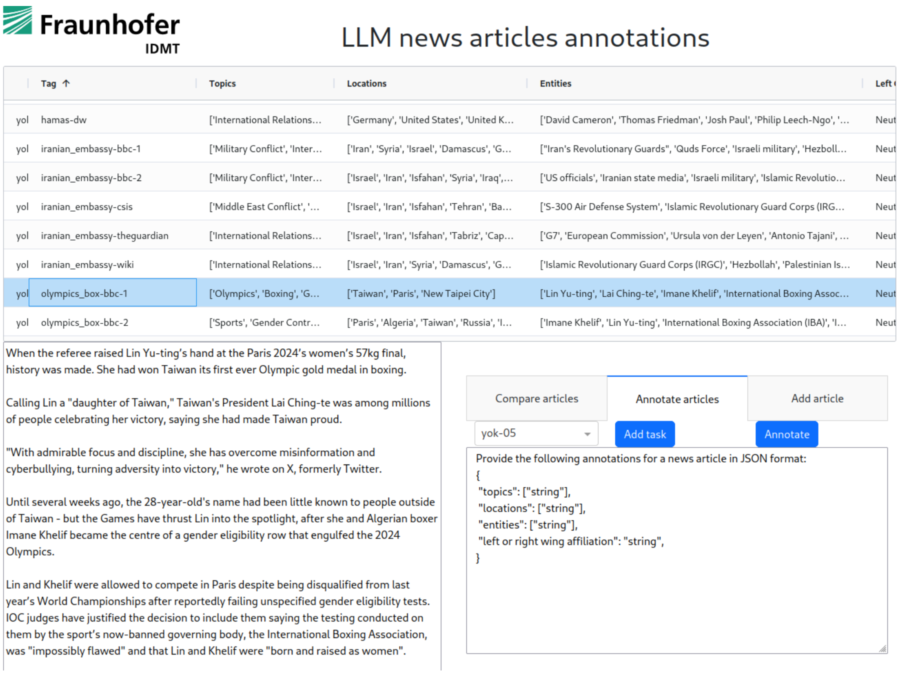



*Figure 14. Debug view of articles already analyzed together with the extracted meta data and the annotation prompt.*

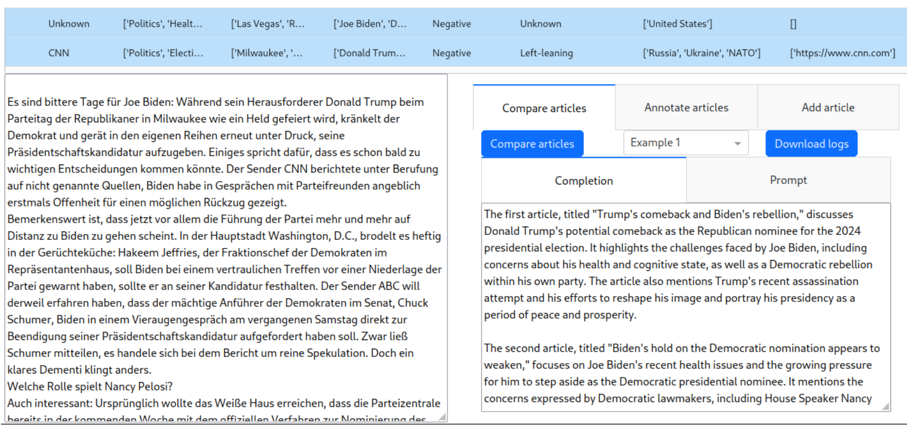



*Figure 15. Example output of a full text article comparison. Note that one of the input texts is German, showcasing the multi-linguality of the system.*

# 8. AI for Healthier Political Debate (T6.4)

## 8.1. Analyzing the Political Positioning of Dutch Belgian News Sources

**Contributing partner:** CEA

### 8.1.1. Introduction

News provides citizens with essential information about political actions and events, contributing to their understanding of society. However, the large amount of content available online makes it difficult for interested stakeholders to grasp the tendencies and orientations expressed in the news. Such an understanding is needed for an adequate interpretation of the political content. Automatic tools can analyze large amounts of data and summarize insights for different stakeholders, including citizens, journalists, political scientists and social scientists. The contribution applies NLP tools to Dutch Belgian news sources and provides such insights.

### 8.1.2. Method

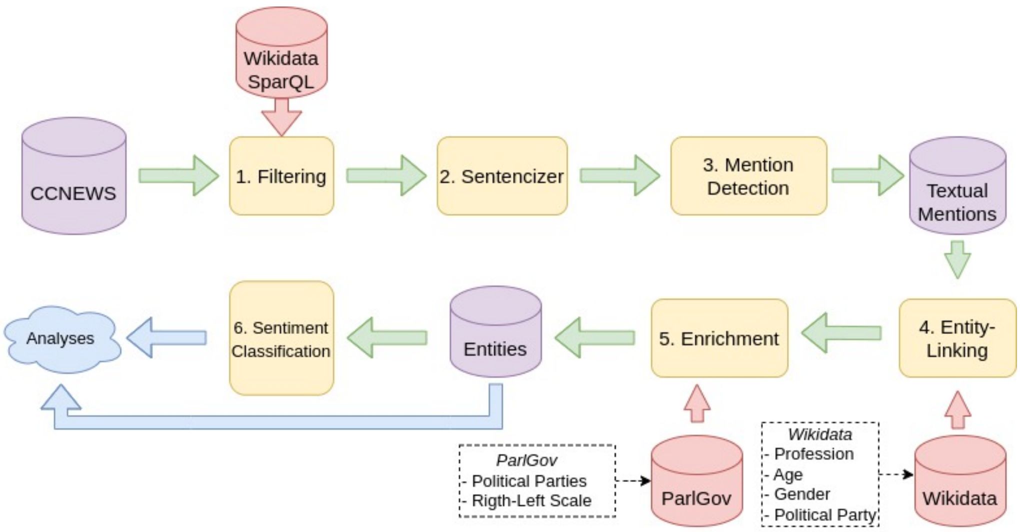



*Figure 16. Proposed news analysis pipeline that starts with political texts, includes a comprehensive NLP processing pipeline including preprocessing, entity detection and sentiment analysis, and links the outputs with external databases to give rich insights about the news content.*

We propose a pipeline for automatic news analysis and summarize it in Figure 16. The pipeline combines different NLP components, such as named entity detection, named entity linking, named entity political characterization (political orientation, demographics), and target-dependent sentiment classification. Importantly, we link the outputs of NLP analysis with external domain-related datasets to provide rich insights about the content of political news. These insights can be aggregated using different axes, such as news sources, types of political events, and demographic attributes (gender, age), to meet the needs of different stakeholders. The pipeline is designed to have minimal dependence on particular languages and this flexibility facilitates comparisons across countries. In D6.3, we introduced a framework combining objective and subjective news analysis and instantiated it for French news sources. Here, we extend this framework to analyze

Dutch-language news sources from Belgium. This work results from Evan Dufraisse's secondment between VRT and CEA and was integrated into the project's UC2.

### 8.1.3. Experiments

**8.1.3.1. Dataset and Preprocessing.** We sampled news articles from Dutch Belgian newspapers from the CC news corpus by collecting news from the "society", "international", "political", and "economy" subcategories. We use articles from other subcategories (e.g., "sports", "lifestyle", "real estate", "art", etc.) as negative examples to train a binary XLM-RoBERTa model that filters relevant articles from others. Then, we use this classifier to create a dataset including 1.36M potentially relevant articles. After entity detection and disambiguation, we obtain 424,811 politician mentions and we keep the 24 outlets with at least 1,000 mentions for counting and sentiment analysis. Sentiment analysis is done using the Dutch model based on the MAD-TSC dataset [151]. This model infers the sentiment expressed for each person's name present in a sentence.

**8.1.3.2. Political Analysis** We discuss the positioning of Belgian newspapers with respect to political tendencies and compare this positioning to the one obtained for French newspapers. We refer the reader to D6.3 for the details of the French analysis used for comparison here.

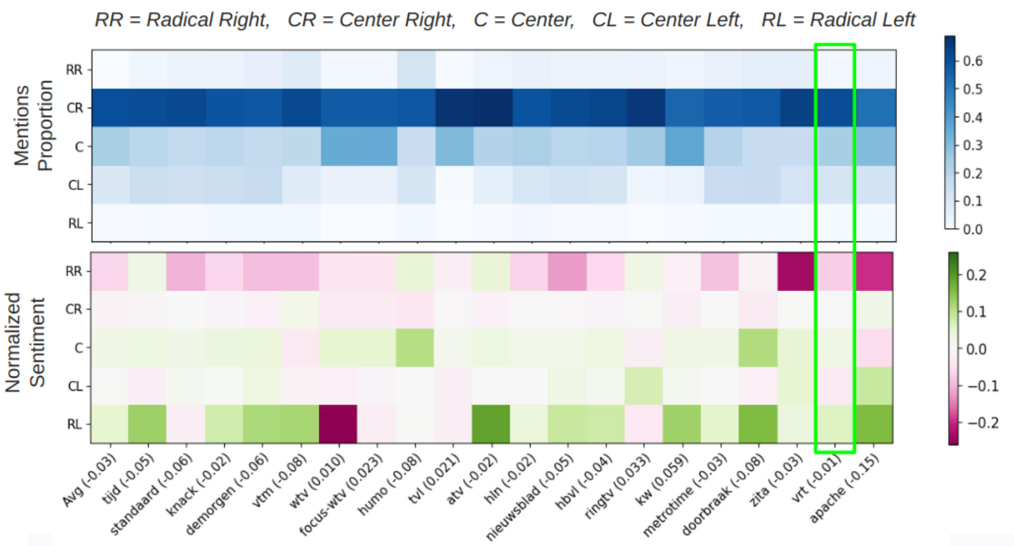



*Figure 17. Deviation of the sentiment associated with major political orientations from the average sentiment of each source. Average sentiment of linked politicians is indicated in parentheses for each source.*

The frequency and sentiment analysis of media outlets provide insights similar to those observed in the French landscape analysis. We again see a higher mention frequency of ruling center-right politicians (see Figure 17). As in the French case, there is a noticeable negative sentiment bias towards the radical right (RR) and a positive sentiment bias towards the radical left (RL). Preliminary experiments done internally by VRT show that journalists find the analysis useful to get insights about the positioning of their own news outlet and that of competing ones. A political expertise or reference on the Belgian media landscape would be necessary to further interpret the results.

#### 8.1.3.3. Demographics Analysis

**Gender Representation**: Similar to the French analysis, the Belgian news landscape shows a significant gender disparity in political representation, as visible in Figure 18, despite the fact that women make up over 40% of parliamentary representatives[10]. Although the disparity in political representation is significantly greater at the local level, with only 15.8% of mayors being women, local news outlets tend to have better representation of women in their coverage on average. Similar to the trend in France, recent years have not shown a clear increase in the proportion of mentions of female politicians (see Figure 19).

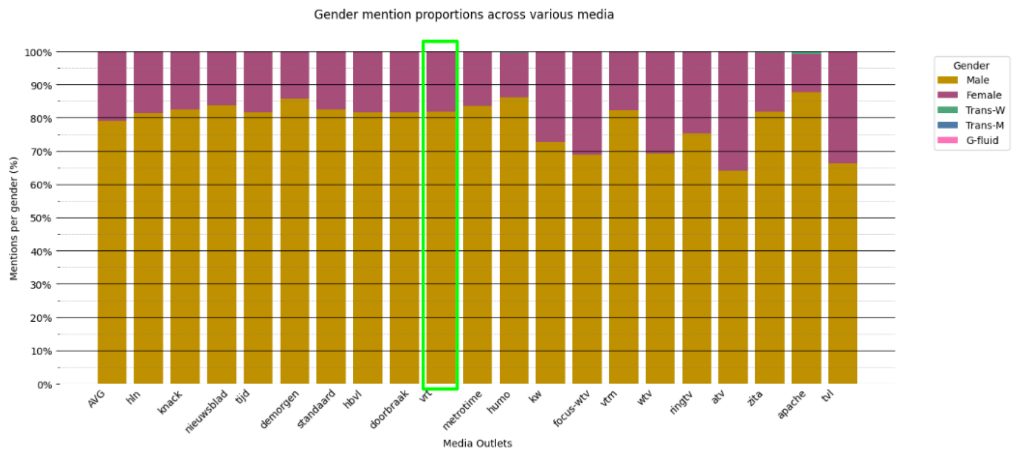



*Figure 18. Gender mention proportions across various media outlets. The bars indicate the percentage of mentions per gender for each media outlet.*

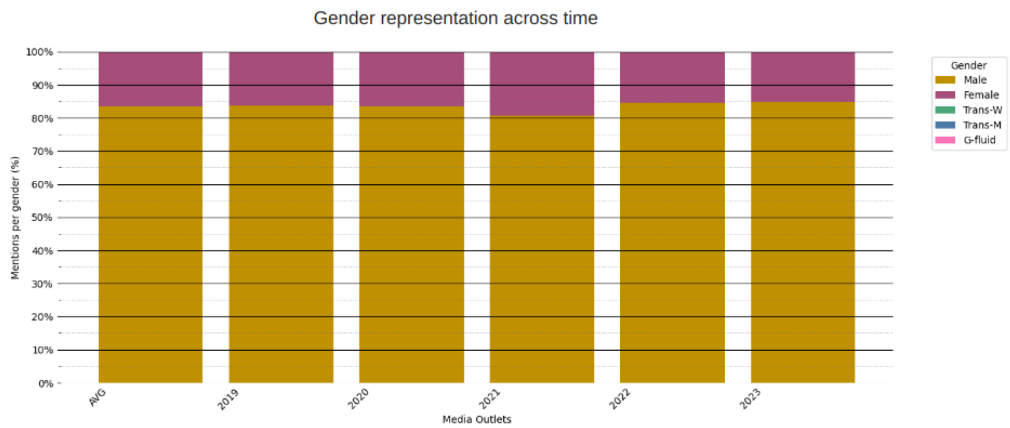



*Figure 19. Gender representation across various media outlets from 2019 to 2023. The bars show the percentage of mentions per gender each year, including an average (AVG) column.*

**Age Representation**: Similar to France, Figure 20 shows that while the average age of Belgian

[10] Lukas Taylor, "Sharp rise in Belgian female politicians in 30 years," The Brussels Times, 2024. https://www.brusselstimes.com/573261/sharp-rise-in-belgian-female-politicians-in-30-years

members of parliament is around 47 years[11] , the average age of mentioned Belgian politicians is approximately 57 years. Local news outlets tend to again differentiate themselves by mentioning younger politicians on average.

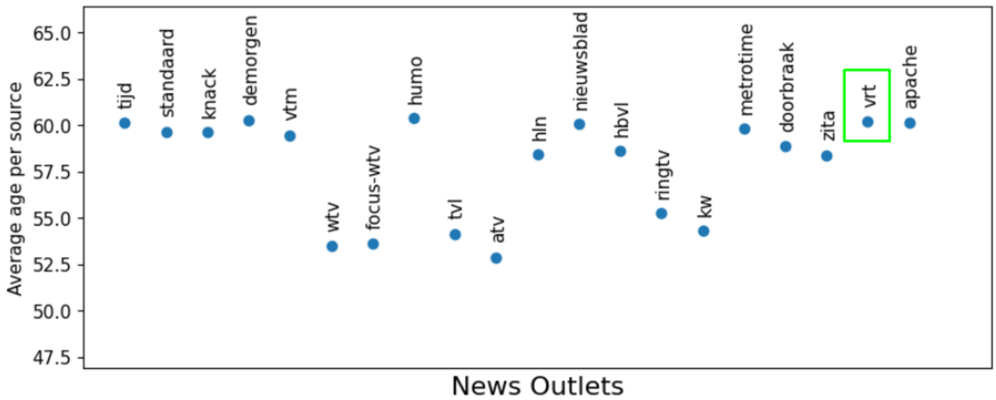



*Figure 20. Average age of politicians mentioned per source.*

### 8.1.4. Conclusion

The proposed analysis shows the political positioning of Dutch-language news sources and elicits their demographic biases. It also compares these results with the ones reported for a similar analysis done for France, highlighting common points and differences. The analysis can be extended to other countries and be used by media organizations and regulators to better understand the content of political news at an aggregate level. However, potential users should keep in mind that the obtained results are not 100% accurate, as is the case with any automatic analysis pipeline.

### 8.1.5. Relevant publications

- E. Dufraisse, A. Popescu, J. Tourille, A. Brun, O. Hamon. "Combining Objective and Subjective Perspectives for Political News Understanding", https://arxiv.org/abs/2408.11174. Under review.

### 8.1.6. Relevance to AI4Media use cases and media industry applications

The pipeline was packaged as a docker and was delivered to VRT as part of their UC2 work. It is tested internally by journalists as a way to automatically audit the positioning of news sources.

## 8.2. Predicting International Political Debates 2

**Contributing partner:** UvA

### 8.2.1. Background - Conceptualization and Methodologies

The proliferation of digital tools and the fast-growing increase in digital materials have created very large digitised and born-digital archives recording international politics. This work presents

[11] IPU Parline, "Data on age: By country," available at https://data.ipu.org/content/belgium?chamber_id=13343

novel methodologies and tools for changing memory and power relations in digital archives of international politics through new ways of reassembling marginalised, non-canonical entities in digital archives. Reassembling digital archives can take advantage of the materiality and the algorithmic processuality of digital collections and reshape them to inscribe lost voices and previously ignored differences.

Digital archives are not fixed and are changed with new research and political questions. They are only identified through new questions. We consider both the extension of archives towards evidence that is otherwise thrown away as well as the provision of new intensive, non-discriminatory viewpoints on existing collections. Visualized in Figure 21 is a relationship-extraction experiment for asylum cases. We can see how the relations are formed around social-media platforms and what role they play in the cases. The graph of social-media relations retells the story of experiences in asylum applications through visual narratives. Unfortunately, this kind of advanced reassembling through new relations comes at great computational cost with the script running for several hours. The results should also not be read as the definite statement on the relations in the archives but as an incomplete but novel perspective on the archive. They show how language models can play a role in re-assembling archives.

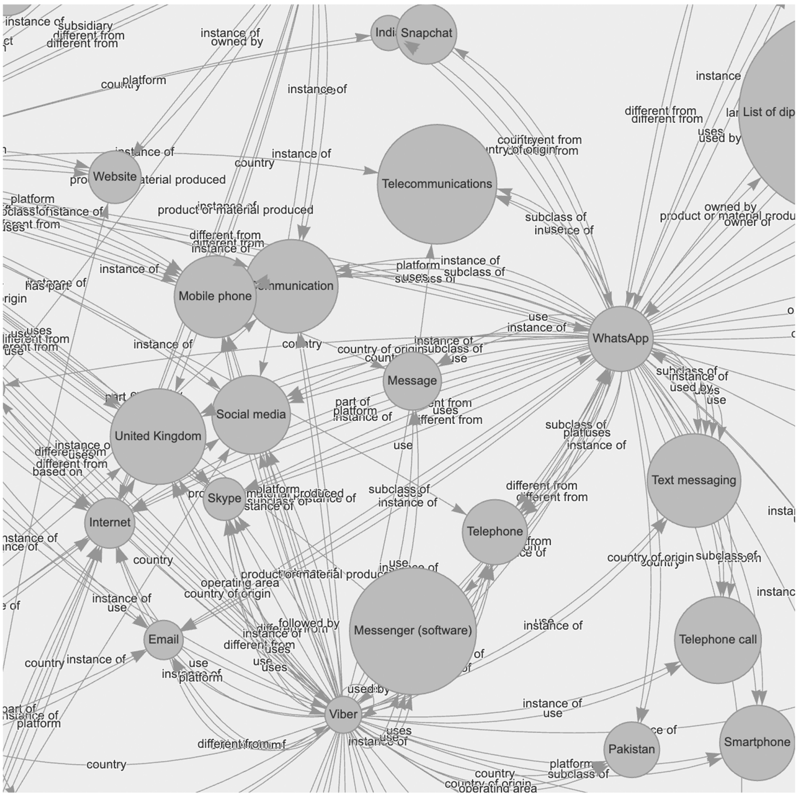



*Figure 21. Network visualisation of entities in Tribunal Decisions*

### 8.2.2. Experiments

Similar to Figure 21, work presents six distinct techniques and strategies to reassemble digital archives and renders these according to three different types of new digital archives. Digital experiments include the dynamic scraping of collections, temporal sampling of digital archives, relationship extraction with LLMs, entity recognition and seeded topic modelling. Table 24 shows the techniques and strategies.

| | Type | Grassroots research | Along the grain | Against the grain |
|---|---|---|---|---|
| Legal Cases (Asylum) | Incidental Document Collection | Stories | Generic Search | Rebuilding the whole archive |
| | Subdocuments | Social Media Evidence | Issue-based search | Visual Narratives from Entities and Relations |
| Public Transparency | Tender and Procurement | Access to Government Practices | e-Repository for diverse stakeholders | Virtual Collection |
| | Verbatim Parliamentary Transcripts | Issue-based Democratic Discourse | Historical Metadata Scheme | Syntax Subjects and Actions |
| Web Archive | Real-time Archive | Monitoring Historical Activities | Seeding Sites | Temporal Sampling |
| | Time-indexed Archive | Unknown Histories | Seeding Keywords | Human-machine Interactions |

*Table 24. Strategies to reassemble digital archives*

Table 24 shows how legal cases and public transparency materials can be reassembled into useful archives for research, focusing on asylum seekers' experiences and government policies. We highlight the existence of "incidental archives," which are collections of mundane documents available on the web without additional metadata. These archives are accessible due to the affordability of generic search algorithms. Two strategies are proposed for making sense of these materials:

- Rebuilding Archives: This involves collecting all relevant materials and reconstructing them from scratch into new, more organized archives.
- Issue-Based Search: This method focuses on specific issues, such as social-media applications, allowing researchers to extract entities and their relationships using language models to form visual narratives.

Legal cases can thus be reassembled into records of everyday experiences and the difficulties asylum seekers encountered making their claims for protection. While they are bound to attract strong grassroots research interests looking for underrepresented experiences, these are also examples of incidental archives that appear on the Web as mundane document collections.

Also online are many examples of public transparency materials, our second example of incidental archival materials, often published by governments or companies, which can be difficult to interpret. The example of UK's Hansard is used to demonstrate how these materials can be broken down into smaller units, like sentences, to analyze political relations and actions. This approach helps researchers understand typical actions within these archives and trace the development of specific discussions, such as those involving GCHQ.

Finally, the article discusses reading web archives with a "temporal grain," which involves analyzing snapshots over time. This method has been used for monitoring NGOs' political campaigning and indexing how their discourses evolve historically. By using temporal sampling and keyword seeding, researchers can better describe the development of debates on security within the UK parliament, combining machine and human analysis. Political web archives are built around snapshots with temporal metadata, which we have used in two research cases for real-time archiving as well as historical indexing. We have developed a temporal sampling that provides a more even distribution of larger and smaller websites than the Internet Archive. We have also used seeded keywords to enable human-machine collaborations that describe political developments better than either machine or humans alone could have done.

#### 8.2.3. Conclusion

The work has presented several techniques and strategies to reassemble digital archives, taken from diverse data-science projects in international politics.

Table 24 also shows that reassembling digital archives remains partial and incomplete. There are many more archives to cover and many more non-archives to transform through grassroots research.

#### 8.2.4. Relevant publications

- Blanke, T. Reassembling digital archives—strategies for counter-archiving. Nature Humanities Social Science Communications 11, 201 (2024). https://www.nature.com/articles/s41599-024-02668-4.

#### 8.2.5. Relevant software/datasets/other outcomes

N/A

#### 8.2.6. 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. It is especially relevant here as it also showcases some shortcomings of these techniques that provide valuable lessons to the community.

### 8.3. Argument-based Detection and Classification of Fallacies in Political Debates

**Contributing partners:** UCA

#### 8.3.1. Introduction

Fallacies are arguments that employ faulty reasoning. Given their persuasive and seemingly valid nature, fallacious arguments are often used in political debates. Employing these misleading arguments in politics can have detrimental consequences for society, since they can lead to inaccurate conclusions and invalid inferences from the public opinion and the policymakers. Automatically detecting and classifying fallacious arguments represents therefore a crucial challenge to limit the spread of misleading or manipulative claims and promote a more informed and healthier political discourse. Our contribution to address this challenging task is twofold:

- First, we extend the ElecDeb60To16 dataset of U.S. presidential debates annotated with fallacious arguments, by incorporating the most recent Trump-Biden presidential debate. We include updated token-level annotations, incorporating argumentative components (i.e., claims and premises), the relations between these components (i.e., support and attack), and six categories of fallacious arguments (i.e., Ad Hominem, Appeal to Authority, Appeal to Emotion, False Cause, Slippery Slope, and Slogans).
- Second, we perform the twofold task of fallacious argument detection and classification by defining neural network architectures based on Transformers models, combining text, argumentative features, and engineered features. Our results show the advantages of complementing transformer-generated text representations with non-textual features.

Whilst most of the computational approaches targeting fallacious argumentation focus on the pure classification of such nefarious content [152, 153, 154, 155, 156, 157], the originality of our contribution is that it proposes, to the best of our knowledge, the first neural architecture to both detect fallacious arguments and classify them in political debates, and one of the very few approaches to tackle this task in general, outperforming competing approaches [158, 159].

### 8.3.2. Method

**8.3.2.1. ElecDeb60to20 Dataset** To address the task of detecting and classifying fallacious arguments within political debates, we decided to rely on the ElecDeb60To16 dataset [160, 155]. It comprises televised debates from U.S. presidential election campaigns spanning from 1960 to 2016. These debates were sourced from the website of the Commission on Presidential Debates[12], which openly provides transcripts of debates broadcasted on television and featuring the prominent candidates for presidential and vice-presidential nominations in the United States. Considering the most recent presidential election between Trump and Biden occurred in 2020, we expanded the dataset with the transcripts of the debates of this election campaign to include updated annotations, incorporating argumentative components such as Claims and Premises; as well as the relations between these components, i.e., Support and Attack; and six fallacy categories. As a result of this annotation update, the dataset is renamed as ElecDeb60to20[13], reflecting the coverage of debates spanning from 1960 to 2020. We performed the annotations based on the following fallacy categories:

1. Ad Hominem: When the argument becomes an excessive attack on an arguer's position.
2. Appeal to Emotion: The unessential loading of the argument with emotional language to exploit the audience's emotional instinct.
3. Appeal to Authority: It occurs when the arguer relies on the endorsement of an authority figure or a group consensus without providing sufficient evidence. It may also involve the citation of non-experts or the majority to support their claim.
4. Slippery Slope: This fallacy implies that an improbable or exaggerated consequence could result from a particular action.
5. False Cause: The misinterpretation of the correlation of two events for causation.
6. Slogan: It is a brief and striking phrase used to provoke excitement of the audience, and is often accompanied by another type of fallacy called *argument by repetition*.

Two annotators, with expertise in computational linguistics, independently annotated the new portion of the dataset (Trump vs. Biden debates) by identifying argumentative components, relations, and fallacies. A set of 50 sentences randomly extracted from the debates was annotated to assess Inter-Annotator Agreement (IAA), and the results, visualized in Table 25, indicate a substantial level of agreement between the annotators.

| **Measure** | **Value** |
|---|---|
| Observed Agreement | 0.857 |
| Krippendorff's $\alpha$ | 0.757 |

*Table 25. IAA agreement over 50 sentences randomly extracted from the 2020 Trump-Biden debates.*

Table 26 summarizes the Trump vs. Biden's debates annotations per category and argumentative features. We tokenized the annotated fallacious arguments to compute the average number of words in each category. Slogans is the shortest with 5.0 tokens on average, whereas SlipperySlope was the longest with 20.5 tokens on average. In the last debate, the most used fallacies are *AppealtoEmotion* and *AdHominem*, confirming the trend of the previous debates. Behind this strategy,

[12] https://www.debates.org/voter-education/debate-transcripts/
[13] https://github.com/pierpaologoffredo/FallacyDetection

there are many references to the COVID-19 pandemic and some personal issues of the two candidates exploited during the debates.

| Category | Freq | AvgTok | Arg. Feature | Freq |
|---|---|---|---|---|
| Ad Hominem | 62 | 4,6 | Claims | 1513 |
| AppealtoAuthority | 17 | 18,6 | Premise | 332 |
| AppealtoEmotion | 147 | 6,81 | Support Rel. | 400 |
| FalseCause | 0 | 0 | Attack Rel. | 112 |
| SlipperySlope | 4 | 20,5 | | |
| Slogans | 2 | 5 | | |
| **Total** | 232 | 9,25 | **Total** | 2357 |

*Table 26. Distribution of annotated fallacies per category and argumentative features of Trump vs. Biden's debates.*

The train and test set split (90% and 10% respectively) was performed considering the entire new dataset ElecDeb60to20. The distribution of fallacy labels is as follows: AppealtoEmotion (59.94%), AppealtoAuthority (15.20%), AdHominem (13.58%), FalseCause (46.93%), Slipperyslope (3.97%), and Slogans (2.63%).

**8.3.2.2. Fallacy Detection** We rely on the *BIO*/*IOB* data format, and specific tags are assigned to annotate the fallacies, such as B-AdHominem, I-AdHominem, I-AppealtoAuthority, etc. The fallacy detection and classification tasks consist therefore in assigning one of these thirteen predefined labels to each token. We build a contextual framework that includes the sentence containing the fallacy, as well as the *preceding* and *following* sentences. When the fallacious sentence is the first or last in the dialogue, the preceding or following sentence is excluded.

**8.3.2.3. Models** To address the above-mentioned tasks, we employ transformer-based architectures[14].

**BERT + (Bi)LSTM(s)** The simplest models consist of a pre-trained BERT model followed by either *(i)* an LSTM layer and a dense layer, or *(ii)* a BiLSTM layer with 0.2 dropout, an LSTM layer, and a dense layer. The weights of the transformer are kept frozen during training. The text serves as input for the transformer, and we extract the last hidden states (i.e., the embedded representation of each token). This output is then passed on to the subsequent layers. In the case where argumentative features are included in the model, we concatenate the last hidden states of the transformer with the one-hot-encoded representation of the argument components and relationships. This concatenated feature representation is then fed into the next RNN-based layers. All models used Adam optimizer with default PyTorch parameters.

**Fine-tuned models for token classification** We fine-tuned transformers models that were designed for token classification tasks. The distinguishing factors among the models, in addition to their architectures, lie in the utilization of distinct corpora for pre-training.

- BertForTokenClassification.
- DebertaForTokenClassification.
- ElectraForTokenClassification.

[14] https://huggingface.co/docs/transformers/tasks/token_classification

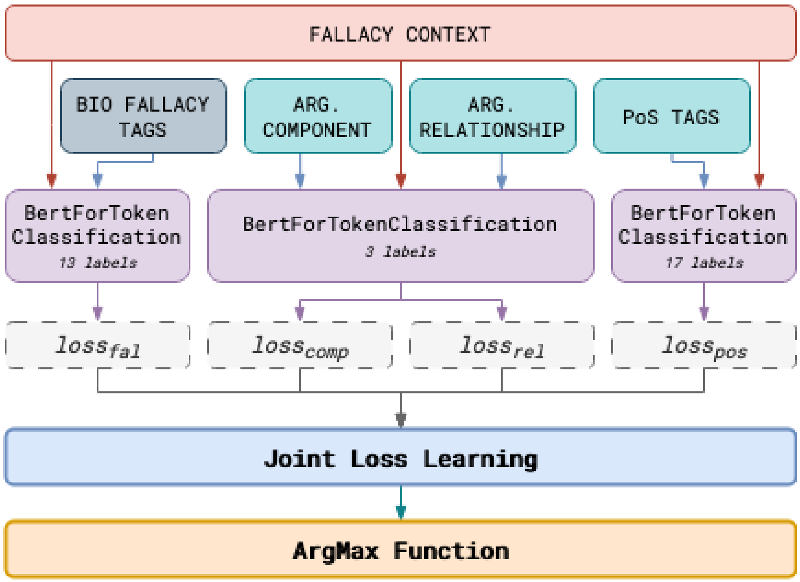



*Figure 22. MultiFusion BERT with $joint_{loss}$ approach.*

- DistilbertForTokenClassification.

**MultiFusion BERT** Among the previously mentioned models, BertFTC showed the best performance in our task. We enhance its capabilities to detect fallacious text by integrating and "fusing" additional features, namely argumentative components (Claim, Premise), argumentative relations (Support, Attack), and Part-of-Speech (PoS) tags.

Figure 22 illustrates the proposed model, called MultiFusion BERT, for the detection and classification of fallacies in political debates. MultiFusion BERT computes logits (L) for each feature by employing a specialized TokenForClassification Transformer model adapted to the number of labels: 3 for components and relations, and 17 for part-of-speech tags. The architectures for argumentative features for components and relations share the same parameters, enabling us to obtain *logits* for both components and relations. An additional model, based on the number of PoS tags (i.e., 17), is used to obtain *logits* for PoS features. Consequently, distinct losses are computed for each model: fallacy loss ($loss_{fal}$), component loss ($loss_{cmp}$), relation loss ($loss_{rel}$), and part-of-speech loss ($loss_{pos}$). These individual losses are combined by multiplying them with an arbitrary $\alpha$ value of 0.1, yielding a unified average loss referred to as the $joint_{loss}$ [161]. In our study, we opted for empirically investigating the optimal alpha value that yielded superior performance, as evidenced by our experiments. The back-propagation function incorporates all losses in the following way:

$$joint_{loss} = \alpha * \frac{(loss_{fal} + loss_{cmp} + loss_{rel} + loss_{PoS})}{N_{loss}}$$

, where $N_{loss}$ denotes the number of losses considered by the model. We conducted an exploration of various values for the $\alpha$ parameter.

### 8.3.3. Experiments

Table 27 presents the results of the tested models for fallacies detection in the political debates. It can be seen that MultiFusion BERT, incorporating argumentative features (components and rela-

| Model | Avg macro F1 Score |
|---|---|
| BERT + LSTM | 0.4697 |
| BERT + LSTM (comp. and rel. features) | 0.5142 |
| BERT + BiLSTM + LSTM | 0.5495 |
| BERT + BiLSTM + LSTM (comp. and rel. features) | 0.5614 |
| BertFTC `bert-base-uncased` | 0.7096 |
| BertFTC `dbmdz/bert-large-cased-finetuned-conll03-english` | **0.7237** |
| DebertaFTC `microsoft/deberta-base` | 0.7222 |
| ElectraFTC `bhadresh-savani/electra-base-discriminator-finetuned-conll03-english` | 0.4033 |
| DistilbertFTC `distilbert-base-cased` | 0.7010 |
| DistilbertFTC `distilbert-base-uncased` | 0.7047 |
| MultiFusion BERT (comp., rel. and PoS features) | **0.7394** |

*Table 27. Average macro F1 scores for fallacy detection (BIO labels are merged) using different models. The scores are based on an average of 3 runs, except for BERT + (Bi)LSTM(s) models, which were evaluated using 10 runs. (FTC stands for "ForTokenClassification)*

| Features | | | Avg macro |
|---|---|---|---|
| **Components** | **Relationships** | **PoS** | **F1 Score** |
| ✓ | | | 0.6922 |
| | ✓ | | 0.6922 |
| | | ✓ | 0.7212 |
| ✓ | ✓ | | 0.7278 |
| ✓ | | ✓ | 0.7166 |
| | ✓ | ✓ | 0.7166 |
| ✓ | ✓ | ✓ | **0.7394** |

*Table 28. Average macro F1 scores for fallacy detection (BIO labels are merged) using MultiFusion BERT and different features. The scores are based on an average of 3 runs.*

tions) as well as PoS tags, significantly outperformed the other models (the performance increase with respect to BertFTC is of 2.12%).

Table 28 presents the results obtained by MultiFusion BERT using all possible combinations of features. Incorporating argumentative components, relations, and PoS features individually or in pairs resulted in a decline in performance compared to the best baseline results (i.e., BertFTC *dbmdz/ bert-large-cased-finetuned-conll03-eng.*), with an average degradation of 4.35% across the different configurations (excluding the one considering all three features). In contrast, when all three features are included a significant improvement in model performance is observed highlighting the importance of considering all of them together for fallacy detection.

Table 29 shows that the identification of tokens labeled as Slogans exhibits the poorest results, despite being relatively easier to recognize for humans. This can be due to the limited presence of examples/tokens in both the training and the test set. On the contrary, tokens labeled as "Slippery Slope" and "False Cause" (with 332 and 321 examples, respectively) are much better classified by the model, showing the highest performances (0.89 and 0.84). The definition of "Slippery Slope" revolves around portraying improbable or exaggerated consequences arising from a specific action, and argumentative components are often used to the cause, as well as semantic nuances well captured by the model.

| Label | precision | recall | f1-score | support |
|---|---|---|---|---|
| AdHominem | 0.99 | 0.77 | 0.87 | 739 |
| AppealtoAuthority | 0.90 | 0.78 | 0.83 | 1'049 |
| AppealtoEmotion | 0.82 | 0.77 | 0.79 | 2'224 |
| FalseCause | 0.82 | 0.86 | 0.84 | 321 |
| Slipperyslope | 0.90 | 0.88 | 0.89 | 332 |
| Slogans | 0.00 | 0.00 | 0.00 | 49 |
| O | 0.90 | 0.95 | 0.93 | 7'914 |
| accuracy | | | 0.89 | 12'628 |
| macro avg | 0.76 | 0.72 | **0.74** | 12'628 |
| weighted avg | 0.89 | 0.89 | 0.89 | 12'628 |

*Table 29. Classification report of Fallacy Detection and Classification with B and I labels merged.*

We analyze the normalized confusion matrix of true labels against predicted labels. The confusion matrix reveals that the model tends to over-predict instances in the non-fallacious category. Moreover, False Cause and Appeal to Emotion are the classes that the models misinterpret the most as non fallacious. In a smaller proportion, the model misclassifies instances of Appeal to Authority as Appeal to Emotion.

### 8.3.4. Conclusion

The main contributions of this work are to:

- Provide a new annotated corpus (ElecDeb60to20) as an extension of the ElecDeb60to16 dataset by incorporating the Trump vs. Biden 2020 presidential debate along with argumentative annotations and fallacies.
- Propose and evaluate MultiFusion BERT, a transformer-based architecture that combines the debate text, the argumentative features (i.e., components and relations), and engineered features to perform the fallacy detection and classification task.

### 8.3.5. Relevant publications

- Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. Argument-based Detection and Classification of Fallacies in Political Debates. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 11101–11112, Singapore. Association for Computational Linguistics, . [162]

### 8.3.6. Relevant software/datasets/other outcomes

- The ElecDeb60to20 dataset is available at https://github.com/pierpaologoffredo/FallacyDetection.

### 8.3.7. Relevance to AI4Media use cases and media industry applications

This study contributes to UC2 "AI for News" as it provides insights into detecting and classifying various categories of fallacies in political debates. This capability is important for media organizations aiming to ensure the integrity of their reporting and prevent the spread of misinformation.

The models proposed in this work support the production of more reliable news content, and ultimately help maintain public trust in media.

It also aligns with UC4 “AI for Social Sciences and Humanities” and UC1 “AI for Social Media and Against Disinformation” by addressing the need to identify misleading arguments. The ability to recognize fallacies is vital in combating the spread of harmful narratives. This research serves as a step towards effectively detecting, explaining, and countering fallacies and disinformation, thereby promoting a more informed and discerning public discourse.

# 9. Perceptions of hyper-local news (T6.5)

## 9.1. Frame analysis of TV transcripts in Dutch

**Contributing partners:** IDIAP, NISV

### 9.1.1. Introduction

In the current media landscape, understanding the framing of information is crucial for critical consumption and informed decision making. Framing analysis is a valuable tool for identifying the underlying perspectives used to present information, and has been applied to a variety of media formats, including television programs. However, manual analysis of framing can be time-consuming and labor-intensive. This is where large language models (LLMs) can play a key role. Frame recognition is a technique for understanding how media content shapes our perception of the world by helping media audiences identify the underlying perspectives used to present information. Identifying frames in the news allows us to understand which aspects of the story are being emphasized. For example, given two news stories on the same topic, such as the war between Russia and Ukraine, one may emphasize the economic impact on Europe due to this event (economic frame), while the other may show the personal case of a family displaced by the war and left homeless (human interest frame). Understanding these frames can help readers and viewers of media to act as more critical media consumers and to make informed decisions about the information consumed.

The use of LLMs for frame analysis is a relatively new field of research, but it is attracting increasing interest as LLMs are effective in identifying frames in text and can be applied to a variety of media formats, such as headlines, tweets, or news articles [163, 164, 165]. This has ignited a growing interest in their potential application to spoken content, such as television transcripts, which is the focal point of our work. Framing analysis plays a pivotal role in comprehending how news programs shape our perception of reality. By identifying the underlying frames employed in newscasts, we can gain insights into the perspectives presented and the potential influence on public opinion [166]. However, conventional framing analysis methods are often labor-intensive [167], thus restricting their applicability to large-scale analysis. LLMs offer a viable solution to this challenge towards automating the frame detection process, enabling more efficient and comprehensive analysis of media content.

### 9.1.2. Methodology

In this section, we explain our methodology for framing analysis, describing the steps we followed: on one hand, the collection and translation of the data; and on the other hand, the human annotation and LLM classification.

**Data.** The dataset we used is a selection of 2,000 news media items from broadcasts of the public Dutch television news programs EenVandaag (1,000 items) and Nieuwsuur (1,000 items).

EenVandaag [15] (OneToday) is a daily evening program broadcasted on Dutch public television channel NPO1. EenVandaag has the format of a news program with current issues and background information behind the news. The program is about 30 minutes long and deals with various news topics during an episode. The program has multiple presenters introducing various news items, and also interview experts live in the studio.

Nieuwsuur [16] (News Hour) is also an evening program and is broadcasted on NPO2, another

[15] https://eenvandaag.avrotros.nl/

[16] https://nos.nl/nieuwsuur

Dutch public television channel. The broadcasts are between 30 and 45 minutes long, and also have the format of a news program with current issues and background information behind the news. This program also has multiple presenters and live interviews with experts.

We chose these two current newscasts as they provide a good overview of Dutch news items on a daily basis. These shows also provided a large corpus of items over the years, with little change in the show format. Due to this, the data is very consistent.

For analysis, the spoken words in the video recordings of these programs are transcribed. This was done with the open-source, Kaldi automatic transcriber [168]. This software can automatically transcribe Dutch spoken language into text. This pre-processing step resulted in a dataset of 2,000 texts covering news between 2014 and 2018, varying in length, with an average number of 499 words for Nieuwsuur and 664 words for EenVandaag.

To be able to use GPT-3.5 in an optimal way, and to obtain results in which the language would not be detrimental to model performance, we translated the content into English using the Deepl API [17].

**Annotation.** For this task, we engaged a person with university-level education and a basic knowledge in the field of framing. This was needed to make sure that the annotator was aware of the potential for human bias and was thus able to look at the texts in the most objective way possible. More specifically, the annotator had a background in language and cultural studies.

To carry out the labeling of the frames in the data, we first designed a codebook with several definitions of the frame analysis concept. This was then used to train the annotator on the task. In addition, we created an interactive environment through Google Forms to do the annotation. Using Google Apps Script[18], we created a script that allowed us to generate forms automatically.

In the annotation form, we first showed the piece of text to annotate, followed by the definitions of the 5 types of frames proposed by [169]:

- Attribution of responsibility. This frame presents an issue or problem in such a way as to attribute responsibility for its cause or solution to a government, an individual, or a group.
- Human interest. This frame brings a human face or an emotional angle to the presentation of an event, issue, or problem.
- Conflict. This frame emphasizes conflict between individuals, groups, or institutions as a means of capturing audience interest.
- Morality. This frame puts the event, problem, or issue in the context of religious tenets or moral prescriptions.
- Economic. This frame reports an event, problem, or issue in terms of the consequences it will have economically on an individual, group, institution, region, or country.

After these definitions, we asked the annotator (1) to define the main frame; (2) to define an alternative frame if there was one; (3) to copy-paste sentences that helped the annotator chose the main frame; and (4) to add free text in a section called comments, in case that the annotator had something to explain.

**Classification with GPT-3.5** The generative language model is used as follows. This type of model, given an input (called prompt) is able to generate text that continues that prompt (called output.) As a simple example, if we ask a generative model "how many sides does a triangle has?", it will generate an output through a series of tokens (a token can be a word, or a smaller unit, so a

[17] https://www.deepl.com/es/pro-api?cta=header-pro-api/
[18] https://www.google.com/script/start/

word can be formed by more than one token), and those tokens have a probability, which reflects how confident the model is of the answer, based on the text it has been trained with. In this simple example, the answer would be “three” with a 100% probability of that token. Based on this idea (i.e., that the generative model produces an answer and gives a probability to the answer), we defined a prompt in which we first pass the definitions of the different frame types, then we pass the text to classify, and finally we ask the model, among the 5 frame type options that we gave, which was the most likely frame. Figure 23 shows an example of the prompt used. The model gives a probability to each of the 5 frame options, so the frame with the highest probability is the predominant frame identified by the model. Furthermore, the fact of being able to access the probabilities given to the other frame types, allows us to study cases where more than one frame was possible, because the second or third options had a high probability as well.

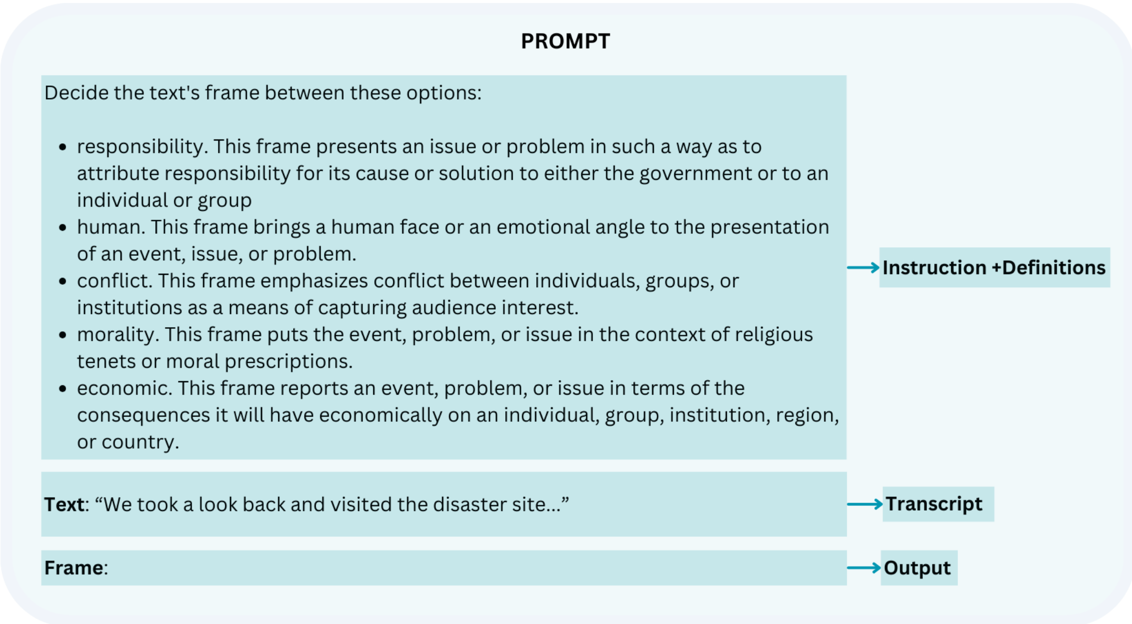



*Figure 23. Example of the prompt used for frame classification, given a transcript.*

#### 9.1.3. Experiments

Once the labeling was done by both the human annotator and the LLM (GPT3.5), we were able to compare the cases of agreement and disagreement for both television programs, EenVandaag and Niewsuur.

In the case of EenVandaag, the agreement between annotator and GPT-3.5 is 483 out of 1,000 items. This corresponds to an accuracy of 48.3%. Of those cases of agreement, 303 are human interest, 162 are conflict, 16 are economic, 1 is morality and 1 is attribution of responsibility. The confusion matrix is shown in Figure 24.

In the case of Niewwsuur, the agreement between annotator and GPT-3.5 is 387 out of 1,000 cases. This corresponds to an accuracy of 38.7%. Of these 387 cases, 197 are classified as human interest, 173 as conflict, and 17 as economic. Figure 25 shows the confusion matrix.

The results were then analyzed by 3 experts in frame analysis, as well as the person who made the annotation, to provide an assessment of the results and to make hypotheses about the reasons

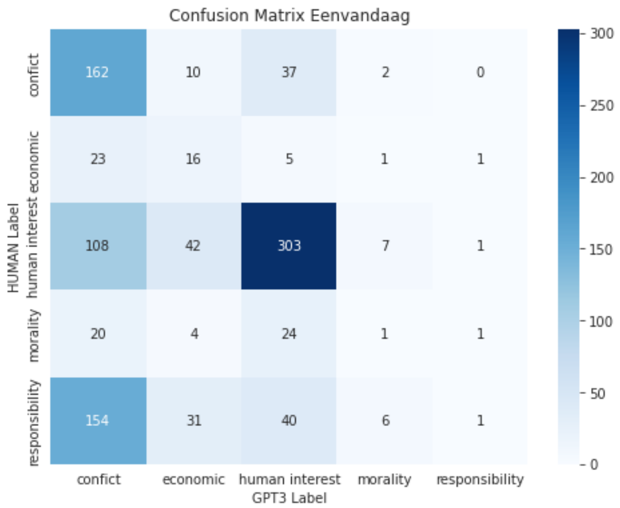



*Figure 24. Agreement between human annotator and GPT-3.5 on classification of EenVandaag transcripts into five categories: Conflict, Economic, Human Interest, Morality and Responsibility.*

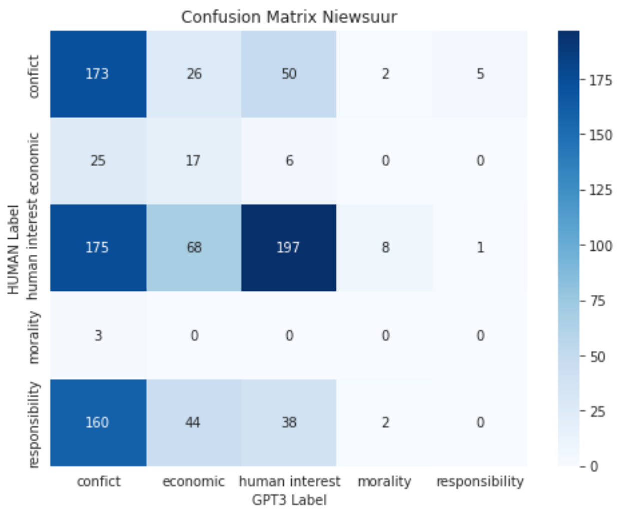



*Figure 25. Agreement between human annotator and GPT-3.5 on classification of Niewwsuur transcripts into five categories: Conflict, Economic, Human Interest, Morality and Responsibility.*

behind the results. In summary, all the experts offered insights into why GPT-3.5 might confuse human interest and attribution of responsibility labels with the conflict label. They highlighted the interconnected nature of these frames and the potential co-occurrence of elements in news stories as contributing factors to the disagreement. Additionally, they mentioned the importance of clear framing cues to avoid such confusion. Additionally, the annotator's views were centered on the idea that the nuances within the content might be challenging for the machine to distinguish accurately, resulting in the misclassification of frames like responsibility, conflict, and human interest.

In addition, we asked the experts and the annotator to provide alternative frame types for future research. The experts and annotator provided varied perspectives on alternative frames they considered appropriate. One expert endorsed Semetko and Valkenburg's typology [169], recognizing the existence of five frames but emphasized that journalists choose frames based on specific issues rather than a fixed set. This expert suggested the "politics-are-responsible frame," focusing on the role of administrators, governments, and politicians in causing or solving issues, and noted that this frame reflects left-wing and right-wing perspectives beyond the traditional conflict frame. They also identified the "system frame," which addresses structures responsible for issues, such as the invisible hand in economics, and is often used to deflect individual responsibility, especially in situations like the banking crisis. Additionally, the expert mentioned the "people frame," referring to general public opinion or desires but cautioned against viewing "the people" as a single entity, acknowledging diverse perspectives within the population.

Another expert suggested adding more positive frames, such as reconciliation instead of conflict, and reframing human interest to emphasize vulnerability, aligning it with morality and responsibility. They noted that human interest, economic, and conflict frames serve as overarching themes, while morality and responsibility are more implicit or sentimental. Furthermore, the annotator found the morality frame difficult to identify due to its religious connotations, and rarely used the economic frame despite its easy identification. She observed that some news pieces appeared neutral and suggested a "no frame" category for such cases, as existing frames sometimes did not fit well. Overall, the experts and annotator emphasized the need for flexibility and consideration of issue-specific frames, positive framing, and recognition of implicit themes in news stories.

### 9.1.4. Conclusion

In conclusion, the main contributions of our work are two-fold:

- We developed and validated a framework that leverages LLMs for automated frame detection in TV transcripts. A dataset was created with transcripts of two current affairs programs on Dutch television. These TV programs were classified by GPT-3.5 using a prompt composed by the definition of different frame types, the transcription item, and finally asking the model to choose the predominant frame based on the definitions.

- We created an annotation system based on Google Forms where an expert read the transcription to be annotated as well as the frame definitions, and answered a series of questions that justified their answer. After that annotation, we analyzed, together with experts in media and frame analysis, the results obtained from the agreement/disagreement between human and machine, and showed that while human and machine agreement is reasonable, there is clearly still room for improvement in this emerging research topic.

### 9.1.5. Relevant publications


- D. Alonso del Barrio, M. Tiel, and D. Gatica-Perez. Human Interest or Confict? Leveraging LLMs for Automated Framing Analysis in TV Shows. In Proc. ACM Int. Conference on

Interactive Media Experiences (IMX), Stockholm, Jun. 2024. [170]. Available at https://zenodo.org/records/11616987.

#### 9.1.6. Relevant software/datasets/other outcomes

- A dataset of labeled transcripts with respect to frames, derived from the two TV Dutch Programs dataset, was produced by NISV. However, we currently do not have authorization to make the data publicly available.

#### 9.1.7. Relevance to AI4Media use cases and media industry applications

This line of research has been carried out jointly between IDIAP and NISV for UC4: AI for Social Sciences and Humanities, since the automatic identification of frames by LLMs is a line of research to facilitate the work of journalists or social science professionals working with this frame concept.

### 9.2. Local News Analysis Based on Perspectives from the Migrant Community: The Case of Lausanne, Switzerland

**Contributing partners:** IDIAP

#### 9.2.1. Introduction

Media studies show that news consumption habits differ by demographic group, where factors such as age, socioeconomic status, and educational background indicate what media sources and platforms a person turns to and with what frequency.

This mixed-methods research aimed at studying the relationship that the migrant community has with the hyperlocal media environment of a city in French-speaking Switzerland (Lausanne). To our knowledge, relatively little research has studied migrant communities' news consumption patterns in Switzerland, and thus we contribute to fill this gap for a demographic that has been understudied. It is also relevant in the context of Lausanne, a city where over 43% of its residents do not have a Swiss passport [171].

#### 9.2.2. Methodology

Our methodology followed two main stages: On one hand, the development of two focus groups where we were able to discuss with migrants in the city of Lausanne and get their opinion on news consumption; on the other hand, the collection of a dataset with articles from a local newspaper as well as the application of NLP techniques to complement the results obtained from the focus groups.

**Focus groups.** For the correct development of the focus groups, we defined the structure and the target population of the focus group, where for the first focus group, we invited migrant women with at least three years living in Switzerland and who spoke French, while in the second focus group the population was made up of migrant men with more than three years living in Switzerland and who spoke French. Likewise, we defined the logistics such as the place, date and time, as well as the recording devices and the material used for the correct execution of the focus groups. These were the main sections of the discussions in the focus groups:

- **News Consumption Habits.** We posed questions about the news consumption habits of each participant. Namely, we asked what media sources they consumed and at what frequency, when did they began consuming Swiss media (after they arrived in Switzerland), and if they specifically interacted with local media outlets.

- **Feelings Associated with News Consumption.** This section accounted for the bulk of the discussion and gave participants an opportunity to talk about what their preferences about the media they consume. We asked if certain news sources are more or less trustworthy to them and why, and what content they wish they could see more or less of.

- **Hands-On Annotation of News Articles.** Along with the discussion, we prepared a hands-on activity for the participants. We chose six articles recently published covering a range of topics, which had varying article lengths, and that we believed could provoke different emotions, positive and negative alike. Each participant would be given a copy of each article and had to rank each article (based on the title and skimming the article alone) in order of interest or relevance to them. After the ranking, each participant would read the article they selected as most interesting or relevant to them, and annotate key phrases or content that was of particular importance. Room for notes was allotted on each printed article.

- **Discussion of Hands-On Activity.** The final discussion section allowed participants to share their thought process in completing the hands-on activity, which elicited additional insights.

**Technical Methods.** Our work also analyzed a collection of digital news articles. The section first outlines the methods used to collect data, and then the technical pipeline applied to the data. The pipeline consisted of five main NLP tasks: named entity recognition, topic modeling, information retrieval, sentiment analysis, and text readability analysis. Some of the methods described in this section were motivated by findings that arose from the focus group, as the idea was that the group discussions would guide our technical work.

- **Data.** To construct our article dataset, we extracted the title, subtitle, author, publication date, content, url, category, and tag of each article using Beautiful Soup, a Python library. The category and tag refer to the organizational method used to group content on the website. There are four categories and 22 tags, and each article is assigned exactly one of each, with each tag belonging to a specific category. The final dataset includes a total of 2,666 articles.

- **Named Entity Recognition.** Named entity recognition (NER) concerns the task of extracting and categorizing named entities from a text, such as locations, people, and organizations. As we specifically analyzed a local newspaper, it was relevant to see just how local the content described in the articles was, by checking which entities were mentioned, and if there was any kind of prevalence in entities outside the sphere and scope of Lausanne.

- **Topic Modeling.** Although the newspaper already has an organizational structure that assigns each article a category and tag, they are not necessarily indicative of topics discussed in the articles. Thus, we decided to perform topic detection on the article collection to find groupings that are more indicative of content. As a pre-processing step, we removed all articles related to film reviews. Since the content of the film was discussed in its review, it could influence the detected topics even though the film content is not actual article content. This reduced our dataset to 2,436 articles.

- **Information Retrieval.** Discussions in the focus group indicated that there were some specific topics that participants wished they could see more of in the news, and some content that they did not enjoy. Thus, in our technical analysis we decided to include a module on information retrieval, i.e., the process of collecting the most relevant documents in our corpus for a particular query. This allows us to query the specific content mentioned by participants in an attempt to corroborate that what they believe is lacking was indeed missing, or to retrieve documents containing content that the focus group participants specifically disliked.

- **Sentiment Analysis.** Participants in the focus group disclosed that they perceive news as being overall more negative than positive, which could result in emotional distress when interacting with the media. Some reflected that what they choose to read depends on the time-of-day and their state of mind, because starting or ending the day with negative news can weigh too heavy on them afterwards. To verify whether the newspaper is among the media that leans towards negative content we used sentiment analysis to study polarity evoked in text. We were not concerned with detecting the exact emotions of each article published, but rather detecting whether each article contained negative, positive, or neutral content to see how it could affect its readership. Vader (a sentiment analysis tool with a package installation in Python) was used to extract polarity scores [172].
- **Text Readability.** Text readability refers to the complexity of a document's language, i.e., it determines how easy or difficult it may be to read and comprehend a document. Determining this may consist of an analysis of lexical, semantic, and syntactic features of a document. The text readability of articles published became of interest after some focus group participants shared that reading news in French was a part of their language learning journey. Some participants also shared that, in general, they may skip over more difficult-to-read content in favor of text that appears easier, since their French may not yet be at a level to comfortably read any text. Although the perspectives shared in focus groups are not always generalizable, we hypothesize that many migrants who do not speak French as their first language likely share this experience. Hence, we found this step in the pipeline to be a contribution to understanding this specific community's needs.

### 9.2.3. Experiments

As mentioned in the methodology, several NLP techniques were applied according to what the focus groups participants discussed about. In this subsection, we summarize the information retrieval and sentiment analysis results.

**Information retrieval results**. The module was to retrieve specific content mentioned in the focus groups by participants. The first query was one related to 'nature', as one participant shared that not enough content related to teaching basic topics about nature appeared in the news anymore. The second query was regarding one specific concern expressed in the women's focus group, namely that some content criticizes women and/or feminism. To create the queries, we gathered a list of words that were related to the two topics. These are shown in Table 30.

| | Subject | Query | Results |
|---|---|---|---|
| Query 1 | Nature | 'nature, graine, plante, jardin, fleur' | 339 |
| Query 2 | Feminism | 'femmes, sexisme, féminisme, égalité' | 2,428 |

*Table 30. Queries used for information retrieval along with number of results yielding non-zero cosine similarity.*

The results column in the table indicates how many results from the corpus of articles yielded a non-zero cosine similarity. From here, we examined the top 10 results of each query. For query 1, from the titles alone, it was difficult to determine what the exact content of the articles would contain as they are ambiguous. However, after close reading of the texts associated with the titles, it was clear that none of the articles actually teach anything about nature. Instead, there are some articles about cooking, some about places to visit that are in nature, and some about books. Nevertheless, no articles actually provided the kind of content that our focus group participant was looking for.

Regarding the top 10 articles retrieved using query 2, the content in some of these articles is indeed critical of the feminist movement. For example, article ranked 2 in the top-10 list criticizes women who claim to be feminists for not speaking out enough about the struggles of women in Afghanistan and Iran; while article ranked 7 in the top-10 list critiques the 2019 women's strike in Switzerland. In this case, the information retrieval technique produced some results very related to the query, even if not all of them were highly relevant results.

**Sentiment Analysis Results**. Figure 26 shows the distribution of positive and negative scores at each level: article content, article subtitles, and article titles. Surprisingly, for each category, the scores had a higher magnitude for the positive than the negative polarity measures. It is also notable that the titles of the articles had higher magnitudes in both positive and negative sentiment than the article content itself. Perhaps this is due to the fact that titles try to capture reader attention which is easier to do when there is either positive or negative sentiment than when the sentiment is neutral. Figure 27 corroborates the higher polarity of titles and subtitles as compared to full article content, by showing the proportion of articles with each polarity score that come from titles, subtitles, and content.

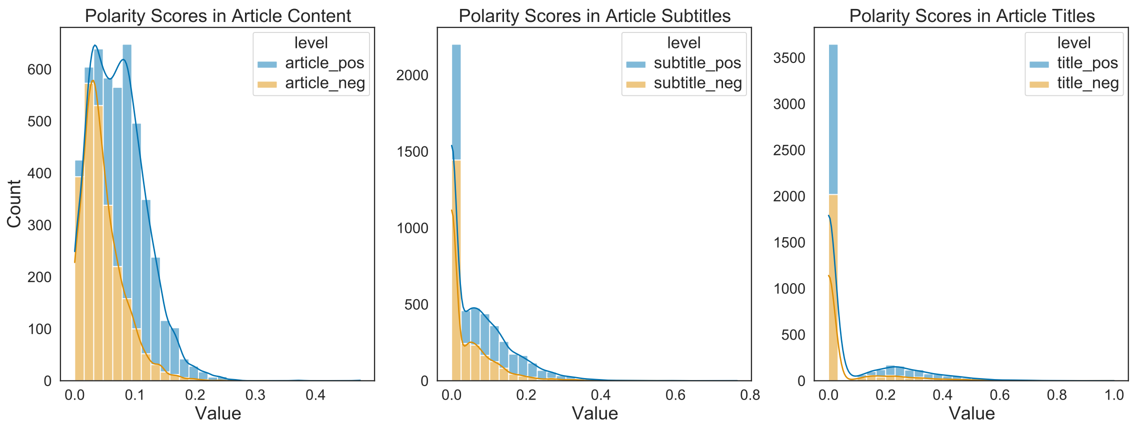



*Figure 26. Distribution of polarity scores on a logarithmic scale by full article text, subtitles, and title respectively. Surprisingly, the positive sentiment scores overall have a higher magnitude at every level than the negative scores.*

### 9.2.4. Conclusion

In conclusion, the main contributions of our mixed-methods work are the following:

- We conducted two focus groups with local migrants in Lausanne, Switzerland (one women group, and men group) to identify some of the information needs and information access practices of the local migrant community, and use this as a guide for analyzing online news articles of a hyper-local newspaper, to understand if the actual content meets the needs of this specific group of readers.
- We reported qualitative insights collected from the focus groups to present the needs and perceptions of the participants, as well as quantitative results found using NLP techniques applied to online articles.

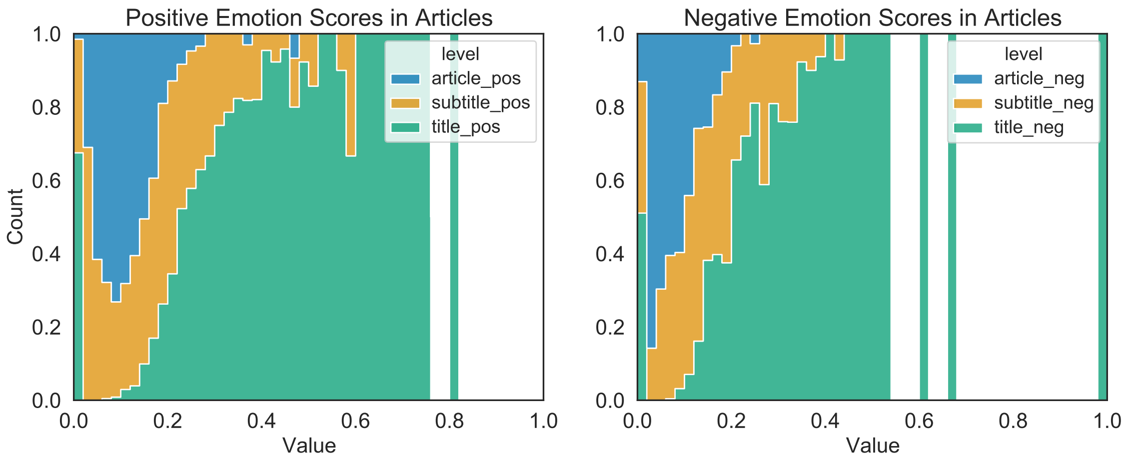



*Figure 27. Share of each polarity score belonging to the article content, subtitles, and titles. Higher polarity scores are found for the titles and subtitles than for the full article itself. This is true for both positive and negative sentiment.*

We believe that our human-centered approach is promising to identify positive uses of NLP to support community access to local information.

### 9.2.5. Relevant publications

- A paper reporting this work is under preparation.

### 9.2.6. Relevant software/datasets/other outcomes

- This work generated a dataset of online articles from a local Swiss newspaper, as well as the transcripts of two focus groups. The data is not publicly available.

### 9.2.7. Relevance to AI4Media use cases and media industry applications

This work used NLP techniques to analyze local news, and it is thus relevant for UC4: AI for Social Sciences and Humanities. It is particularly useful to understand the specificities of local news outlets compared with national ones in terms of topics, events, represented persons, etc.

## 9.3. Understanding Local News in German-Speaking Switzerland: the Case of Zürich

**Contributing partner:** IDIAP

### 9.3.1. Introduction

Digitalization has transformed the way news are disseminated and consumed [173, 174]. In recent years, news outlets have struggled to keep up with the rapidly changing media landscape, with the advent of social media platforms and the proliferation of online news sources [174]. Understanding the dynamics of local news is critical, especially in the context of political, social, and economic issues that affect local communities [175, 176, 177].

The use of NLP methods for analyzing news content and sentiment [178, 179] is increasing. However, there's a noticeable research gap, in terms of understanding main trends, for the case of Swiss local newspapers, particularly those in the German-speaking region. Our research contributes to this domain.

### 9.3.2. Methodology

We collected and curated a dataset of local news from Zürich, and conducted a series of text analyses to understand how the news were presented on a weekly basis in two districts of the Canton of Zürich. The dataset consists of around 14,500 full articles from two local newspapers *Tagblatt der Stadt Zürich* and *Winterthurer Zeitung* in the canton of Zürich, with all the content in German.

We implemented an analysis pipeline that included topic modeling; named entity recognition; and sentiment analysis at three levels (article, headline and sentence). First, topic modeling was executed using BERTopic, followed by a manual analysis of the topic categories to assign a label to each category. Second, named entity recognition was executed using the spaCy library, which distinguishes between person, location, organization, and miscellaneous entities. The NER analysis was conducted on both the complete text bodies of the data sets and their corresponding topics identified by topic modeling. Finally, sentiment analysis was executed using a sparkNLP library for German texts as well as 3 other models from huggingface, as a comparison to the performance of the first model. Furthermore, a qualitative analysis of 100 headlines for each newspaper was conducted to validate the accuracy of the findings.

### 9.3.3. Experiments

The following are the most salient features of the results obtained in each of the NLP techniques used for each of the newspapers.

**Topic modeling results.** In the case of *Tagblatt der Stadt Zürich*, the three detected topics with the highest frequency across all document categories were: "Gastronomy in Zurich" with 2,933 articles, "Christmas in Zurich" with 576 articles and "Public Transport in Zurich" with 2,232 articles. The articles and topic labels were chosen based on the following keywords returned by the model for each topic (English translations are provided in parentheses for some words):

*Public Transport*: 'escher', 'alfred escher', 'alfred', 'bahnhofstrasse', 'osiris', 'anna', 'strasse' (street), 'bus', 'Herkules' (hercules), 'bahn' (train).

*Christmas*: 'weihnachten' (Christmas), 'nikol', 'cachelin', 'dezember' (December), 'moll', 'glühwein' (mulled wine), 'roman', 'papa moll' (fictitious father in a famous Swiss comic series), 'heiligabend' (Christmas Eve), 'papa' (dad).

*Gastronomy*: 'bar', 'drinks', 'gast' (guest), 'geräuschpegel' (noise level), 'getrunken' (drank), 'bedienung' (service), 'waldhaus' (forest cabin), 'hotel', 'flims', 'rechnung' (check).

The prominence of "Public Transport" and "Gastronomy" as top topics that significantly impact the daily lives of the Zürich population is to be expected. These subjects inherently reflect essential aspects of urban life.

In the case of *Winterthurer Zeitung*, the three detected topics with the highest frequency are:"Music" with 5,498 articles, "Art" with 6,878 articles, and "Entertainment" with 2,625 articles. The articles and topic labels were chosen based on the following keywords returned by the model for each topic:

*Art*: 'ausstellung', 'kunst', 'künstler', 'werk', 'museum', 'bild', 'sammlung', 'zeigen', 'winterthur', 'jungkunst'

*Music*: 'konzert', 'band', 'musik', 'musikalisch', 'musiker', 'spielen', 'orchester', 'song', '30', '20'

*Entertainment*: 'theater', 'stück', 'sommertheater', 'komödie', 'bühne', 'publikum', 'casinotheater', 'geschichte', 'premiere', 'spielen'

Interestingly, none of the chosen top topics from *Winterthurer Zeitung* ("Music", "Art" and "Entertainment") were present in the topic modeling results of *Tagblatt der Stadt Zürich*. This disparity in topics between the two newspapers could suggest that the newspapers are catering to distinct interests and preferences.

**NER results.** For the articles of *Tagblatt der Stadt Zürich*, a total of 37,513 different entities were found by the model, 12,036 in the category "location", 10,701 in "person", 5,503 in "organization", and 9,273 in "miscellaneous". The top 10 entities for each category are shown in Table 31. It is important to clarify that a single entity may be placed in different categories depending on the context of the sentence in which it appears. For instance, the term "Zürcher" could be interpreted as referring to a person from Zürich, or used as an adjective to describe an organization based in Zürich (e.g., "das Zürcher Unternehmen" - the Zürich company).

| Location | Count | Person | Count | Organisation | Count | Miscellaneous | Count |
|---|---|---|---|---|---|---|---|
| Zoo Zürich | 501 | Sohn | 271 | SVP | 306 | Zeit | 1064 |
| Franken | 417 | Corine Mauch | 166 | Kirche | 259 | Tag | 267 |
| Deutschland | 409 | Corona | 135 | VBZ | 253 | Box | 233 |
| Europa | 320 | Wieso | 133 | FCZ | 228 | Internet | 216 |
| Hause | 278 | Richard Wolff | 111 | SP | 212 | deutschen | 211 |
| Bund | 269 | Samichlaus | 109 | FDP | 204 | Weihnachten | 209 |
| Bahnhofstrasse | 244 | Vaters | 94 | EU | 155 | sozialen | 207 |
| Bern | 238 | Mauro Tuena | 93 | Universität Zürich | 152 | Tages | 169 |
| Limmat | 220 | Nein | 84 | GC | 117 | Facebook | 157 |
| Oerlikon | 206 | Josef | 82 | ÖV | 113 | Virus | 151 |

*Table 31. Ten most frequent entities for each category in Tagblatt der Stadt Zürich.*

For the articles of the *Winterthurer Zeitung*, a total of 60,099 entities were identified, comprising 16,406 locations, 17,202 persons, 11,592 organizations and 15,899 miscellaneous entities. Table 32 shows the ten most frequent terms in each category.

| Location | Count | Person | Count | Organisation | Count | Miscellaneous | Count |
|---|---|---|---|---|---|---|---|
| Altstadt | 434 | Winterthur | 694 | Winterthur | 645 | Zeit | 715 |
| Franken | 329 | Michael Künzle | 169 | Winterthurer Zeitung | 345 | Internet | 203 |
| Töss | 307 | Sohn | 127 | SVP | 336 | Tag | 186 |
| Effretikon | 303 | Josef Lisibach | 107 | FDP | 280 | Weihnachten | 147 |
| Seuzach | 302 | Petrus | 93 | SP | 267 | sozialen | 138 |
| Wülflingen | 285 | Barbara Günthard-Maier | 89 | FC Winterthur | 179 | soziale | 132 |
| Andelfingen | 277 | Stefan Fritschi | 66 | Kantonspolizei Zürich | 179 | Anschluss | 117 |
| Eulachstadt | 265 | Peter Jiricek | 60 | FCW | 177 | Dieses Jahr | 103 |
| Illnau-Effretikon | 258 | Lui Eigenmann | 53 | Partei | 168 | Stadthausstrasse | 101 |
| Neftenbach | 229 | Christa Meier | 532 | EHC Winterthur | 163 | Millionen Franken | 93 |

*Table 32. Ten most frequent entities for each category in Winterthurer Zeitung.*

The results shown in in Tables 31 and 32 mostly correspond to real, highly local entities. Please note that for the "location" and "miscellaneous" categories, terms that included "Zürich", "Schweiz", "Stadt" or "Winterthur" were excluded from the Tables. It is worth noting that the term "Franken" is likely mistakenly categorized as a location entity for both newspapers, whereas it should be correctly listed as a miscellaneous entity representing the Swiss currency (Swiss Franc.)

As a second example of error, "Corona" (the German term for Covid-19 in the media) is categorized as a person in Table 31. Finally, as another interesting result by comparing the most frequent locations mentioned in Tables 31 and 32, it is evident that the locations mentioned in *Winterthurer Zeitung* have a more local focus and are predominantly based around the Winterthurer municipality in the canton of Zürich.

**Sentiment analysis.** Table 33 shows a summary of the sentiment analysis results. M1-M4 correspond to different computational models for sentiment analysis. For each model, the absolute numbers and the percentages are shown (green: positive sentiment; red: negative sentiment; blue: neutral sentiment). At the full article level, the majority of articles were classified as neutral. However, examining the headlines for *Tagblatt der Stadt Zürich*, a majority was categorized as negative. The qualitative analysis revealed that many articles had the title "Werbung" (ad), which the model somehow associated with a negative sentiment. Otherwise, the model's results and the results of the qualitative analysis differed by 22% for *Tagblatt der Stadt Zürich* and by 36% for *Winterthurer Zeitung*.

| **Document Type** | **Tagblatt der Stadt Zürich** | | | | **Winterthurer Zeitung** | |
|---|---|---|---|---|---|---|
| | M1 | M2 | M3 | M4 | M1 | M2 |
| Full Article Body | 3527 (58.92%)<br>2459 (41.08%) | 399 (6.67%)<br>906 (15.14%)<br>4681 (78.2%) | 425 (7.1%)<br>1110 (18.54%)<br>4451 (74.36%) | 213 (3.56%)<br>657 (10.98%)<br>5116 (85.47%) | 6384 (74.54%)<br>2181 (25.46%) | 614 (7.17%)<br>817 (9.54%)<br>7143 (83.4%) |
| Titles | | 409 (6.84%)<br>4026 (67.29%)<br>1548 (25.87%) | | | | 2100 (24.52%)<br>2469 (28.83%)<br>3962 (46.26%) |
| Without Ads | | 203 (7.22%)<br>458 (16.29%)<br>2151 (76.49%) | | | | |
| Ads | | 197 (6.2%)<br>448 (14.1%)<br>2533 (79.7%) | | | | |

*Table 33. Results of sentiment analysis for various levels of text as units. Values are denoted both as absolute figures and as relative percentages within parentheses. Green = positive, red = negative, blue = neutral.*

### 9.3.4. Conclusion

In summary, the main contributions of this work were two-fold:

- We collected and curated datasets from two local German-speaking newspapers, which gave the opportunity to understand hyper-local news content for a specific European region.
- Using the datasets, we carried out a series of analyses with NLP techniques including topic modeling, named entity recognition and sentiment analysis. These allowed to identify key themes and topics in the news as well as to analyze sentiment. The results clearly show the locality of the information produced and circulated by these media outlets.

### 9.3.5. Relevant publications

- A Bachelor's thesis report by the Junior Fellow working on this topic was written: T. Nikray, Understanding Zurich Local News: A Data Science Approach, ETH, 2023.

### 9.3.6. Relevant software/datasets/other outcomes

- This work generated a dataset of around 14,500 online articles in German from two local newspapers in the canton of Zürich. The data is not publicly available.

### 9.3.7. Relevance to AI4Media use cases and media industry applications

This project uses NLP techniques to analyze local news and it is relevant for UC4: AI for Social Sciences and Humanities. By examining unique features of local news, it aims to understand its value compared to national news, and potentially reveal how media covers local stories and biases.

## 9.4. Local News Characterization: the Case of the Swiss Romandy Region

**Contributing partner:** IDIAP

### 9.4.1. Introduction

Local media occupy a special place within the media ecosystem. Historically, they have long represented regional differences and the divergent interests of readers from one city to another, thus becoming markers of geographical distinction [180].

In our era, local media have retained this specificity of representing the communities they serve, addressing societal needs often neglected by mass media. Local press is thus perceived as a potential actor in addressing the crisis of trust that national and international media are experiencing. Local press plays a major role in maintaining the link between citizen communities and democratic institutions. Considered as "good neighbors," local newspapers fulfill the role of disseminating community life [181].

However, the sustainability of this ecosystem is being challenged worldwide, particularly due to the precarious economic model in the face of digitization and changing reader habits. Newsrooms are struggling to maintain quality journalism, which is necessary for democratic functioning, while being economically threatened by increased competition. In this difficult context, Switzerland had remained relatively stable over the last decades despite changes in business plans (advertising), but this stability has unfortunately been questioned in the past few years [182].

### 9.4.2. Methodology

In the context of French-speaking Swiss (a.k.a. Romandy) news, we conducted research to answer the following research questions:

- What automated analysis of news articles can extract significant knowledge about local issues?
- How can we link qualitative analyses and quantitative measures to provide a substantiated view of the dynamics of local media?
- What differences can be observed between local press and national media in their coverage?

To answer these questions, we implemented a comprehensive analysis framework combining quantitative measure indicators and qualitative analyses to understand the specifics of local journalism.

#### 9.4.2.1. Quantitative Measures

The first part of our framework concerns quantitative measures. We implemented a broad coverage of different elements of the articles, including:

- Contextual elements: publication frequency, journalist contributions, tags associated with articles, and article length.
- Textual features: extraction of quotes, acronyms as proxies for mentioned organizations, named entities, and measures of lexical richness and readability of the articles.

These initial measures allowed us to evaluate the quality of the dataset and then support the comparative analysis between the local and national scales. Finally, we implemented a topic modeling pipeline to organize the articles into thematic groups, allowing us to target topics that represent significant differentiation for scale comparison.

#### 9.4.2.2. Qualitative Analysis

The second part corresponds to the qualitative analysis of the articles, following the iterative process of Grounded Theory to extract significant elements from the selected articles [183]. We annotated each article according to the three-dimensional model of Fairclough [184]:

- Textual level: focusing on the linguistic features of the text, such as vocabulary, grammar, and structure.
- Discourse level: examining how the text is produced, distributed, and consumed, considering the context and the processes involved in its creation and interpretation.
- Cultural practices level: exploring the broader social and cultural context, analyzing how the text relates to societal structures, power relations, and ideologies.

This annotation allows us to link the elements extracted from journalistic coverage with those that contribute to the reader's understanding of the information. The iterative process allowed us to extract a theory regarding the specificity of local media treatment compared to larger scales.

#### 9.4.2.3. Data

We curated two datasets composed of articles from Swiss newspapers published in French between June 2019 and June 2022:

- Local Dataset: It includes 130,000 articles in French from three local titles: Arc Info, La Cote, and Le Nouvelliste. The articles were collected via direct RSS access with access to the articles and associated metadata.
- National Dataset: It includes 200,000 articles in French from 20 min, 24 heures, and La Tribune De Genève, collected from the database constituted by Common Crawl, CCNews [185].

After examining the data, it was clear that the quality of both datasets was very disparate and required significant processing to make them suitable for computational analyses. We corrected collection errors and removed artifacts and boilerplates, which were hindering the performance and the interpretation of the analysis.

### 9.4.3. Experiments

In this section, we present the framework that allowed us to extract relevant elements of differentiation, focusing on the most important results.

#### 9.4.3.1. Comparison of Newspapers

First, we compared the different newspapers in our dataset using quantitative measures implemented in Python. Regarding contextual data, we observed very few interpretable differences between the newspapers, each having its own specificities. However, some limitations emerged, particularly due to access to sources during data collection for the CCNews bot. These quantitative measures were notably useful for evaluating the quality of the dataset and correctly targeting relevant articles for our analysis.

An interesting phenomenon can be observed concerning the length of the articles: there is a correlation with the period of the 2021 Olympic Games, where the average length is shorter because articles reporting sport competition results are shorter.

#### 9.4.3.2. Textual Data Analysis

The analysis of textual data revealed significant differences between the newspapers. A first differentiation appeared during the extraction of French quotes using regex analysis. Local newspapers contain a higher number of quotes per article, and these quotes are also longer as seen in Table 34.

| Metrics - mean (std) | Local | Arc Info | La Cote | Le Nouvelliste |
|---|---|---|---|---|
| number of quotes | 5.71 (6.13) | 5.774 (6.47) | 4.998 (5.03) | 6.045 (6.32) |
| length of quotes | 84 (84.9) | 70.93 (83.4) | 90.99 (85.9) | 92.31 (84.3) |
| | National | 24 heures | TDG | 20 min |
| number of quotes | 4.33 (5.52) | 4.465 (5.94) | 3.945 (5.38) | 4.436 (5.39) |
| length of quotes | 50.1 (66.8) | 56.24 (76.3) | 49 (71.1) | 48.04 (60) |

*Table 34. Average number of quotations and their average length in the articles in the newspaper corpus.*

For the acronyms and named entities extracted, we used a pre-trained NER model that has the ability to be a generalist, GLiNER, meaning that we can define ourselves the categories to extract [186]. We chose four specific types relevant for our analysis: Persons, Organizations, Locations, and Events. The extracted entities and acronyms show a differentiation between scales, with national and international figures in the national press, and a significant sample of local figures in the local press.

#### 9.4.3.3. Lexical Richness and Readability

We implemented several measures of lexical richness and readability of the articles:

- Lexical Richness : Type-Token ratio, Herdan's Index, Dugast's Uber index, Yule's K characteristic, Measure of Textual Lexical Diversity (MTLD), Hypergeometric Distribution Diversity (HDD), and Moving Average Type-Token Ratio (MATTR)

- Readability : Flesch Reading Ease (FRE), Gunning-Fog Index (GFI), Coleman-Liau Index (CLI), Flesch-Kincaid Grade Level (F-K GL), and Automated Readability Index (ARI).

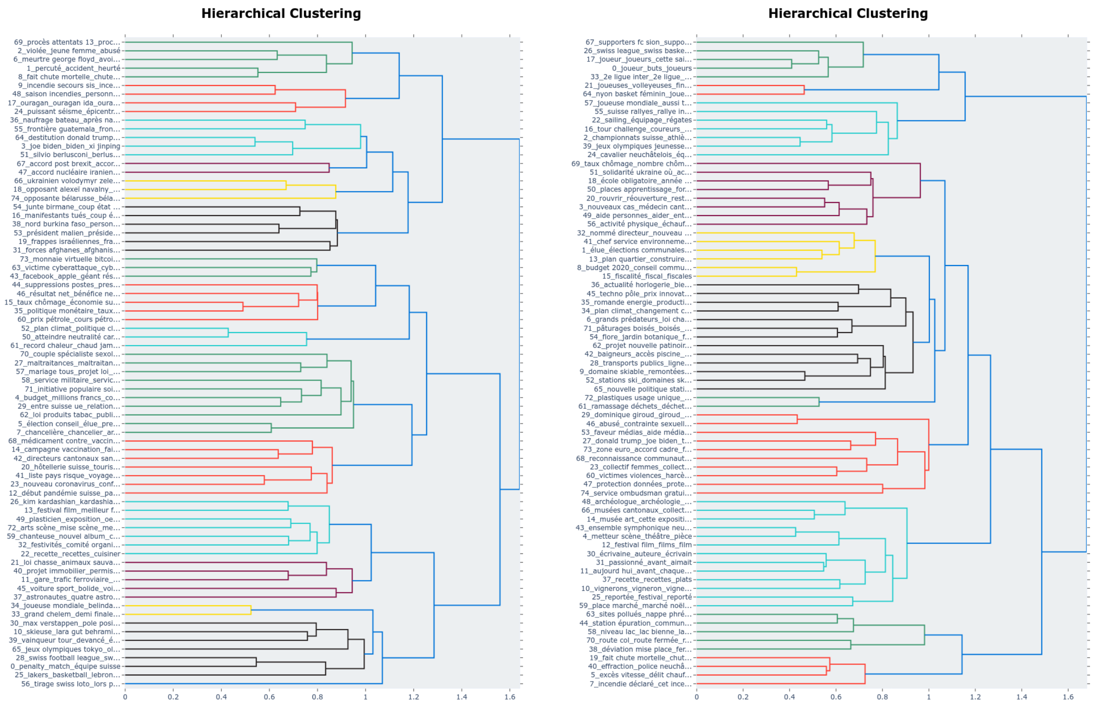
These measures did not show significant differences at the dataset level, but they were useful for comparing topics. The overall readability measures showed a high-school level, which is quite high compared to the broad inclusion assumption of the press.

#### 9.4.3.4. Topic Analysis

To analyze the distribution of articles and infer topics, we used a Topic Modeling pipeline with BERTopic [187]. The pipeline consists of several key steps, each crucial for identifying and interpreting topics:

1. **Embedding Generation**: We utilized a pre-trained french BERT model tailored to our use case to generate high-quality embeddings from the article texts. These embeddings capture the semantic nuances of the documents, providing a dense and informative representation of their content.
2. **Dimensionality Reduction**: We applied UMAP to reduce the dimensionality of the generated embeddings. UMAP is a non-linear dimensionality reduction algorithm that preserves the local and global structure of the data, facilitating visualization and subsequent processing. This step is essential for making the data more manageable and improving the performance of clustering algorithms.
3. **Clustering**: We used HDBSCAN to cluster the articles based on their reduced embeddings. HDBSCAN is a density-based clustering algorithm that identifies clusters of varying shapes and sizes while effectively handling noise in the data. This allows us to group articles into coherent clusters representing distinct topics.
4. **Topic Interpretation**: For topic interpretation, we employed KeyBERT, a BERT-based method for extracting the most representative keywords for each cluster. KeyBERT uses BERT embeddings to identify the most relevant and informative terms, facilitating the interpretation and labeling of the identified topics.

Figure 28 shows the hierarchical distribution of the obtained clusters after applying the pipeline to both datasets.




*(a) National Articles* *(b) Local Articles*

*Figure 28. Hierarchies of the inference of the BERTopic model on the Local and National datasets.*

### 9.4.3.5. Qualitative Analysis

Following the topic modeling, we selected articles within similar topics shared by the two scales of media. We chose six main themes: Sports, Environment, Animals & Climate, Politics, Business & Economy, Cultural Events, and Justice & *Faits Divers*. These themes were chosen for their observability in both distributions and their known differences in treatment between media scales.

We conducted a stratified sampling of 100 articles for each theme, analyzed according to the iteration of grounded theory to extract a generalization. The main lines of the extracted theory are articulated around three main axes:

**1. Proportionality of temporal and geographical scales.** This observed phenomenon is described by a proportionality of institutions, organizations, and geographical scales with the distribution scale of newspapers (expected), but also on the mentioned temporal scales. In local journalism, sports results reporting is limited to the match played or the one that precedes/follows, while national newspapers tend to place the result in the context of a season or competition. Similarly, for cultural or festive events, the mentioned dates are very close for the local press, while the national press talks about planned events further in advance. This can be explained by the proximity relationship with readers, which allows speaking to people who regularly follow local coverage, limiting the need for context, or playing on immediate interest.

**2. Promotion of direct consumption versus advertising campaign.** Many articles in the analyzed local dataset are related to selling something, even bordering on advertising. Thus, the boundary between articles of interest for regional events and figures, and advertisements for local products and events is very thin. Local journalism even details the description of the service or consumption object. It is worth noting that there is still an effort to inform about these types of articles in addition to the advertisement itself. This can be seen especially for local artisans' products or tickets to nearby cultural and sports events. In contrast to this advertising, the analyzed national articles showed more integration within an advertising campaign rather than wanting to sell a specific object. Thus, we find interviews with actors to promote an upcoming film, or discussions about companies' strategies to gain market share. We observed more long-term storytelling rather than immediate consumption.

**3. Pedagogy and popularization of technicality in contrast with direct communication.** This theory highlights the desire to address a larger number of irregular readers on the part of national newspapers. In the analyzed articles, we saw that a lot of jargon is used for context setting with very technical vocabulary, especially on scientific, economic, or justice topics. However, this is counterbalanced by an effort to popularize to anchor these topics as serious, with reported expert opinions that support the serious treatment of these topics. This contrasts with the local style on these topics, which remains focused on presenting the facts while limiting the use of technical terms, especially on biological topics. The focus is on the actors and what they say, with fewer expert opinions and more explanation. This can probably be explained by the scope of the reported topics, as national topics have more significant implications, precision seems to be more necessary than in articles of pedagogy on scientific topics.

The three presented theories are nevertheless not exhaustive of our observations but illustrative of unexpected elements.

### 9.4.4. Conclusion

This research has provided a comprehensive analysis of the Romandy local news ecosystem, highlighting the unique characteristics and challenges faced by local media in comparison to national outlets. By leveraging both quantitative and qualitative methodologies, we have been able to draw a few significant insights into the dynamics of local journalism.

- *Quantitative Analysis:* Our quantitative measures, including contextual elements, textual features, and topic modeling, have revealed distinct patterns in local versus national news coverage. Notably, local newspapers tend to include more and longer quotes, indicating a deeper engagement with local voices. The use of named entities and acronyms further differentiated the scales, with local media focusing more on regional figures and events.
- *Qualitative Analysis:* Through the iterative process of Grounded Theory, we identified three main axes of differentiation:
    1. *Proportionality of Temporal and Geographical Scales:* Local journalism tends to focus on immediate and nearby events, reflecting a closer relationship with the community. In contrast, national media provide broader temporal and geographical contexts.
    2. *Promotion of Direct Consumption vs. Advertising Campaigns:* Local articles often blur the lines between news and advertisements, promoting local products and events directly. National media, however, integrate advertisements within broader storytelling and long-term campaigns.

3. *Pedagogy and Popularization vs. Direct Communication:* National newspapers employ technical jargon and expert opinions to address a wider audience, while local media focus on straightforward communication, emphasizing the voices of local actors.

### 9.4.5. Relevant publications

An article is currently under preparation. The dataset (see below) is described in: V. Bros and D, Gatica-Perez, The Suisse Romande Local News Dataset, Idiap Technical Report, 2023.

### 9.4.6. Relevant software/datasets/other outcomes

The dataset has been made public at `https://zenodo.org/records/10256911`. The associated code will be made public once the paper is finalized.

### 9.4.7. Relevance to AI4Media use cases and media industry applications

This research implemented a framework of analysis for a human-centered subject, which is particularly relevant for the UC4: AI for Social Sciences and Humanities. The dataset can be used to fine-tune textual models for the understanding of local and national news content.

## 9.5. Media Event Temporal Analysis in Swiss News

**Contributing partners:** IDIAP, CEA

### 9.5.1. Introduction

The media coverage of a topic is influenced by its interaction with other concurrent events in the media sphere and civil society. Topics that are covered extensively and over a relatively long period undergo notable evolution. This research focused on this temporal evolution, with an application to various media topics suitable for this analysis.

Analyzing the temporal component is crucial for understanding the different phases of media treatment and the accompanying changes in public opinion. This can provide insights into the tipping points that might shift the perception of themes or topics [188]. In the current context of large-scale media analysis, this temporal component remains relatively underexplored and challenging, due to the complexity of the task and the need to obtain high-quality data for analysis (broad and continuous coverage of sources and topics, in particular) [189].

### 9.5.2. Methodology

We propose an analytical framework with an application to the Swiss media ecosystem. The choice of Swiss media was guided by our prior knowledge of this ecosystem and its specificity. The Swiss media landscape is mainly divided into three languages according to the distribution regions, allowing for a comparative analysis to discover distinctive elements. Moreover, the diversity of Swiss media provides broad media coverage from multiple angles, adding particular interest to this study.

The objective of this research is to examine the possibility of studying events covered by the media according to a temporal component, filling this often neglected aspect of analysis by large-scale AI methods. We have thus conducted this research to answer the following questions:

- How to systematize the temporal analysis of media events in the multilingual Swiss press?

- What patterns of temporal evolution can be extracted from the media coverage of thematic, singular, and recurring events in the Swiss press?

For this research, we collected a set of articles published in Swiss media from the Common Crawl database, CCNews [185]. We first retrieved all articles from domain names associated with Swiss media. However, after investigation, the data quality was rather disparate depending on the sources, due to the access difficulties that the CCNews bot may have encountered during its collection.

We first selected articles corresponding to the official languages of Switzerland: German, French, and Italian (we also considered Romansh, but the volume of data was too low to obtain relevant results and interpretations). Following this initial filtering, we found that some sources publishing articles in these languages had been collected unsatisfactorily. We then sorted the sources according to the quality of the collected articles, the frequency, and the quantity of data. Finally, we processed all the articles to eliminate obvious artifacts and boilerplates that could skew the analyses. The resulting dataset contains more than 1.7 million articles published between 2019 and 2022 in the three languages, with a distribution shown in Table 35.

| Language | Number of Articles | Number of Sources |
|---|---|---|
| German | 1,248,880 | 82 |
| French | 332,934 | 31 |
| Italian | 122,807 | 16 |
| Romansh | 8,002 | 3 |
| Total | 1,712,623 | 125 |

*Table 35. Description of the dataset by languages.*

Considering the chosen data, we needed to define the media events that a priori incorporate a phenomenon of temporal evolution, to be able to carry out this analysis, and particularly the identification of phases. We identified three types of concepts borrowed from media studies:

1. **Topics**: This term encompasses broad themes that span a long time, potentially including opinions and debates among stakeholders.

2. **Events**: We consider significant, one-time events whose media coverage can be characterized in phases. We could identify an evolution such as the phases of emergence, peak, decline, and end.

3. **Recurring Events**: These are associated with regular events (similar to journalistic "evergreens") with potentially a cyclical aspect in the treatment of the topic.

Based on these three definitions, we adapted our filtering methods to match their respective characteristics.

Once the articles were selected to analyze the coverage of an event in the media, we applied an analysis pipeline to extract significant elements of this media treatment. We defined three aspects analyzed by the framework:

- **The evolution of the importance of media coverage**, measured by the number of publications for each topic and the number of sources that covered it.

- **The dynamics of the event's stakeholders**, including their appearance date in the articles and the prevalence of each important figure for the topic.
- **The emotion and opinion associated with the topic and actors**, as well as the evolution of opinion regarding these figures.

For these three aspects, we identified the phases using a changepoint detection algorithm, which delimits the phases that we will relate to our knowledge of the chosen topics for the study.

#### 9.5.3. Experiments

As an initial step for our experiments, we filtered articles according to the three definitions of media events outlined earlier.

##### 9.5.3.1. Selection of Articles by Themes

For the theme-based definition, the goal was to group articles covering the same subject, ensuring they were semantically similar. We employed the BERTopic pipeline for this selection [187]. The pipeline is the same we described in the previous section, and is reproduced here for completeness purposes:

1. **Embedding Generation**: We utilized a pre-trained multilingual BERT model adapted to our use case to generate high-quality embeddings from the article texts. These embeddings capture the semantic nuances of the documents, providing a dense and informative representation of their content.
2. **Dimensionality Reduction**: We applied UMAP to reduce the dimensionality of the generated embeddings. UMAP is a non-linear dimensionality reduction algorithm that preserves the local and global structure of the data, facilitating visualization and subsequent processing. This step is essential for making the data more manageable and improving the performance of clustering algorithms.
3. **Clustering**: We used HDBSCAN to cluster the articles based on their reduced embeddings. HDBSCAN is a density-based clustering algorithm that identifies clusters of varying shapes and sizes while effectively handling noise in the data. This allows us to group articles into coherent clusters representing distinct topics.
4. **Topic Interpretation**: For topic interpretation, we employed KeyBERT, a BERT-based method for extracting the most representative keywords for each cluster. KeyBERT uses the previously computed embeddings to identify the most relevant and informative terms, facilitating the interpretation and labeling of the identified topics.

By combining these techniques, the BERTopic pipeline enabled us to group articles semantically and identify the main themes covered, meeting our need for theme-based filtering of media events. This pipeline was applied to each language, and we identified topics suitable for temporal analysis based on keyword coherence, topic relevance (sufficient and prolonged coverage), and the number of data points available for analysis. We selected topics such as space research, the British royal family, Hong Kong, cybersecurity, and international relations with North Korea as common subjects across the three languages. These topics were selected to ensure diversity and check the robustness of the method.

#### 9.5.3.2. Selection of Singular and Recurring Events

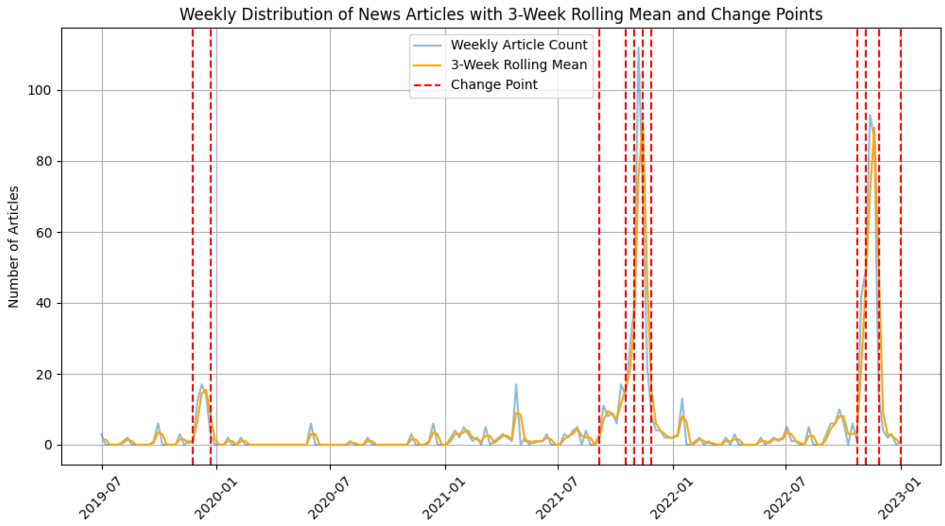
For the other two definitions corresponding to specific and recurring events, we opted for the BM25 algorithm to select topics that a priori matched the definitions. The BM25 algorithm is a probabilistic information retrieval model widely used to evaluate document relevance to a search query. BM25 is an enhanced version of the probabilistic relevance model, considering term frequency in the document and inverse document frequency in the corpus.

BM25 adjusts each term's contribution based on its frequency in the document and its rarity in the corpus, better modeling document relevance to the query. We defined a strategy to adapt this algorithm to our use, prioritizing articles highly relevant to the topic, even at the risk of missing some, thus favoring precision over recall. We defined a strategy for each topic, preferring those extractable topics with a single distinctive query term. For specific events, we selected topics like COP 26, Brexit, 5G, and Crédit Suisse, and the Pierre Maudet affair (local politician). For recurring events, we chose Christmas, grape harvests, feminist strikes, and the World Economic Forum in Davos.

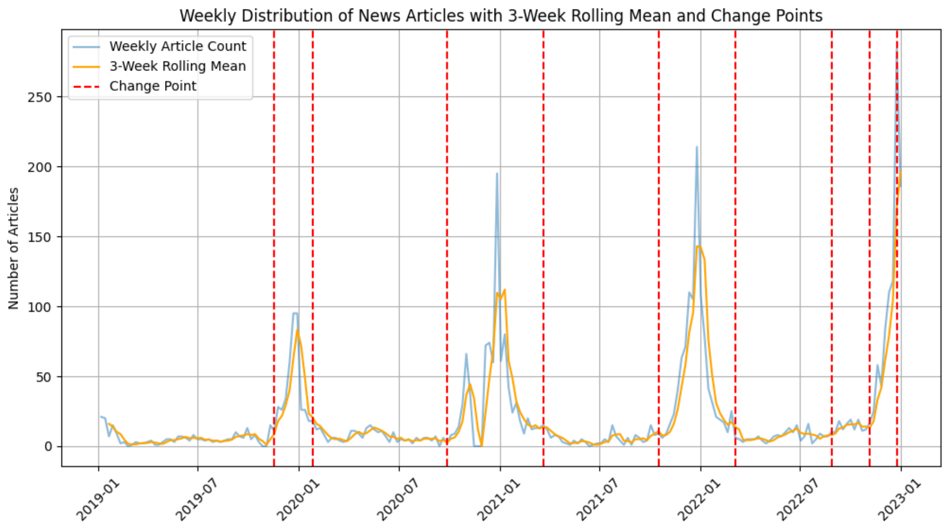
#### 9.5.3.3. Temporal Analysis Pipeline

After filtering, we applied our temporal analysis pipeline to the different topics. First, we studied the distribution of articles and the number of sources covering each topic. We then identified phases using the change point detection algorithm.

Change point detection is a statistical method used to identify moments when the properties of a time series change significantly. It segments the series into homogeneous periods by evaluating the probability of each point being a change point. Applying this algorithm, we segmented the time series of articles into distinct phases corresponding to periods of homogeneous media coverage based on frequency and volume. Figure 29 illustrates the results for some topics.




*(a) Distribution of articles for the COP events*




*(b) Distribution of articles for the Christmas recurring event*

*Figure 29. Distribution with phase identification of the coverage of the COP and Christmas events by the Swiss media.*

### 9.5.3.4. Entity Extraction and Sentiment Analysis

To extract important figures and entities related to the topics, we used a pre-trained generalist NER model, GLiNER, allowing us to define the entities we wanted to extract [186]. We defined four categories: Persons, Locations, Events, and Organizations. We identified prevalent entities and analyzed their temporal evolution to determine their appearance and connections with other entities.

Finally, we evaluated the sentiments associated with predominant entities for each topic to identify potential fluctuations based on current events. We used a combined approach of entity-linking and sentiment evaluation to determine how the opinion associated with entities evolved over time.

By combining these three aspects of analysis for each extracted topic, we reconstructed the different phases of media coverage for the selected and targeted topics. Detailed analyses for each topic will be available in the forthcoming paper publication.

### 9.5.4. Conclusion

In this research, we developed and implemented a comprehensive framework for the temporal analysis of media events within the Swiss media ecosystem. Our approach addresses the often-overlooked temporal dimension in large-scale media analysis, providing valuable insights into the evolution of media coverage over time. The key contributions of our research are summarized as follows:

- *Systematization of Temporal Analysis*: We proposed a systematic method for the temporal analysis of media events, specifically tailored to the multilingual context of Swiss media. This method allows for the identification and characterization of different phases in media coverage, enhancing our understanding of how topics evolve over time.
- *Advanced Topic Filtering and Clustering*: Utilizing advanced techniques such as the BERTopic pipeline and the BM25 algorithm, we effectively filtered and clustered articles into coherent topics. This enabled us to categorize media events into broad themes, specific events, and recurring events, facilitating targeted temporal analysis.
- *Changepoint Detection for Phase Identification*: By applying a changepoint detection algorithm, we segmented the time series of articles into distinct phases. This approach allowed us to identify significant shifts in media coverage, providing a granular view of the temporal dynamics of various topics.
- *Entity Extraction and Sentiment Analysis*: We employed advanced NLP techniques, including Named Entity Recognition and sentiment analysis, to extract and analyze key entities associated with each topic. This enabled us to track the temporal evolution of important figures and assess the sentiment dynamics related to these entities.
- *Multilingual Analysis*: Our framework was applied across the three main languages of Swiss media – German, French, and Italian – allowing for a comparative analysis that highlights distinctive elements in media coverage across different linguistic regions.

### 9.5.5. Relevant publications

An article is currently under preparation.

### 9.5.6. Relevant software/datasets/other outcomes

The associated code will be made public once the paper is finalized.

### 9.5.7. Relevance to AI4Media use cases and media industry applications

This research was a collaboration between IDIAP and CEA through the work of a Junior Fellow. The implementation of the framework of analysis for a human-centered subject is particularly relevant for the UC4: AI for Social Sciences and Humanities. The understanding of event reporting gives social scientists the means to analyze news in depth and provide in-depth interpretation of these texts.

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

## 10.1. Studying the Generalization of Media Memorability Difficulty Prediction Methods

**Contributing partner:** UNSTPB

### 10.1.1. Introduction and methodology

Memorability is one of the essential components and concepts related to human perception of visual media data, and the MediaEval Predicting Video Memorability task [190] series of benchmarking tasks plays an important role in bringing attention in the computer vision community to the study of this concept. UNSTPB's work is developed in the context of this benchmarking task, targeting the MediaEval 2023 PVM task, proposing an approach that uses sample-level analysis in order to understand and predict which samples are "hard-to-predict" / "hard-to-classify" from a memorability standpoint.

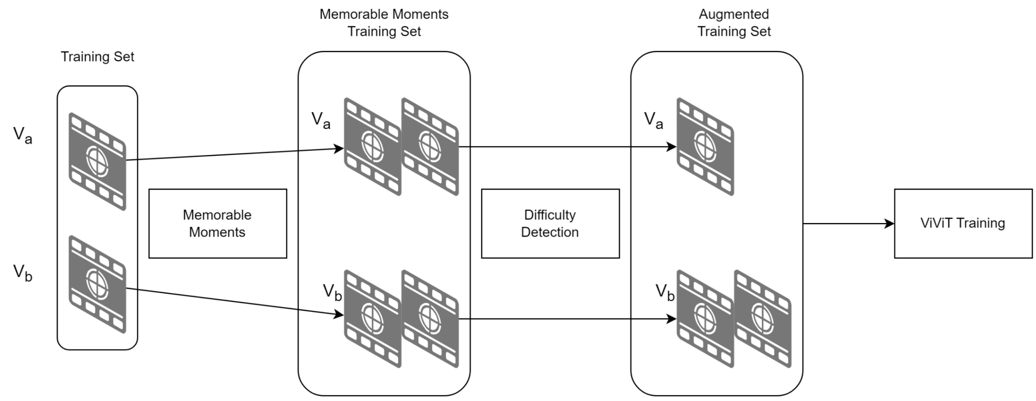



*Figure 30. Illustration of the video memorability prediction pipeline. The training videos are segmented according to the Memorable Moments schema. Following this, the Difficulty Decision module classifies the videos according to how "hard-to-predict" from a memorability standpoint the videos are. Videos that are considered "hard-to-predict" are assigned more memorable segments compared with those that are easier to predict. Following this, the video segments are used for training a ViViT-like approach.*

Our approach is built upon two of our previous works. The first one deals with methods of selecting the most important segments of video in defining memorability (which we called "Memorable Moments") [191]. The second one is a work that analyzes previous runs sent by participants at the MediaEval PVM task in order to detect the videos that are the most problematic, on average, for memorability predictor methods and find common features and methods of describing those videos [192]. The pipeline is presented in Figure 30. We propose extracting the two most Memorable Moments from each video in the training dataset. Following this, we apply several methods of detecting which videos are harder to predict by memorability predictors. For the hard-to-predict videos, we keep both extracted segments, while for the easier to predict videos, we only keep the top most significant memorable moment. In theory, this should imbalance the dataset and put more accent on videos that would otherwise be considered hard-to-predict. We theorize that this should increase the overall performance of any method, as it creates more focus towards learning difficult samples.

Concretely, for each video in the dataset, composed of $N$ frames: $V = [f_1, f_2, ...F_N]$, we use the annotations as provided by the competition organizers, and provide a score of 1 for the frames during which annotators had a video recall moment (while also accounting for possible delays in them signaling the recall moment). This will result, for each video, in a recall score assigned to all the frames in the video as follows: $R_V = [S_1, S_2, ...S_N]$. We select video segments that have the top-2 $S_i$ values as centroids, and keep them for the following steps of the pipeline.

Following our results presented in [192], we select the methods of difficulty assessment that presented the best results. The first of these methods is based on the ground-truth memorability score of each video, and perhaps unsurprisingly, concluded that videos with mid-level memorability (not too memorable, not too forgettable), are the hardest to classify by a large number of memorability predictors. The other methods use a series of discriminative features in order to determine which videos are harder to predict, as follows: sharpness computed via the Laplacian operator [193] (sharper videos are harder to classify with regards to memorability), contrast computed in RGB space [194] (higher contrast videos are harder to classify), and dynamism computed via the Farnebäck method [195]. We extract each of the four features (ground truth score, sharpness, contrast, and dynamism), and split the training set into four quartiles, according to the value of each feature. The top quartile, $Q_1$, represents videos that should be easiest to predict, while the bottom quartile, $Q_4$, represents those that would be hardest to predict. We will then keep two Memorable Moments segments only for the videos that belong to the bottom quartile, $Q_4$, for each of the four discriminative features.

The final step in the training pipeline is represented by the ViViT transformer network. For these experiments we chose a ViViT network that can process 15 frames in a window, using a tubelet embedding approach, 8 parallel self-attention blocks, and 8 repeatable transformer blocks.

#### 10.1.2. Experimental results

In evaluating the proposed method, we used the training / testing setup and split as provided by the organizers of the MediaEval Predicting Video Memorability 2023 task. Our experiments are aimed at determining which of the four proposed augmentation schemes improves the results of the ViViT approach that does not use an imbalanced dataset at training. The final results are presented in Table 36. As presented, two of the data augmentation schemes show improvement in memorability prediction, namely the augmentation based on ground truth (GT) scores, with an improvement of 5.81%, and the one based on dynamism, with an improvement of 5.26%, while the others decrease the overall performance compared with the ViViT setup where the dataset is not skewed towards hard-to-predict samples.

#### 10.1.3. Conclusions

This work analyzes a set of data augmentation schemes, that seek to increase the number of "hard-to-predict" samples with the goal of improving the results of video memorability prediction. We test four augmentation schemes, based on the ground truth memorability (augmenting the number of samples that have mid-level ground truth memorability scores), sharpness (sharper videos were found to be harder to classify and were augmented), contrast (videos with higher contrast were found to be harder to classify and were augmented), and dynamism (low-dynamism videos are harder to classify and were augmented). This augmentation scheme is applied to the quartile of videos that is considered hardest to classify according to each criterion. Final results show

| | SRCC | | | |
|---|---|---|---|---|
| Augmentation Method | devset | dev % imp | testset | test % imp |
| ViViT - baseline | 0.651 | - | 0.361 | - |
| ViViT + GT score | 0.668 | 1.07% | **0.382** | **5.81%** |
| ViViT + sharpness | 0.628 | -3.53% | 0.291 | -19.39% |
| ViViT + contrast | 0.631 | -3.07% | 0.265 | -26.59% |
| ViViT + dynamism | 0.680 | 4.45% | 0.380 | 5.26% |

*Table 36. Results of the proposed augmentation methods, according to Spearman's Rank Correlation Coefficient (SRCC) metric in the MediaEval PVM 2023 benchmarking task.*

that ground-truth and dymanism-based augmentation schemes improve the overall score, with a maximum improvement of 5.81% for ground-truth augmentation.

#### 10.1.4. Relevant publications

M.G. Constantin, B. Ionescu: AIMultimediaLab at MediaEval 2023: Studying the Generalization of Media Memorability Methods, `https://ceur-ws.org/Vol-3658/paper32.pdf`. In Working Notes Proceedings of the MediaEval 2023 Workshop, Amsterdam, The Netherlands, 2024.

#### 10.1.5. Relevant software/datasets/other outcomes

No other relevant software published at this point.

#### 10.1.6. Relevance to AI4Media use cases and media industry applications

This research can be applied to the training of systems for modality-dependent sentiment analysis. This contribution deals with the prediction of video memorability, one of the most important concepts influencing human perception of social and multimedia data, which is fundamental for the media industry.

### 10.2. Prompting Visual-Language Models for Dynamic Facial Expression Recognition

**Contributing partner:** QMUL

#### 10.2.1. Introduction

Recently, vision-language pre-training (V-LP) models such as CLIP have emerged as promising alternatives for visual representation learning. The main idea of the V-LP models is to align large connections of images and raw texts using two separate encoders, so as to learn semantic information between visual and textual data. The success of V-LP models has led to a growing interest in using them for various downstream computer vision tasks, such as video understanding , image synthesis, and semantic segmentation. Given the powerful representation learning capabilities of V-LP models, a relevant question is how to best exploit their potential for the dynamic facial expression recognition (DFER) task. One proposed approach is to fine-tune the image encoder of a V-LP model on DFER datasets. However, this approach faces two challenges that must be addressed. Firstly, while the original CLIP model can recognize objects and scenes in images, it may not be as effective in recognizing subtle facial expressions, which require more fine-grained descriptors and modelling of the similarities between expressions. Secondly, learning robust temporal facial features to understand emotions is crucial for DFER. Unfortunately, the standard CLIP visual encoder encodes static images and therefore lacks ways of capturing temporal information.

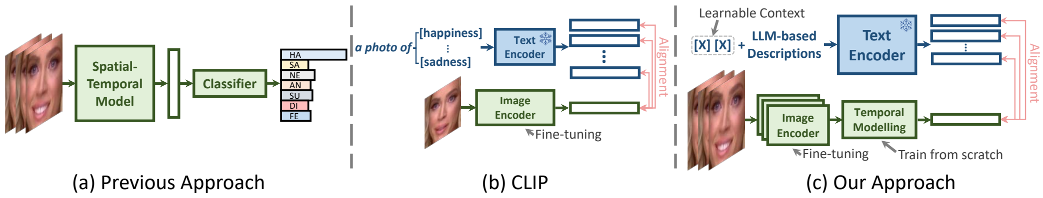



*Figure 31. Illustration of the (a) previous approach for DFER, which relies on a classifier to predict the facial expression label. (b) Standard vision-language model CLIP. (c) Based on the CLIP, we propose a DFER-CLIP, which further models temporal facial features and incorporates fine-grained descriptors.*

### 10.2.2. Methodology

To address these challenges, based on the CLIP model, we propose a novel architecture, namely the DFER-CLIP model. An overview of the differences between DFER-CLIP with previous approaches and CLIP is shown in Figure 31 and the detailed architecture is shown in Figure 32. The proposed method mainly consists of a visual part and a textual part. Regarding the visual part, based on the CLIP image encoder, we introduce a temporal model, which consists of several Transformer encoders, for modelling the temporal facial features. The final video-level facial features are obtained by a learnable class token. Regarding the textual part, considering that different facial expressions have both common properties and unique or special properties at the level of local behaviour, we utilize descriptions related to facial behaviour instead of class names for the text encoder. As a result, the text embedding can provide more detailed and precise information about the specific movements or positions of muscles involved in each expression. Furthermore, inspired by CoOp, we adopt the learnable prompt as a context for descriptors of each class – this does not require experts to design context words and allows the model to learn relevant context information for each expression during training. To evaluate our DFER-CLIP model, we conducted experiments on three datasets. The results show that the temporal model can clearly enhance performance, and adopting the fine-grained expression descriptions with learnable context is superior to class-level prompts. Furthermore, compared with the current supervised DFER methods, the proposed DFER-CLIP achieves state-of-the-art results on DFEW, FERV39k, and MAFW benchmarks.

Regarding the visual part, on top of the CLIP image encoder, the temporal model consisting of several ViT encoder layers is adopted for modelling the temporal relationship. Each encoder also consists of a multi-headed self-attention and feed-forward network, which are all trained from scratch. The frame-level features are first learnt by the shared CLIP visual encoder. Then all of the frame-level features along with an additional learnable class token will feed into the temporal model, in which the learnable position embedding is added to encode the temporal position.

Concretely, given a facial video, we sample $T$ frames of the size $H \times W$ so as to form an input $x \in \mathbb{R}^{T\times 3\times H\times W}$. For each frame $x_i$, we first utilize a shared CLIP image encoder $f(\cdot)$ to extract feature vectors $\boldsymbol{f}_i^v \in \mathbb{R}^L$, where $i \in \{1, 2, \cdots, T\}$, $L$ is the length of the feature vectors. Then $T$ feature vectors will feed into a temporal model $g(\cdot)$ for learning temporal features, and the final visual representation $\boldsymbol{f}^V \in \mathbb{R}^L$ can be obtained:

$$\boldsymbol{f}_i^v = f(x_i) \tag{8}$$

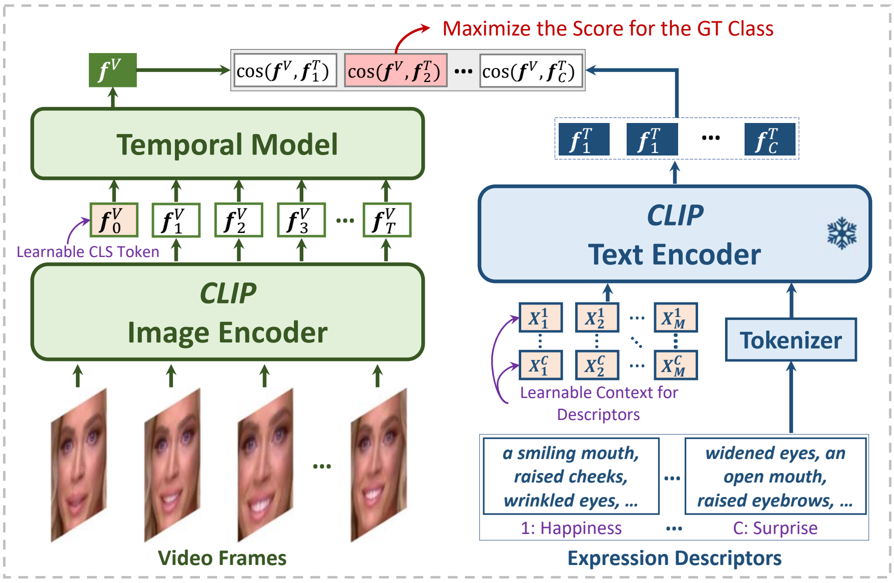



*Figure 32. The structure of the proposed DFER-CLIP model. cos() denotes cosine similarity. M denotes the learnable context number. C denotes expression classes.*

$$\boldsymbol{f}^V = g(\boldsymbol{f}_0^v + \boldsymbol{e}_0, \boldsymbol{f}_1^v + \boldsymbol{e}_1, ..., \boldsymbol{f}_T^v + \boldsymbol{e}_T) \tag{9}$$

where $\boldsymbol{f}_0^v$ is a special learnable vector for the class token, $\boldsymbol{e}$ represents a learnable position embedding added to encode the temporal position.

Regarding the textual part, we utilize descriptions related to facial behaviour instead of class names for the text encoder. Furthermore, we adopt the learnable prompt as a context for descriptors of each class. The form of the prompt is as follows:

$$\boldsymbol{P}_k = [p]_k^1[p]_k^2 \cdots [p]_k^M[Tokenizer(description)]_k \tag{10}$$

where $M$ is a hyperparameter specifying the number of context tokens, $k \in \{1, 2, \cdots, C\}$, $C$ is the number of the facial expression classes, and each $[p]_k^m$, $m \in \{1, 2, \cdots, M\}$, is a vector with the same dimension as word embeddings. Here, we adopt class-specific context where context vectors are independent of each description. By forwarding a prompt $\boldsymbol{P}_k$ to the text encoder $h(\cdot)$, we can obtain $C$ classification weight vector $\boldsymbol{f}_k^T \in \mathbb{R}^L$ representing a visual concept:

$$\boldsymbol{f}_k^T = h(\boldsymbol{P}_k) \tag{11}$$

Then the prediction probability can be computed as:

$$p(y = k|x) = \frac{exp(cos(\boldsymbol{f}^V, \boldsymbol{f}_k^T)/\tau)}{\sum_{k'=1}^{C} exp(cos(\boldsymbol{f}^V, \boldsymbol{f}_{k'}^T)/\tau)} \tag{12}$$

where $\tau$ is a temperature parameter learned by CLIP and $cos(\cdot, \cdot)$ denotes cosine similarity.

During the training phase, the CLIP text encoder is fixed and we fine-tune the CLIP image encoder. The temporal model, learnable class token, and learnable context are all learned from scratch. DFER-CLIP is trained end to end and the cross-entropy loss is adopted for measuring the distance between the prediction and the ground-truth labels.

| Methods | DFEW | | FERV39k | | MAFW | |
|---|---|---|---|---|---|---|
| | UAR | WAR | UAR | WAR | UAR | WAR |
| C3D | 42.74 | 53.54 | 22.68 | 31.69 | 31.17 | 42.25 |
| P3D | 43.97 | 54.47 | 23.20 | 33.39 | - | - |
| I3D-RGB | 43.40 | 54.27 | 30.17 | 38.78 | - | - |
| 3D ResNet18 | 46.52 | 58.27 | 26.67 | 37.57 | - | - |
| R(2+1)D18 | 42.79 | 53.22 | 31.55 | 41.28 | - | - |
| ResNet18-LSTM | 51.32 | 63.85 | 30.92 | 42.95 | 28.08 | 39.38 |
| ResNet18-ViT | 55.76 | 67.56 | 38.35 | 48.43 | <u>35.80</u> | 47.72 |
| EC-STFL [MM'20] | 45.35 | 56.51 | - | - | - | - |
| Former-DFER [MM'21] | 53.69 | 65.70 | 37.20 | 46.85 | 31.16 | 43.27 |
| NR-DFERNet [arXiv'22] | 54.21 | 68.19 | 33.99 | 45.97 | - | |
| DPCNet [MM'22] | 57.11 | 66.32 | - | - | - | - |
| T-ESFL [MM'22] | - | - | - | - | 33.28 | <u>48.18</u> |
| EST [PR'23] | 53.94 | 65.85 | - | - | - | - |
| LOGO-Former [ICASSP'23] | 54.21 | 66.98 | 38.22 | 48.13 | - | - |
| IAL [AAAI'23] | 55.71 | 69.24 | 35.82 | 48.54 | - | - |
| CLIPER [arXiv'23] | <u>57.56</u> | <u>70.84</u> | <u>41.23</u> | <u>51.34</u> | - | - |
| M3DFEL [CVPR'23] | 56.10 | 69.25 | 35.94 | 47.67 | - | - |
| AEN [CVPRW'23] | 56.66 | 69.37 | 38.18 | 47.88 | - | - |
| DFER-CLIP (Ours) | **59.61** | **71.25** | **41.27** | **51.65** | **39.89** | **52.55** |

*Table 37. Comparison with the state-of-the-art methods.*

### 10.2.3. Experimental results

To evaluate the effectiveness of our DFER-CLIP, we compare our results with several state-of-the-art methods on the DFEW, FERV39k, and MAFW benchmarks. Consistent with previous methods, the experiments on DFEW and MAFW are conducted under 5-fold cross-validation and use training and test set on FERV39k. Furthermore, we train models three times with different random seeds and then use the average for more stable and reliable results. The comparative performance in Table 37 demonstrates that the proposed DFER-CLIP outperforms the compared methods both in unweighted average recall (UAR) and Weighted Average Recall (WAR). Specifically, compared with the previous best results, our method shows a UAR improvement of 2.05%, 0.04%, and 4.09% and a WAR improvement of 0.41%, 0.31%, and 4.37% on DFEW, FERV39k, and MAFW, respectively. It should be pointed out that FERV39k is the current largest DFER benchmark with 38,935 videos. Given this substantial scale, making significant enhancements becomes a formidable task.

### 10.2.4. Conclusions

This work presents a novel visual-language model called DFER-CLIP for in-the-wild dynamic facial expression recognition. In the visual part, based on the CLIP image encoder, a temporal model consisting of several Transformer Encoders was introduced for modelling the temporal facial expression features. In the textual part, the expression descriptors related to facial behaviour were adopted as the textual input to capture the relationship between facial expressions and their underlying facial behaviour, in which the expression descriptors were generated by large language models like ChatGPT. The learnable contexts for these descriptors were also designed to help the model learn relevant context information for each expression during training. Extensive experiments demonstrate the effectiveness of each component in DFER-CLIP. Moreover, the proposed method achieves state-of-the-art results on three benchmarks.

### 10.2.5. Relevant publications

Zengqun Zhao and Ioannis Patras. "Prompting Visual-Language Models for Dynamic Facial Expression Recognition.", https://papers.bmvc2023.org/0098.pdf Proceedings of the British Machine Vision Conference (BMVC). 2023.

### 10.2.6. Relevant software/datasets/other outcomes

- The Pytorch implementation can be found in https://github.com/zengqunzhao/DFER-CLIP

### 10.2.7. Relevance to AI4Media use cases and media industry applications

This research can be applied to the training of systems for affective analysis. Such systems could be useful in media and advertising, as the reaction of viewers to media stimuli can be measured rather than relying on self-reporting.

## 10.3. EmoCLIP: A Vision-Language Method for Zero-Shot Video Facial Expression Recognition

**Contributing partner:** QMUL

### 10.3.1. Introduction and methodology

Facial Expression Recognition (FER) is a crucial task in affective computing, but its conventional focus on the seven basic emotions limits its applicability to the complex and expanding emotional spectrum. To address the issue of new and unseen emotions present in dynamic in-the-wild FER, we propose a novel vision-language model that utilises sample-level text descriptions (i.e. captions of the context, expressions or emotional cues) as natural language supervision, aiming to enhance the learning of rich latent representations, for zero-shot classification. Then, during inference, we use class-level descriptions for each emotion. Specifically, we generate descriptions of each emotion in relation to the typical facial expressions associated with it. In the case of compound emotions, we propose manipulating the latent representation of the categories' descriptions in the embedding space rather than creating additional prompts.

An overview of the proposed method, EmoCLIP, can be seen in Figure 33. We follow the CLIP [196] contrastive training paradigm to optimise a video ($E_V$) and a text ($E_T$) encoder jointly.

Given a video-text pair $x = \{x^V, x^T\}$, we obtain the video-text embeddings using the respective encoders so that $\mathbf{z}^V = E_V(x^V)$ and $\mathbf{z}^T = E_T(x^T)$, where $\mathbf{z}^V, \mathbf{z}^T \in \mathbb{R}^D$. The $\mathbf{z}^V$ and $\mathbf{z}^T$, obtained for each video-text pair in the mini-batch $B$, are utilised to generate a $B \times B$ matrix of cosine similarities, the diagonal of which corresponds to the $B$ positive pairings and off-diagonal $B^2 - B$ negative pairings. During inference, the cosine similarity of text and image embedding in the joined

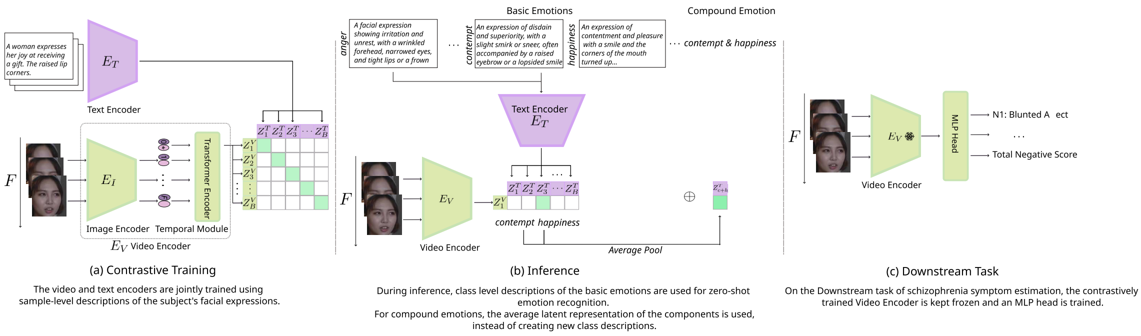



*Figure 33. Overview of our method, EmoCLIP. During training (a), we use joint training to optimise the cosine similarity of video-text embedding pairs in the mini-batch. Sample-specific descriptions of the subject's facial expressions are used to train the model. During inference (b), we perform zero-shot classification using class-level descriptions for each of the emotion categories. For compound emotions (c), we take the average latent representation of the components and concatenate them to the set of representations for each new compound emotion.*

latent space is used as the basis for the classification. The prediction probability is then defined as:

$$P(y = i|x) = \frac{e^{\langle \mathbf{z}^V, \mathbf{z}_i^T \rangle / \tau}}{\sum_{j=1}^{N} e^{\langle \mathbf{z}^V, \mathbf{z}_j^T \rangle / \tau}},$$

where $\tau$ is a learnable temperature parameter in CLIP and $\langle \cdot, \cdot \rangle$ is the cosine similarity.

**Class Descriptions** for basic emotions, in the form of natural language obtained from LLMs, rather a prompt in the form of *'an expression of* {*emotion*}*'*, are used for inference. For example, the description for *"contempt"* is: *An expression of disdain and superiority, with a slight smirk or sneer, often accompanied by a raised eyebrow or a lopsided smile.*

**Compound Emotions** are a complex combination of basic emotions, such as "happily surprised". For any new compound emotion, we calculate its latent representation $\mathbf{z}_n^T$ as the average of the latent representations of its $C$ component emotions so that:

$$\mathbf{z}_n^T = \frac{1}{C} \sum_{c=0}^{C} \mathbf{z}_c^T$$

The resulting representations are concatenated to the set of basic emotion representations for inference.

### 10.3.2. Experimental results

To evaluate the effectiveness of the proposed method, we compare it with pre-trained CLIP [196] and FaRL [197] models in a Zero-shot setting. As both of these methodologies are trained on static images, we take the average of the latent representations of all frames in a video to compute the video embedding and use that to calculate the cosine similarity with the text description embeddings. We show the performance of our method against the CLIP and FaRL baselines, with a frozen CLIP backbone and the finetuned image-text encoders on the 11 class classification of MAFW [1] in Table 38. Furthermore, we present experimental results on the classification of 43 compound emotions in the MAFW dataset in Table 39. We evaluate the performance of our proposed method, EmoCLIP, against a baseline approach of using concatenated prompts, as well as CLIP and FaRL baselines. Specifically, we concatenate the class descriptions for each compound

| Mode | Architecture | Contrastive Pre-training | UAR | WAR |
|---|---|---|---|---|
| Supervised | C3D [1] | - | 31.17 | 42.25 |
| | Resnet18_LSTM [1] | - | 28.08 | 39.38 |
| | VIT_LSTM [1] | - | 32.67 | 45.56 |
| | C3D_LSTM [1] | - | 29.75 | 43.76 |
| | T-ESFL [1] | - | 33.28 | **48.18** |
| | EmoCLIP (LP) | - | 30.26 | 44.231 |
| | EmoCLIP (Frozen backbone) | MAFW [class descriptions] | **34.24** | 41.46 |
| Zero-shot | CLIP [196] | Laion-400m | 20.40 | 21.16. |
| | FaRL - ViT-B/16 [197] | Laion Face-20M | 14.07 | 7.70 |
| | EmoCLIP | MAFW [sample descriptions] | **25.86** | **33.49** |

*Table 38. Performance of the proposed method on the MAFW [1] dataset on 11-class single expression classification against other SOTA architectures in a supervised and zero-shot setting.*

emotion and use this as class prompt input. We demonstrate that EmoCLIP outperforms the baseline approach for all metrics. Moreover, we note that in the 43 emotions classification, both CLIP and FaRL perform significantly worse than EmoCLIP and have performance comparable to random (where only the majority class is predicted).

| Mode | Architecture | Repr. Avg | UAR | WAR | F1 | AUC |
|---|---|---|---|---|---|---|
| Supervised | C3D [1] | - | **9.51** | 28.12 | 6.73 | 74.54 |
| | Resnet18_LSTM [1] | - | 6.93 | 26.6 | 5.56 | 68.86 |
| | VIT_LSTM [1] | - | 8.72 | 32.24 | **7.59** | 75.33 |
| | C3D_LSTM [1] | - | 7.34 | 28.19 | 5.67 | 65.65 |
| | T-ESFL [1] | - | 9.15 | **34.35** | 7.18 | **75.63** |
| Zero-shot | Random | - | 2.38 | 7.72 | 0.34 | 50.00 |
| | CLIP [196] | ✗ | 4.72 | 5.25 | 2.44 | 51.89 |
| | CLIP [196] | ✓ | 4.14 | 5.35 | 2.46 | **53.07** |
| | FaRL [197] | ✗ | 3.03 | 4.66 | 2.16 | 51.01 |
| | FaRL [197] | ✓ | 4.00 | 5.75 | 2.56 | 51.10 |
| | EmoCLIP | ✗ | 5.24 | 15.34 | 3.80 | 51.30 |
| | EmoCLIP | ✓ | **6.58** | **18.53** | **4.78** | 52.59 |

*Table 39. Zero-shot classification on the 43 compound expressions of the MAFW [1] dataset. Supervised methods are included as a reference.*

Finally, we evaluate the performance of our proposed method using sample-level descriptions from MAFW [1] on four widely used video FER datasets and compare it with the CLIP baseline as shown in Table 40. Additionally, in line with previous works in zero-shot emotion classification [198, 199, 200, 201], we train our architecture using class-level descriptions and evaluate using leave-one-class-out (loco) cross-validation. We note that we cannot directly compare with these architectures, as they involve either different modalities (e.g. audio, pose) [198, 199, 201, 202] or a different task [200], we adopt however, their experimental set-up using our architecture to show how natural language supervision and semantically rich class descriptions can help improve zero-shot FER performance.

We observe that the EmoCLIP trained on MAFW [1] sample-level descriptions show impressive generalisation ability on all datasets that we evaluate. Specifically, for AFEW [203], MAFW [1] and DFEW [204], we see that the EmoCLIP model is outperforming both the loco experiment and the CLIP [196] baseline. Furthermore, the generalisation of the method is resistant to domain shift from unseen datasets, as we observe from the significant performance increase between the

CLIP [196] baseline and EmoCLIP. We note that for FERV39K [205], the loco experiment has a higher performance than the sample-wise training. However, it is very important to stress that the FERV39K [205] is significantly larger than the base dataset (over 3x more samples); therefore, methods trained on it would have an advantage, particularly as in the loco experiment, there is no domain shift.

| | | | DFEW (7 classes) | | AFEW (7 classes) | | FERV39K (7 classes) | | MAFW (11 classes) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Architecture | Training labels | UAR | WAR | UAR | WAR | UAR | WAR | UAR | WAR |
| Supervised | EmoCLIP | [class description] | **58.04** | **62.12** | **44.32** | **46.19** | **31.41** | 36.18 | **34.24** | 41.46 |
| | EmoCLIP (LP) | [class] | 50.29 | 62.09 | 33.74 | 38.85 | 30.58 | **43.54** | 30.26 | **44.21** |
| Zero-shot | CLIP [196] | [image caption] | 19.86 | 10.60 | 23.05 | 11.80 | 20.99 | 17.10 | 20.04 | 21 |
| | EmoCLIP (leave-one-class-out) | [class description] | <u>22.85</u> | <u>24.96</u> | <u>35.11</u> | <u>27.57</u> | **39.35** | **41.60** | <u>24.12</u> | <u>24.74</u> |
| | EmoCLIP | [video caption] | **36.76** | **46.27** | **36.13** | **39.90** | <u>26.73</u> | <u>35.30</u> | **25.86** | **33.49** |

*Table 40. Evaluation of EmoCLIP using sample descriptions vs class-level description as natural language supervision, on four video FER datasets.*

### 10.3.3. Conclusions

In this work, we presented a novel contrastive pre-training paradigm for video FER, trained on video-text pairs with sample-level descriptions without any class-level information. While contrastive learning and natural language supervision have been used in other domains, zero-shot emotion recognition remains surprisingly unexplored, with works focusing on creating class prototypes with simpler word encoding methods [198, 201, 199, 202]. Emotional prototypes, however, disregard the intra-class variation that is inherently present in FER tasks. To overcome the limitations of training on coarse emotional categories, EmoCLIP is trained on sample-level descriptions. We evaluate our method on four popular FER video datasets [203, 204, 205, 1] and test using zero-shot evaluation on the basic emotions as well as compound emotions. Our method outperforms the CLIP baseline by a large margin and shows impressive generalisation ability on unseen datasets and emotions. To our knowledge, this is the first work to train with sample-level descriptions for FER and to propose zero-shot evaluation using semantically rich class descriptions in the domain.

### 10.3.4. Relevant publications

N. M. Foteinopoulou and I. Patras, "EmoCLIP: A Vision-Language Method for Zero-Shot Video Facial Expression Recognition.", https://arxiv.org/abs/2310.16640, The 18th IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2024

### 10.3.5. Relevant software/datasets/other outcomes

The Pytorch implementation can be found in: https://github.com/NickyFot/EmoCLIP/

### 10.3.6. Relevance to AI4Media use cases and media industry applications

This research can be applied to the training of systems for affective 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.

# 11. Other relevant activities

In addition to scientific and technical work, WP6 partners contributed to organizing relevant activities in the areas covered by the tasks. These include the organization of workshops, evaluation campaigns, special sessions. This section summarises these contributions and provides details regarding their intended purpose and topics.

Continuing a series started in 2022 and 2023, T6.2 partners co-organized the third edition of the **Multimedia Against Disinformation (MAD) workshop**[19] in conjunction with the ACM ICMR 2024 conference. These workshops target several key aspects of disinformation detection, such as multimodality, the analysis of disinformation campaigns, explaining disinformation, temporal and cultural aspects, multimedia verification systems, and ensembling techniques. In the latest edition of this workshop, i.e. MAD'24 held in Phuket, Thailand on 10 June 2024, 11 papers were accepted, and 2 keynote speakers were invited. The workshop included four sessions on (1) the evaluation of AI models, (2) the detection of synthetic audio-visual content, (3) AI for image and video analysis, and (4) AI-based automatic fact-checking. The event was co-organized by AI4Media and veraAI[20]. A summary of the workshop is available here.

WP6 partners co-organized the **Meet the Future of AI - Generative AI and Democracy** event[21] on 19 June 2024 in Brussels, Belgium. The event comprised two main sessions focused on "GenAI in latest elections and other disinformation campaigns" and "Harnessing GenAI against disinformation and for trust building". These topics are tightly related to WP6 since they focus on the effects of AI in politics and society. The event was co-organized with the following projects: Titan, veraAI, AI4Trust, AI4Debunk and AI-CODE. A summary of the event findings is available here.

WP6 partners co-organized a workshop focused on **Assessing and Enhancing Fairness in AI systems** during the 4th AI Community Workshop organized by the AI NoEs [22]. The event was held in Thessaloniki, Greece on 26 June 2024. The workshop focused on the following themes: (1) interdisciplinary initiatives that seek to make AI fairer, such as initiatives to reduce dataset biases by design, (2) effects of biases in high-impact AI applications (face recognition, recommenders, automatic scoring, media analysis), and (3) representational biases in large multimodal and language models. The workshop attracted an audience of approximately 20 people in the room plus a dozen connected remotely. The video recording of the meeting is available here.

Finally, individual WP6 partners also presented their work at a number of other events. Specifically, IDIAP participated in the Perspectives on AI Symposium called AI and Democracy: Opportunities and Risks, in Martigny, Switzerland, February 21, 2024 [23]. IDIAP also participated in the AI and Human Rights event organized by the Scandinavian Forum in Zurich, Switzerland, November 2, 2023 [24]. Adrian Popescu participated in a session of the CLAIRE X NoEs All Questions Answered Series on AI for Citizens. The recording of the session is available here.

[19] https://mad2024.aimultimedialab.ro/
[20] https://www.veraai.eu/
[21] https://www.ai4media.eu/event/meet-the-future-of-ai-generative-ai-and-democracy/
[22] https://www.vision4ai.eu/community-workshop-2024/
[23] https://perspectives-on-ai.idiap.ch
[24] https://swecham.ch/events-2/partner-event-flynn-coleman-ai-and-human-rights/

# 12. Conclusions and Perspectives

## 12.1. Ongoing work

Below, we briefly summarize the ongoing work associated with each reported technical task.

### 12.1.1. Policy recommendations on content moderation (Task 6.1)

**KUL** continues exploring the influence of the European regulatory framework on the deployment of AI-based content moderation. Additionally, KUL is currently busy preparing publications about content moderation of deepfakes, disinformation and AI generated outputs and the specific status of content from "media" outlets online. Furthermore, the AI Media Observatory will keep monitoring and curating relevant considerations about the use of AI for content production and for content moderation.

### 12.1.2. Manipulation and synthetic content detection in multimedia (Task 6.2)

**UNITN** is currently working on a novel approach involving a graph Transformer generative adversarial network (GTGAN) to learn effective graph node relations in an end-to-end fashion for challenging graph-constrained architectural layout generation tasks.

**CERTH** is refining the proposed model architectures for deepfake detection for robustness and explainability, to increase their trustworthiness to the levels required for deployment in the wild. The team is also working on refining and expanding the Frequency MLP-Mixer architecture for synthetic speech detection. This will involve exploring various modifications to the frequency-domain processing component, potentially incorporating more advanced spectral analysis techniques or adaptive filtering methods. Additionally, the model's performance will be evaluated on a wider range of datasets beyond ASVspoof 2019, including more recent benchmarks and real-world scenarios, to assess its generalizability and robustness.

**UNIFI** is currently investigating the distinctive feature given by the camera surface frames by analyzing how their different characteristics can be combined and exploited. In particular, the idea is to understand if they can improve performances specifically in terms of generalization in a cross-forgery scenario for deepfake detection.

### 12.1.3. Hybrid, privacy-enhanced recommendation (Task 6.3)

**FhG** will continue on integrating the metadata and relative article differences generated by the LLMs in a graph-based recommender system. Our goal is to have a unique cross-vendor news recommender system that presents different views and opinions on the same topics.

### 12.1.4. AI for Healthier Political Debate (Task 6.4)

**CEA** will investigate multimodal news analysis. The focus will be on the visual framing of newspaper articles by mining objective and subjective cues from the photos associated with the articles.

**UCA** will continue work on argument mining by testing more recent large foundation models and extending the datasets to content coming from diversified sources.

### 12.1.5. Perceptions of hyper-local news (Task 6.5)

**IDIAP** is in the process of finalizing the research mentioned in this report and writing the related papers. The ongoing work involves refining some final experiments and completing the writing process. The team plans to continue the analysis of complex concepts in journalistic treatment using AI methodologies. This includes identifying possible groupings in an unsupervised manner by guiding the extraction of embeddings.

#### 12.1.6. Measuring and Predicting User Perception of Social Media (Task 6.6)

**QMUL** is focusing on the use of Large Multimodal models and how these can be used to incorporate domain knowledge so as to go beyond the typical classification-based paradigm.

**UNSTPB** continues working on the MediaEval Benchmarking Initiative for Multimedia Evaluation[25]. It offers a series of interesting and challenging tasks related to multimedia data processing, analysis, and retrieval. Throughout the years, Task 6.6 has closely collaborated with MediaEval, resulting in a series of interesting tasks and challenges focused on the human perception of social media items. In this context, UNSTPB is currently working on writing and submitting a paper to the International Journal of Computer Vision that collects and analyzes all the contributions to past editions of the MediaEval Predicting Media Memorability task, a benchmarking task that ran from 2018 to 2024 [190]. Our ongoing work proposes to gather all the runs and methods of the participants over the years and analyze their results, trends, high-level conclusions, methods, and features, as well as propose new fusion methods based on these runs in order to increase the overall performance on the memorability dataset.

### 12.2. Conclusions

In this deliverable, we presented the research results obtained in tasks T6.1 to T6.6 from M37 to M48. The contributions addressed a large spectrum of open challenges associated with the effects of AI technologies on individuals and society as a whole. They led to many high-quality outcomes, as highlighted by the numerous publications in high-impact venues discussed in the deliverable.

T6.1 shed light on the complex content moderation landscape and the multitude of stakeholders involved. It also showed how the EU policy and legal landscape is shifting from an intermediary liability regime to a regime with liability and accountability rules imposed on online intermediaries to guarantee a safe and healthy online environment. The arrival of GenAI tools also shaken content moderation policies. Much is yet to be discovered about the new content moderation risks arising from GenAI. The obligations imposed by the AI Act and the DSA aim to strike a fair balance between fundamental rights. The ongoing implementation of the DSA and the AI Act represent critical components of this new regulatory landscape. However, the rapid evolution of content moderation challenges—exacerbated by the rise of AI—demands continued research and analysis. Future studies are essential to understand the impacts of these regulations, to refine content moderation strategies, and to navigate the delicate balance between safeguarding users and upholding fundamental rights.

Task 6.2 focused on AI techniques that improve the generation and detection of synthetic media. Regarding synthetic media generation, work focused on three tasks: a) improving virtual cloth try-on tools by introducing a method based on an interactive transformer, b) proposing edge-guided GANs for more realistic semantic image synthesis, and c) automatically generating human reactions in videos by using an innovative interaction transformer.

Regarding synthetic media detection, the task investigated various techniques to improve the discriminative capabilities and generalization capabilities of deepfake detectors. These techniques include a) using the representations from the intermediate encoder blocks of CLIP, b) incorporating attribution to the standard detection objective, which can improve the explainability of results, c) training detectors with adversarial augmentation to increase robustness, d) using appropriate features to detect inconsistencies in the surfaces of a deepfake scene, which are invisible to the

[25] https://multimediaeval.github.io/

human eye, e) adapting emerging models from the computer vision literature to synthetic audio detection. In addition, Task 6.2 conducted a comprehensive investigation of model compression and transfer learning techniques to enable their deployment at the edge (e.g. smartphones). Overall, Task 6.2 led to advances that have made synthetic media detection methods better performing, more robust and cost-effective, and therefore more practical to deploy in a variety of operational conditions. At the same time, it has been found that maintaining high detection accuracy in the wild is still an open challenge, and therefore, more research will be needed in the future in order to keep up with the increasing risk of hyper-realistic synthetic media generation capabilities.

Task 6.3 presented an update of the use of the relatively novel text processing capabilities of LLMs to use that for a comparative evaluation of opinion pieces, new ways to extract relevant metadata from a text by *just asking* instead of training specialized classifiers. This will all be used in a novel, graph based and explainable news recommender system.

Task 6.4 studied various aspects of the automatic analysis of political debates. The contributions presented in this deliverable provide useful complements to previous work with focus on: a) the analysis of political news published in two European countries from objective and subjective perspective, b) the automatic update of insights extracted from digital archives of international political debates to reflect new questions and highlight previously marginalized perspectives, and c) the detection and classification of fallacies in political debates, with the introduction of a long-term dataset including US presidential debates and proposal of a transformer-based method to address the task. Overall, T6.4 proposed innovative tools that improve the understanding of political debates reflected in the news, the social media, and digital archives. These tools are usable by media professionals, as shown by the integration of some of them in AI4Media use cases, but also by citizens, as demonstrated by the integration in user-facing prototypes.

Task 6.5 studied the ecosystem of local news with AI methodologies. In conclusion, the exploration of hyper-local news through various AI methodologies has produced valuable insights into the dynamics of local media ecosystems. The studies developed in Task 6.5 demonstrate the potential of leveraging advanced computational techniques to analyze and understand some of the specificities of local journalism. By examining frame analysis, the relationship between migrant communities and local media, and the content dynamics in different Swiss regions, the research conducted in this task has highlighted the nuanced ways in which local news addresses and serves its communities. The findings produced by the different analyses underscore the importance of local media in addressing community-specific issues, promoting social cohesion, and ultimately contributing to support democratic processes. Furthermore, the temporal analysis of media events offers a framework for understanding the evolution of media coverage and its impact on public opinion. Overall, these insights contribute to the understanding of how to promote a more informed and engaged citizenry, emphasizing the important role of local news in contemporary media landscapes.

Task 6.6 targets the study of user perception of social and multimedia data, measured with the help of two very important emotional concepts: video memorability and facial expression and emotion recognition. The study of memorability aims to discover methods of leveraging content that would otherwise be hard to classify and to augment those videos by selecting more significant segments and using them in training, leading to methods of training AI memorability models that would be more robust and opens the door for retrieving results in an interpretable way. Regarding facial expression classification, we studied methods using vision-language pre-training models and zero-shot learning approaches. Experimental results show state-of-the-art performance for these approaches, with good generalization capabilities for the models.

## 12.3. Open Challenges

WP6 studied a variety of topics associated with the impact of AI on citizens and society. These works contributed to the state of the art but novel challenges appeared during the project lifetime. They are due to the rapid advancement of AI technologies, and foundation models in particular. These large models reshaped the research priorities and found practical applications very rapidly. We discuss below open challenges associated with WP6 tasks.

**T6.1-related challenges**:

- The articulation between the European legal framework, designed to preserve fundamental freedoms and regulate the development of digital platforms and services in agreement with democratic values, and the technical capabilities for content moderation offered by AI systems, driven by the platforms' business models, is a complex matter. From a legal research perspective, the challenge is to provide actionable interpretations of the regulations to platforms and end users. From a technical perspective, it is necessary to introduce AI systems that integrate legal requirements by design.

- Online platforms are required to introduce content moderation tools that ensure lawful operation within the EU. Human-based moderation is currently deployed by it has high cost, and there AI-based tools are increasingly used. This raises questions about their performance level and their conformity with legal requirements. Auditing tools that combine human expertise and automatic analysis should be deployed to evaluate the content moderation efforts independently.

**T6.2-related challenges**:

- The quality of generated content has progressed significantly in recent years with the availability of very large models. However, the realism of synthetic data and the associated biases remain open problems. For instance, generated content exhibits cultural, domain-related and spatio-temporal biases that arise from the imbalanced distribution of the training set. As a consequence, generation is particularly challenging for scarcely-represented and/or specialized domains whose associated data are rarely seen during model pretraining.

- In real-world scenarios, synthetic content detection most often happens in black-box configurations since the generation is unknown and/or unavailable. Combined with the growing quality of generated content, this creates an asymmetry of means that increases the difficulty of the detection process. The task is likely to become even more difficult in the future with the progress made in content generation.

- As the prevalence of synthetic media increases in the digital realm, so does the challenge of verifying that a media item is authentic by overcoming the so-called Liar's Dividend, i.e. the situation where people discredit genuine media content as synthetic. This is particularly concerning when there is media evidence of public figures' (e.g. politicians) wrongdoings, which such figures quickly try to discredit as being a deepfake. To overcome this challenge, the accuracy and reliability of detection approaches, especially with respect to the number of false positives, needs to considerably improve. Additionally, "active" protection approaches such as watermarking could provide an additional alternative verification approach.

**T6.3-related challenges**:

- Most existing recommendation systems that are deployed in commercial settings aim to maximize the time spent by the user on the online platforms. This objective might lead to proposing content that is aligned with the users' existing preferences at the expenses of novelty and diversity. This functioning is not problematic when recommending entertainment-related content, but has adverse effects when dealing with socially-important topics, such as politics. The challenge is to create recommenders that offer an adequate balance between catering to user's preferences and introducing diversity in the proposed content.

- LLMs and VLMs can be used successfully in recommendation tasks. However, their usage raises trust and privacy concern since personal data needs to be processed by the centralized platforms whose functioning remains opaque. To harness the power of large models while preserving privacy, future research could investigate hybrid recommenders in which content pre-selection is done by an centralized model while result refinement is performed locally.

**T6.4-related challenges**:

- The proposed contributions focus on particular aspects of the political debate understanding. While some NLP components were integrated to move toward a comprehensive analysis, more efforts are required to achieve this goal. Equally important, efforts were focused on texts and it would be useful to work toward an automatic multimedia content analysis. Finally, the packaging of the insights provided by AI algorithms should be prioritized to better serve the needs of different media stakeholders (journalists, editors, political scientists, citizens, etc.).
- The advent of large models has lowered the entry barrier for analyzing political texts since many of the individual tasks can now be solved using zero-shot of few-shot approaches for both textual and visual content. However, strong challenges remain regarding the factuality of the proposed insights, the consideration of multilingual and/or cultural specificities.

**T6.5-related challenges**:

- The local news analysis is an under-researched topic despite the fact that local sources have a strong impact on users' daily lives. The proposed contributions provided interesting results for Switzerland. It would be useful to adapt the analysis to other countries and/or cultural contexts. Another interesting direction is the comparison of local, national, and European contexts. This would allow a finer grained-understanding of common and specific topics.
- The access to local news sources remains problematic because their digitization is more reduced compared to national outlets. Even when digital archives are available, they might not be included in standardized collections, such as CCNews[26]. We also note that access to full content is often paywalled and the obtained insights are based on the freely available parts of the articles. Beyond technical aspects, there are legal issues regarding the access to editorialized content that are only partly solved by research-targeted exceptions.

**T6.6-related challenges**:

- The project addressed the perception of social media by human subjects. Legal challenges arise under the AI Act since some of the tasks might be considered in the unacceptable risks category. The challenge here is to establish whether research in this area falls under the exceptions foreseen by the regulation.
- Pretrained visual-language models proved useful for solving media perception-related tasks, and contributions going beyond a zero-shot setting were proposed. Further efforts are required to see how well VLMs compare to models built specifically for each task, and how one could get the best of the two worlds.

[26] https://commoncrawl.org/blog/news-dataset-available

**T6.7-related challenges**:

- The proposed contributions enable users to understand how their online profile might affect them in impactful real-life situations. The algorithms were implemented at the edge to preserve user privacy. However, the question of whether they qualify as social scoring under the AI Act or not remains open. This question affects any privacy feedback and awareness work involving a comparison among users.
- The availability of VLMs has increased the performance of photo-related inferences but these models do not ensure privacy. A quality gap in their favor appears when comparing predictions with those of edge algorithms. This gap might be a challenge for the adoption of the proposed application prototypes, a recurring problem of privacy feedback and awareness tools.

The WP6 participants hope that these challenges will be met at least in part. Addressing them would facilitate the adoption of AI technologies in the media industry and beyond and ensure that its impact on society and citizens remains globally positive.

## References


[1] Y. Liu, W. Dai, C. Feng, W. Wang, G. Yin, J. Zeng, and S. Shan, "MAFW: A Large-scale, Multi-modal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild," in Proceedings of the 30th ACM International Conference on Multimedia, MM '22, Oct. 2022.

[2] 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.

[3] 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, pp. 1–11, 2019.

[4] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," Journal of machine learning research, vol. 9, no. 11, 2008.

[5] S. Zhu, R. Urtasun, S. Fidler, D. Lin, and C. Change Loy, "Be your own prada: Fashion synthesis with structural coherence," in International Conference on Computer Vision, pp. 1680–1688, 2017.

[6] Y. Men, Y. Mao, Y. Jiang, W.-Y. Ma, and Z. Lian, "Controllable person image synthesis with attribute-decomposed gan," in Conference on Computer Vision and Pattern Recognition, pp. 5084–5093, 2020.

[7] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3d people models," in Conference on Computer Vision and Pattern Recognition, pp. 8387–8397, 2018.

[8] Z. Lahner, D. Cremers, and T. Tung, "Deepwrinkles: Accurate and realistic clothing modeling," in European Conference on Computer Vision, pp. 667–684, 2018.

[9] E. Gundogdu, V. Constantin, A. Seifoddini, M. Dang, M. Salzmann, and P. Fua, "Garnet: A two-stream network for fast and accurate 3d cloth draping," in International Conference on Computer Vision, pp. 8739–8748, 2019.

[10] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black, "Drape: Dressing any person," ACM Transactions on Graphics, vol. 31, no. 4, pp. 1–10, 2012.

[11] M. Sekine, K. Sugita, F. Perbet, B. Stenger, and M. Nishiyama, "Virtual fitting by single-shot body shape estimation," in Int. Conf. on 3D Body Scanning Technologies, pp. 406–413, Citeseer, 2014.

[12] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, "Clothcap: Seamless 4d clothing capture and retargeting," ACM Transactions on Graphics, vol. 36, no. 4, pp. 1–15, 2017.

[13] 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.

[14] H. Dong, X. Liang, X. Shen, B. Wu, B.-C. Chen, and J. Yin, "Fw-gan: Flow-navigated warping gan for video virtual try-on," in International Conference on Computer Vision, pp. 1161–1170, 2019.

[15] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "Viton: An image-based virtual try-on network," in Conference on Computer Vision and Pattern Recognition, pp. 7543–7552, 2018.

[16] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in European Conference on Computer Vision, pp. 589–604, 2018.

[17] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo, "Towards photo-realistic virtual try-on by adaptively generating-preserving image content," in Conference on Computer Vision and Pattern Recognition, pp. 7850–7859, 2020.

[18] M. Minar, T. Tuan, H. Ahn, P. Rosin, and Y. Lai, "Cp-vton+: Clothing shape and texture preserving image-based virtual try-on," in Conference on Computer Vision and Pattern Recognition Workshops, vol. 2, p. 11, 2020.

[19] J. Xu, Y. Pu, R. Nie, D. Xu, Z. Zhao, and W. Qian, "Virtual try-on network with attribute transformation and local rendering," IEEE Transactions on Multimedia, 2021.

[20] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," IEEE Transactions on pattern analysis and machine intelligence, vol. 11, no. 6, pp. 567–585, 1989.

[21] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in European Conference on Computer Vision, pp. 128–142, 2002.

[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.

[23] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," IEEE transactions on pattern analysis and machine intelligence, vol. 19, no. 5, pp. 530–535, 1997.

[24] T. Issenhuth, J. Mary, and C. Calauzenes, "Do not mask what you do not need to mask: a parser-free virtual try-on," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, 23–28, 2020, pp. 619–635, Springer, 2020.

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, vol. 13, no. 4, pp. 600–612, 2004.

[26] 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.

[27] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in NeurIPS, 2016.

[28] B. Ren, H. Tang, F. Meng, R. Ding, P. Torr, and N. Sebe, "Cloth interactive transformer for virtual try-on," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 20, no. 4, 2023.

[29] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in ICCV, 2017.

[30] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in CVPR, 2017.

[31] S. Gu, J. Bao, H. Yang, D. Chen, F. Wen, and L. Yuan, "Mask-guided portrait editing with conditional gans," in CVPR, 2019.

[32] X. Liu, G. Yin, J. Shao, X. Wang, et al., "Learning to predict layout-to-image conditional convolutions for semantic image synthesis," in NeurIPS, 2019.

[33] X. Qi, Q. Chen, J. Jia, and V. Koltun, "Semi-parametric image synthesis," in CVPR, 2018.

[34] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in CVPR, 2019.

[35] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in CVPR, 2018.

[36] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in CVPR, 2016.

[37] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in CVPR, 2017.

[38] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in CVPR, 2018.

[39] H. Tang, X. Qi, G. Sun, D. Xu, N. Sebe, R. Timofte, and L. Van Gool, "Edge guided gans with semantic preserving for semantic image synthesis," in ICLR, 2023.

[40] 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.

[41] V. Sushko, E. Schönfeld, D. Zhang, J. Gall, B. Schiele, and A. Khoreva, "You only need adversarial supervision for semantic image synthesis," in ICLR, 2021.

[42] H. Tang, S. Bai, and N. Sebe, "Dual attention gans for semantic image synthesis," in ACM MM, 2020.

[43] H. Tang, D. Xu, Y. Yan, P. H. Torr, and N. Sebe, "Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation," in CVPR, 2020.

[44] H. Tang and N. Sebe, "Layout-to-image translation with double pooling generative adversarial networks," IEEE TIP, vol. 30, pp. 7903–7913, 2021.

[45] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, "Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation," in CVPR, 2019.

[46] H. Tang, P. H. Torr, and N. Sebe, "Multi-channel attention selection gans for guided image-to-image translation," IEEE TPAMI, 2022.

[47] H. Tang, L. Shao, P. H. Torr, and N. Sebe, "Local and global gans with semantic-aware upsampling for image generation," IEEE TPAMI, 2022.

[48] Y. Wang, L. Qi, Y.-C. Chen, X. Zhang, and J. Jia, "Image synthesis via semantic composition," in ICCV, 2021.

[49] Y. Shi, X. Liu, Y. Wei, Z. Wu, and W. Zuo, "Retrieval-based spatially adaptive normalization for semantic image synthesis," in CVPR, 2022.

[50] Z. Lv, X. Li, Z. Niu, B. Cao, and W. Zuo, "Semantic-shape adaptive feature modulation for semantic image synthesis," in CVPR, 2022.

[51] T. Wang, T. Zhang, B. Zhang, H. Ouyang, D. Chen, Q. Chen, and F. Wen, "Pretraining is all you need for image-to-image translation," arXiv preprint arXiv:2205.12952, 2022.

[52] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," arXiv preprint arXiv:2302.08453, 2023.

[53] W. Wang, J. Bao, W. Zhou, D. Chen, D. Chen, L. Yuan, and H. Li, "Semantic image synthesis via diffusion models," arXiv preprint arXiv:2207.00050, 2022.

[54] H. Tang, G. Sun, L. van Gool, and N. Sebe, "Edge guided gans with multi-scale contrastive learning for semantic image synthesis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 12, pp. 14435–14452, 2023.

[55] M. Baruah and B. Banerjee, "A multimodal predictive agent model for human interaction generation," in CVPR Workshops, 2020.

[56] J. Kundu, H. Buckchash, P. Mandikal, R. M V, A. Jamkhandi, and R. Babu, "Cross-conditioned recurrent networks for long-term synthesis of inter-person human motion interactions," in WACV, 2020.

[57] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in CVPR Workshops, 2012.

[58] T. Hu, X. Zhu, and W. Guo, "Two-person interaction recognition based on key poses," Journal of Computational Information Systems, vol. 10, pp. 1965–1972, 2014.

[59] M. Maghoumi and J. J. LaViola, "Deepgru: Deep gesture recognition utility," in Advances in Visual Computing, 2019.

[60] 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.

[61] 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.

[62] H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, and J. Kautz, "Dancing to music," in NeurIPS, 2019.

[63] 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.

[64] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in CVPR, 2017.

[65] S. Aliakbarian, F. S. Saleh, M. Salzmann, L. Petersson, and S. Gould, "A stochastic conditioning scheme for diverse human motion prediction," in ICCV, 2020.

[66] E. Aksan, M. Kaufmann, P. Cao, and O. Hilliges, "A spatio-temporal transformer for 3d human motion prediction," in 2021 International Conference on 3D Vision (3DV), pp. 565–574, IEEE, 2021.

[67] T. Ma, Y. Nie, C. Long, Q. Zhang, and G. Li, "Progressively generating better initial guesses towards next stages for high-quality human motion prediction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6437–6446, June 2022.

[68] B. Chopin, H. Tang, N. Otberdout, M. Daoudi, and N. Sebe, "Interaction transformer for human reaction generation," IEEE Transactions on Multimedia, vol. 25, pp. 8842–8854, 2023.

[69] G. Iglesias, E. Talavera, and A. Díaz-Álvarez, "A survey on gans for computer vision: Recent research, analysis and taxonomy," Computer Science Review, vol. 48, p. 100553, 2023.

[70] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[71] Z. Lu, D. Huang, L. Bai, J. Qu, C. Wu, X. Liu, and W. Ouyang, "Seeing is not always believing: Benchmarking human and model perception of ai-generated images," in Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2023.

[72] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," Information Fusion, vol. 64, pp. 131–148, 2020.

[73] 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, pp. 8695–8704, 2020.

[74] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? understanding properties that generalize," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16, pp. 103–120, Springer, 2020.

[75] 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, pp. 3247–3258, PMLR, 2020.

[76] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in European conference on computer vision, pp. 86–103, Springer, 2020.

[77] Y. Jeong, D. Kim, S. Min, S. Joe, Y. Gwon, and J. Choi, "Bihpf: Bilateral high-pass filters for robust deepfake detection," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 48–57, 2022.

[78] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24480–24489, 2023.

[79] A. F. Agarap, "Deep learning using rectified linear units (relu)," arXiv preprint arXiv:1803.08375, 2018.

[80] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” The journal of machine learning research, vol. 15, no. 1, pp. 1929–1958, 2014.

[81] I. J. Good, “Rational decisions,” Journal of the Royal Statistical Society: Series B (Methodological), vol. 14, no. 1, pp. 107–114, 1952.

[82] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, “Supervised contrastive learning,” NeurIPS, 2020.

[83] Y. Jeong, D. Kim, Y. Ro, and J. Choi, “Frepgan: robust deepfake detection using frequency-level perturbations,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 1060–1068, 2022.

[84] C. Tan, Y. Zhao, S. Wei, G. Gu, and Y. Wei, “Learning on gradients: Generalized artifacts representation for gan-generated images detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12105–12114, 2023.

[85] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, “On the detection of synthetic images generated by diffusion models,” in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, IEEE, 2023.

[86] C. Chen, Q. Chen, J. Xu, and V. Koltun, “Learning to see in the dark,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3291–3300, 2018.

[87] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” in International Conference on Learning Representations, 2018.

[88] 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, pp. 4401–4410, 2019.

[89] 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, pp. 8110–8119, 2020.

[90] A. Brock, J. Donahue, and K. Simonyan, “Large scale gan training for high fidelity natural image synthesis,” in ICLR, 2019.

[91] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in ICCV, 2017.

[92] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, “Second-order attention network for single image super-resolution,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11065–11074, 2019.

[93] K. Li, T. Zhang, and J. Malik, “Diverse image synthesis from semantic layouts via conditional imle. 2019 ieee,” in CVF International Conference on Computer Vision (ICCV), pp. 4219–4228, 2019.

[94] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” Advances in neural information processing systems, vol. 34, pp. 8780–8794, 2021.

[95] 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, pp. 10684–10695, 2022.

[96] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," arXiv preprint arXiv:2112.10741, 2021.

[97] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in International Conference on Machine Learning, pp. 8821–8831, PMLR, 2021.

[98] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, "Countering malicious deepfakes: Survey, battleground, and horizon," Int. J. Comput. Vis., vol. 130, no. 7, pp. 1678–1734, 2022.

[99] T. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Robust deepfake on unrestricted media: Generation and detection," CoRR, vol. abs/2202.06228, 2022.

[100] B. M. Le, J. Kim, S. Tariq, K. Moore, A. Abuadbba, and S. S. Woo, "Sok: Facial deepfake detectors," CoRR, vol. abs/2401.04364, 2024.

[101] S. Jia, X. Li, and S. Lyu, "Model attribution of face-swap deepfake videos," in 2022 IEEE International Conference on Image Processing, ICIP 2022, Bordeaux, France, 16-19 October 2022, pp. 2356–2360, IEEE, 2022.

[102] X. Li, R. Ni, P. Yang, Z. Fu, and Y. Zhao, "Artifacts-disentangled adversarial learning for deepfake detection," IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 4, pp. 1658–1670, 2023.

[103] A. Jain, P. Korshunov, and S. Marcel, "Improving generalization of deepfake detection by training for attribution," in 23rd International Workshop on Multimedia Signal Processing, MMSP 2021, Tampere, Finland, October 6-8, 2021, pp. 1–6, IEEE, 2021.

[104] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), 2020.

[105] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, vol. 119 of Proceedings of Machine Learning Research, pp. 1597–1607, PMLR, 2020.

[106] D. Shen, Y. Zhao, and C. Quan, "Identity-referenced deepfake detection with contrastive learning," in IH&MMSec '22: ACM Workshop on Information Hiding and Multimedia Security, Santa Barbara, CA, USA, June 27 - 28, 2022 (B. S. Manjunath, J. Butora, B. Tondi, and C. Vielhauer, eds.), pp. 27–32, ACM, 2022.

[107] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, and R. Ji, "Dual contrastive learning for general face forgery detection," in Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI

2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, pp. 2316–2324, AAAI Press, 2022.

[108] A. Kumar, A. Bhavsar, and R. Verma, “Detecting deepfakes with metric learning,” in 8th International Workshop on Biometrics and Forensics, IWBF 2020, Porto, Portugal, April 29-30, 2020, pp. 1–6, IEEE, 2020.

[109] K. Q. Weinberger and L. K. Saul, “Distance metric learning for large margin nearest neighbor classification,” J. Mach. Learn. Res., vol. 10, pp. 207–244, 2009.

[110] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, “Momentum contrast for unsupervised visual representation learning,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pp. 9726–9735, Computer Vision Foundation / IEEE, 2020.

[111] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” CoRR, vol. abs/1807.03748, 2018.

[112] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pp. 1–11, IEEE, 2019.

[113] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-df: A large-scale challenging dataset for deepfake forensics,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pp. 3204–3213, Computer Vision Foundation / IEEE, 2020.

[114] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, “Fakeavceleb: A novel audio-video multimodal deepfake dataset,” in Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual (J. Vanschoren and S. Yeung, eds.), 2021.

[115] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton-Ferrer, “The deepfake detection challenge dataset,” CoRR, vol. abs/2006.07397, 2020.

[116] 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 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pp. 4360–4369, Computer Vision Foundation / IEEE, 2021.

[117] K. Narayan, H. Agarwal, K. Thakral, S. Mittal, M. Vatsa, and R. Singh, “Df-platter: Multiface heterogeneous deepfake dataset,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 9739–9748, IEEE, 2023.

[118] S. Tomar, “Converting video formats with ffmpeg,” Linux Journal, vol. 2006, no. 146, p. 10, 2006.

[119] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “Retinaface: Single-shot multilevel face localisation in the wild,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pp. 5202–5211, Computer Vision Foundation / IEEE, 2020.

[120] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pp. 18699–18708, IEEE, 2022.

[121] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pp. 2185–2194, Computer Vision Foundation / IEEE, 2021.

[122] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton-Ferrer, "The deepfake detection challenge (DFDC) preview dataset," CoRR, vol. abs/1910.08854, 2019.

[123] C. C. Ferrer, B. Pflaum, J. Pan, B. Dolhansky, J. Bitton, and J. Lu, "Deepfake detection challenge results: An open initiative to advance ai," Meta AI, blog, June, vol. 12, 2020.

[124] B. Huang, Z. Wang, J. Yang, J. Ai, Q. Zou, Q. Wang, and D. Ye, "Implicit identity driven deepfake face swapping detection," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 4490–4499, IEEE, 2023.

[125] M. Goldblum, H. Souri, R. Ni, M. Shu, V. Prabhu, G. Somepalli, P. Chattopadhyay, M. Ibrahim, A. Bardes, J. Hoffman, R. Chellappa, A. G. Wilson, and T. Goldstein, "Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks," in Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[126] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," in 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020.

[127] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," in Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (M. Meila and T. Zhang, eds.), vol. 139 of Proceedings of Machine Learning Research, pp. 10096–10106, PMLR, 2021.

[128] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext V2: co-designing and scaling convnets with masked autoencoders," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 16133–16142, IEEE, 2023.

[129] W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "PVT v2: Improved baselines with pyramid vision transformer," Comput. Vis. Media, vol. 8, no. 3, pp. 415–424, 2022.

[130] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer V2: scaling up capacity and resolution," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pp. 11999–12009, IEEE, 2022.

[131] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in IEEE Conference on Computer Vision and Patten Recognition (CVPR), 2020.

[132] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (dfdc) preview dataset," 2019.

[133] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in neural information processing systems, vol. 33, pp. 6840–6851, 2020.

[134] D. E. King, "Dlib-ml: A machine learning toolkit," The Journal of Machine Learning Research, vol. 10, pp. 1755–1758, 2009.

[135] W. Xian, Z. Li, M. Fisher, J. Eisenmann, E. Shechtman, and N. Snavely, "Uprightnet: geometry-aware camera orientation estimation from single images," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9974–9983, 2019.

[136] A. Ciamarra, R. Caldelli, F. Becattini, L. Seidenari, and A. Del Bimbo, "Deepfake detection by exploiting surface anomalies: The surfake approach," in 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), pp. 1024–1033, 2024.

[137] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al., "Mlp-mixer: An all-mlp architecture for vision," Advances in neural information processing systems, vol. 34, pp. 24261–24272, 2021.

[138] Z. Huang, Z. Zhang, C. Lan, Z.-J. Zha, Y. Lu, and B. Guo, "Adaptive frequency filters as efficient global token mixers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6049–6059, 2023.

[139] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," arXiv preprint arXiv:1904.05441, 2019.

[140] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[141] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al., "Cnn architectures for large-scale audio classification," in 2017 ieee international conference on acoustics, speech and signal processing (icassp), pp. 131–135, IEEE, 2017.

[142] N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection," Computer Speech & Language, vol. 20, no. 2-3, pp. 230–275, 2006.

[143] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in European conference on information retrieval, pp. 345–359, Springer, 2005.

[144] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 6367–6371, IEEE, 2022.

[145] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6369–6373, IEEE, 2021.

[146] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," arXiv preprint arXiv:1907.00501, 2019.

[147] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

[148] R. Hadidi, J. Cao, Y. Xie, B. Asgari, T. Krishna, and H. Kim, "Characterizing the deployment of deep neural networks on commercial edge devices," in 2019 IEEE International Symposium on Workload Characterization (IISWC), pp. 35–48, IEEE, 2019.

[149] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in International conference on machine learning, pp. 2790–2799, PMLR, 2019.

[150] C. Eksombatchai, P. Jindal, J. Z. Liu, Y. Liu, R. Sharma, C. Sugnet, M. Ulrich, and J. Leskovec, "Pixie: A system for recommending 3+ billion items to 200+ million users in real-time," in Proceedings of the 2018 world wide web conference, pp. 1775–1784, 2018.

[151] E. Dufraisse, A. Popescu, J. Tourille, A. Brun, and J. Deshayes, "Mad-tsc: A multilingual aligned news dataset for target-dependent sentiment classification," in 61st Annual Meeting of the Association for Computational Linguistics, 2023.

[152] I. Habernal, R. Hannemann, C. Pollak, C. Klamm, P. Pauli, and I. Gurevych, "Argotario: Computational argumentation meets serious games," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, (Copenhagen, Denmark), pp. 7–12, Association for Computational Linguistics, Sept. 2017.

[153] I. Habernal, P. Pauli, and I. Gurevych, "Adapting Serious Game for Fallacious Argumentation to German: Pitfalls, Insights, and Best Practices," in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), (Miyazaki, Japan), European Language Resources Association (ELRA), May 2018.

[154] Z. Jin, A. Lalwani, T. Vaidhya, X. Shen, Y. Ding, Z. Lyu, M. Sachan, R. Mihalcea, and B. Schölkopf, "Logical fallacy detection," arXiv preprint arXiv:2202.13758, 2022.

[155] 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.), pp. 4143–4149, International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[156] P. Vijayaraghavan and S. Vosoughi, "TWEETSPIN: Fine-grained propaganda detection in social media using multi-view representations," in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Seattle, United States), pp. 3433–3448, Association for Computational Linguistics, July 2022.

[157] T. Alhindi, T. Chakrabarty, E. Musi, and S. Muresan, "Multitask instruction-based prompting for fallacy recognition," in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, (Abu Dhabi, United Arab Emirates), pp. 8172–8187, Association for Computational Linguistics, Dec. 2022.

[158] V. Vorakitphan, E. Cabrio, and S. Villata, “Protect: A pipeline for propaganda detection and classification,” in CLiC-it 2021-Italian Conference on Computational Linguistics, 2022.

[159] S. Sahai, O. Balalau, and R. Horincar, “Breaking down the invisible wall of informal fallacies in online discussions,” in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), (Online), pp. 644–657, Association for Computational Linguistics, Aug. 2021.

[160] S. Haddadan, E. Cabrio, and S. Villata, “Yes, we can! mining arguments in 50 years of US presidential campaign debates,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, (Florence, Italy), pp. 4684–4690, Association for Computational Linguistics, July 2019.

[161] V. Vorakitphan, E. Cabrio, and S. Villata, “”Don’t discuss”: Investigating Semantic and Argumentative Features for Supervised Propagandist Message Detection and Classification,” in RANLP 2021 - Recent Advances in Natural Language Processing, (Varna / Virtual, Bulgaria), Sept. 2021.

[162] P. Goffredo, M. Chaves, S. Villata, and E. Cabrio, “Argument-based detection and classification of fallacies in political debates,” in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 11101–11112, Association for Computational Linguistics, Dec. 2023.

[163] D. Alonso del Barrio and D. Gatica-Perez, “Framing the news: From human perception to large language model inferences,” in Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, ICMR ’23, (New York, NY, USA), p. 627–635, Association for Computing Machinery, 2023.

[164] F. Gilardi, M. Alizadeh, and M. Kubli, “Chatgpt outperforms crowd workers for text-annotation tasks,” Proceedings of the National Academy of Sciences, vol. 120, July 2023.

[165] M. Alizadeh, M. Kubli, Z. Samei, S. Dehghani, J. D. Bermeo, M. Korobeynikova, and F. Gilardi, “Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks,” 2023.

[166] R. M. Entman, J. Matthes, and L. Pellicano, “Nature, sources, and effects of news framing,” The handbook of journalism studies, pp. 175–190, 2009.

[167] V. Bhatia, V. P. Akavoor, S. Paik, L. Guo, M. Jalal, A. Smith, D. A. Tofu, E. E. Halim, Y. Sun, M. Betke, et al., “Openframing: open-sourced tool for computational framing analysis of multilingual data,” in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 242–250, 2021.

[168] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The kaldi speech recognition toolkit,” in IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society, Dec. 2011. IEEE Catalog No.: CFP11SRW-USB.

[169] H. Semetko and P. Valkenburg, “Framing european politics: A content analysis of press and television news,” Journal of Communication, vol. 50, pp. 93 – 109, 06 2000.

[170] D. Alonso del Barrio, M. Tiel, and D. Gatica-Perez, "Human interest or conflict? leveraging llms for automated framing analysis in tv shows," in ACM International Conference on Interactive Media Experiences, IMX '24, (New York, NY, USA), p. 157–167, Association for Computing Machinery, 2024.

[171] V. de Lausanne, "Population," 2020.

[172] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text.," Eighth International Conference on Weblogs and Social Media (ICWSM-14), 2014.

[173] B. Martens, L. Aguiar, E. Gomez-Herrera, and F. Mueller-Langer, "The digital transformation of news media and the rise of disinformation and fake news," SSRN Electronic Journal, 4 2018.

[174] M. Djourelova, R. Durante, and G. Martin, "The impact of online competition on local newspapers: Evidence from the introduction of craigslist," 5 2021.

[175] "Community newspapers and community identity," 2001.

[176] L. Shaker, "Community newspapers play significant role in election," Newspaper Research Journal, vol. 32, pp. 6–18, 1 2011.

[177] M. Yamamoto, "Community newspaper use promotes social cohesion," Newspaper Research Journal, vol. 32, pp. 19–33, 1 2011.

[178] T. I. Nadal Burgers and H. Aladeen, "Bias in news reporting: The potential of ai and natural language processing," in 8th Swiss Text Analytics Conference (SwissText), Mar. 2023.

[179] M. A. Khushi Rakhecha, Simran Rauniyar and A. Bhatt, "A survey on bias detection in online news using deep learning," in 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), pp. 396–403, may 2023.

[180] A. Hobbs, A Fleet Street In Every Town: The Provincial Press in England, 1855-1900. Open Book Publishers. 001.

[181] P. M. Poindexter, D. Heider, and M. McCombs, "Watchdog or good neighbor?: The public's expectations of local news," vol. 11, no. 1, pp. 77–88. 021.

[182] "Sustainability of local journalism," 007.

[183] B. G. Glaser and A. L. Strauss, The Discovery of Grounded Theory: Strategies for Qualitative Research. New York, NY: Aldine de Gruyter, 1967.

[184] N. Fairclough, Critical Discourse Analysis: The Critical Study of Language. Longman applied linguistics, Longman, 2010.

[185] J. Mackenzie, R. Benham, M. Petri, J. R. Trippas, J. S. Culpepper, and A. Moffat, "Cc-news-en: A large english news corpus," in Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, (New York, NY, USA), p. 3077–3084, Association for Computing Machinery, 2020.

[186] U. Zaratiana, N. Tomeh, P. Holat, and T. Charnois, "Gliner: Generalist model for named entity recognition using bidirectional transformer," 2023.

[187] M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure.” 71.

[188] M. Neiger and K. Tenenboim-Weinblatt, “Understanding journalism through a nuanced deconstruction of temporal layers in news narratives,” Journal of Communication, vol. 66, pp. 139–160, 02 2016.

[189] B. Calvo Figueras, T. Caselli, and M. Broersma, “Finding narratives in news flows: The temporal dimension of news stories,” Digital humanities quarterly, vol. 15, no. 15, 2021.

[190] M. G. Constantin, C.-H. Demarty, C. Fosco, G. Healy, B. Ionescu, A. Matran-Fernandez, R. S. Kiziltepe, A. F. Smeaton, et al., “Overview of the mediaeval 2023 predicting video memorability task.,” 2023.

[191] 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, January 2023.

[192] 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, pp. 188–192, 2023.

[193] J. Wan, X. He, and P. Shi, “An iris image quality assessment method based on laplacian of gaussian operation.,” in MVA, pp. 248–251, 2007.

[194] 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), vol. 1, pp. 419–426, IEEE, 2006.

[195] 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, pp. 363–370, Springer, 2003.

[196] 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,” 2021.

[197] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, and F. Wen, “General facial representation learning in a visual-linguistic manner,” in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18676–18688, 2022.

[198] A. Banerjee, U. Bhattacharya, and A. Bera, “Learning Unseen Emotions from Gestures via Semantically-Conditioned Zero-Shot Perception with Adversarial Autoencoders,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 3–10, June 2022. Number: 1.

[199] X. Xu, J. Deng, N. Cummins, Z. Zhang, L. Zhao, and B. W. Schuller, “Exploring Zero-Shot Emotion Recognition in Speech Using Semantic-Embedding Prototypes,” vol. 24, 2022. Conference Name: IEEE Transactions on Multimedia.

[200] C. Zhan, D. She, S. Zhao, M.-M. Cheng, and J. Yang, “Zero-shot emotion recognition via affective structural embedding,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1151–1160, 2019.

[201] F. Qi, X. Yang, and C. Xu, "Zero-shot Video Emotion Recognition via Multimodal Protagonist-aware Transformer Network," in Proceedings of the 29th ACM International Conference on Multimedia, MM '21, (New York, NY, USA), pp. 1074–1083, Association for Computing Machinery, Oct. 2021.

[202] X. Xu, J. Deng, Z. Zhang, Z. Yang, and B. W. Schuller, "Zero-shot speech emotion recognition using generative learning with reconstructed prototypes," in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, IEEE, 2023.

[203] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, et al., "Collecting large, richly annotated facial-expression databases from movies," IEEE multimedia, vol. 19, no. 3, p. 34, 2012.

[204] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu, "Dfew: A large-scale database for recognizing dynamic facial expressions in the wild," in Proceedings of the 28th ACM international conference on multimedia, pp. 2881–2889, 2020.

[205] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang, "Ferv39k: a large-scale multi-scene dataset for facial expression recognition in videos," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20922–20931, 2022.