W-TDL: Window-Based Temporal Deepfake Localization

Luka Dragar
University of Ljubljana
Ljubljana, Slovenia
ld8435@student.uni-lj.si

Peter Rot
University of Ljubljana
Ljubljana, Slovenia
peter.rot@fe.uni-lj.si

Peter Peer
University of Ljubljana
Ljubljana, Slovenia
peter.peer@fri.uni-lj.si

Vitomir Štruc
University of Ljubljana
Ljubljana, Slovenia
vitomir.struc@fe.uni-lj.si

Borut Batagelj
University of Ljubljana
Ljubljana, Slovenia
borut.batagelj@fri.uni-lj.si

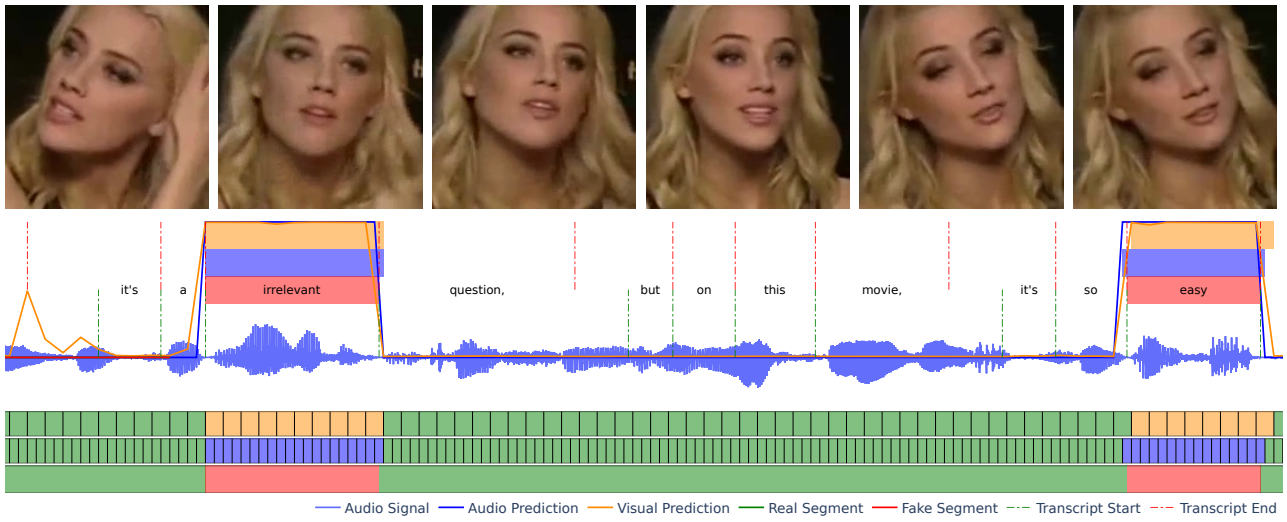

Figure 1: Visualization of audio and video frame predictions

ABSTRACT

The quality of synthetic data has advanced to such a degree of realism that distinguishing it from genuine data samples is increasingly challenging. Deepfake content, including images, videos, and audio, is often used maliciously, necessitating effective detection methods. While numerous competitions have propelled the development of deepfake detectors, a significant gap remains in accurately pinpointing the temporal boundaries of manipulations. Addressing this, we propose an approach for temporal deepfake localization (TDL) utilizing a window-based method for audio (W-TDL) and a complementary visual frame-based model. Our contributions include an effective method for detecting and localizing fake video and audio segments and addressing unbalanced training labels in spoofed audio datasets. Our approach leverages the EVA visual transformer for frame-level analysis and a modified TDL method for audio, achieving competitive results in the 1M-DeepFakes Detection Challenge. Comprehensive experiments on the AV-Deepfake1M dataset demonstrate the effectiveness of our method, providing an effective solution to detect and localize deepfake manipulations.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MRAC '24, October 28-November 1 2024, Melbourne, VIC, Australia
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1203-6/24/10
https://doi.org/10.1145/3689092.3689410

CCS CONCEPTS

• **Applied computing** → *System forensics*; **Investigation techniques**; • **Computing methodologies** → **Computer vision**.

KEYWORDS

Deepfake Detection, Temporal Localization, Audio-Visual Analysis

ACM Reference Format:
Luka Dragar, Peter Rot, Peter Peer, Vitomir Štruc, and Borut Batagelj. 2024. W-TDL: Window-Based Temporal Deepfake Localization. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing (MRAC '24), October 28-November 1 2024, Melbourne, VIC, Australia. Proceedings of the 32nd ACM International Conference on Multimedia (MM'24), October 28-November 1, 2024, Melbourne, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3689092.3689410

1 INTRODUCTION

The quality of synthetic data has reached such a level of realism[8] that differentiating it from real data samples is now increasingly difficult [12]. The inexpensive production of deepfake content, whether in the form of images, videos, or audio, is frequently utilized for malicious purposes, such as spreading misinformation or inflicting targeted reputation damage. As new and increasingly sophisticated techniques for generating deepfakes evolve rapidly [13], detecting deepfakes becomes an increasingly challenging task.

Many competitions have been organized to accelerate the development of deepfake detectors and benchmark state-of-the-art technologies, offering new datasets [2, 7, 11, 16]. However, these initiatives overlook the crucial task of pinpointing the temporal boundaries of manipulations, which is essential to practical applications that require precise detection of when and where manipulations take place within a video or audio.

To address this gap, recent studies [4, 22, 24] and the largest audiovisual data set released in the 1M-DeepFakes Detection Challenge [3] proposed a task called temporal deepfake localization (TDL). In this paper, we present our solution for this challenge, introducing a window-based approach to audio TDL called W-TDL, effectively addressing the problem of unbalanced training labels in spoof datasets with a complementary visual frame-based model that proved highly competitive in the competition. Our contributions can be summarized as follows:

- We propose an effective and highly competitive method for detecting and localizing fake video and audio segments.
- The method effectively addresses the problem of unbalanced training labels in partially spoofed audio datasets.

2 RELATED WORK

Deepfake detection encompasses multiple tasks based on the provided data [15, 19, 25]. We can distinguish among: *(i)* image-, *(ii)* audio-, and *(iii)* video-based deepfake detection. Video-based techniques can be considered a combination of image- and audio-based techniques, where handling a sequence of images with temporal components poses a significantly more computationally demanding task.

Video-based detectors often process individual modalities (i.e., image and audio) separately, analyzing each frame to detect anomalies or inconsistencies, and subsequently, features extracted from both modalities are merged to obtain a final score. After obtaining these characteristics, temporal inconsistencies can be further analyzed (e.g., discrepancies where mouth movements do not align with the word pattern of the audio) [10, 17, 23].

Most video-based deepfake benchmarks only provide labels for entire videos, indicating whether they are genuine or fake. To address this, a new dataset called LAV-DF was recently proposed in the work of [4] to enable benchmarking deepfakes at the localization level, where the goal is to identify the specific segments that were manipulated. Additionally, the authors introduced BA-TFD+, which uses a Multiscale Vision Transformer as its backbone and is trained with contrastive learning.

Xie et al. [22] proposed a partially spoofed audio detection method known as Temporal Deepfake Location (TDL). It introduced two key components: an embedding similarity module to enhance identification between real and fake features and a temporal convolution operation to focus on position information. This method proved to be highly effective on ASVspoof2019 Partial Spoof dataset [21].

3 METHODS

3.1 Overview

Our approach employed two distinct methodologies for visual and audio components. For the visual analysis, we utilized the EVA visual transformer, leveraging its robust image classification capabilities to identify tampered frames within videos. For the audio analysis, we implemented a modified Temporal Deepfake Location (W-TDL) method to accurately detect and localize deepfake segments within audio streams. By combining these two techniques, we aimed to achieve comprehensive and precise forgery detection across visual and auditory data. The following sections detail the specific methodologies and adaptations used for each modality.

3.2 Visual EVA

We employed an image frame-based approach using the EVA visual transformer [9] to detect visually manipulated video frames. EVA, a vanilla Vision Transformer (ViT), is pre-trained to reconstruct masked-out, image-text-aligned vision features based on visible image patches. We selected EVA due to its exceptional performance on ImageNet classification benchmarks and its native resolution of 224×224, which perfectly aligns with our dataset. This alignment ensures that no information is lost during processing, eliminating the need for image rescaling. Specifically, we utilized the `eva_giant_patch14_224.clip_ft_in1k` model from the Timm image library pre-trained on ImageNet 1k. Then, we fine-tuned this model for the binary classification task of predicting whether a given frame is real or tampered, using Cross-Entropy loss as the learning objective.

3.3 Audio W-TDL

A technique called Temporal Deepfake Location (TDL) [22] was used to identify audio deepfake segments. It's composed of a wav2-vec2-XLS-R300M [1] frontend as a feature extractor, which combines these features across audio frames using temporal convolution operations. The method incorporates an embedding similarity module to distinguish real and fake frames in an embedding space. It was chosen due to its impressive performance on the ASVspoof2019 Partial Spoof dataset [21].

We adapted and modified the TDL method to a window-based approach (W-TDL), addressing the limitations of the original frame-level prediction method. By adapting the audio model to process windows, we overcame issues related to imbalanced training labels and eliminated the need for padding. Each window consists of 64 wav2vec2 feature vectors, corresponding to 1.28 seconds of audio. This setup provides a resolution of 20ms, which aligns precisely with most of the audio fake segment labels, ensuring accurate detection and localization of deepfakes.

Table 1: Dataset Split

Splits	Number of Videos	Percentage
Training	746,180	74.62%
Validation	57,340	5.73%
Test	196,480	19.65%

4 EXPERIMENTS

4.1 AV-Deepfake1M Dataset

The AV-Deepfake1M dataset [3] is a large-scale audio-visual deepfake dataset designed to advance state-of-the-art deepfake detection and localization. It was generated using a multistage pipeline, leveraging a subset of real videos from Voxceleb2 [6].

The pipeline begins with extracting transcripts from the real videos using Whisper [18]. Then a Large Language Model (LLM) is utilized to propose modifications to the transcripts, aiming to alter their meaning. These modifications can take one of three forms: "delete," "insert," or "replace," with "replace" being the most commonly used operation at 92.2% out of them all.

The modified transcripts are then used to generate artificial audio using two distinct text-to-speech methods: VITS [14], which is identity-dependent, and YourTTS [5], which is identity-independent. Finally, the visual frames are generated using TalkLip [20], a state-of-the-art method that leverages the subject's original pose and the newly generated fake audio to produce lip-synchronized fake visual frames that align with the input audio.

Notably, this pipeline can produce a variety of high-quality content-driven deepfake and real videos that are together categorized into four distinct types:

(1) **Real**: Non modified videos.
(2) **Fake Audio and Fake Visual**: Both audio and visual frames are manipulated.
(3) **Fake Audio and Real Visual**: Only real audio corresponding to replacements and deletions are manipulated, with synchronized fake audio and real visual segments.
(4) **Real Audio and Fake Visual**: Only visual frames are manipulated, with the original audio remaining unchanged.

The dataset comprises a total of 2,068 unique subjects, with each of the four distinct types of manipulations evenly represented. This balanced distribution results in a comprehensive collection of one million videos. These videos are partitioned subject-wise into training, validation, and test sets (Table 1).

4.2 Video Manipulation

TalkLip's face detection and extraction process involves taking a frame and identifying a bounding box around the person's face. This box is then resized to compact 96×96 pixels, ready for further processing in the pipeline. The new frames, now lip-synced to the modified audio, are generated at this resolution. Finally, the generated frame is resized back to its original bounding box size, ensuring the pixels in the original frame are replaced at that location.

This extraction, rescaling, and generation at lower resolution results in the generated region being visibly lower quality than the rest of the surrounding image, producing visible boundary lines at the transition between the original and generated pixels. However, when the frames are encoded back into a video, the compression hides most of these boundaries, resulting in fakes that are harder to detect.

4.3 Audio Manipulation

VITS [14] and YourTTS [5] generally produce high-quality fake audio. Additionally, the creators ensured that the generated audio included background noise similar to real recordings. They achieved this by using a denoiser to separate the noise from speech and adding the same noise to the generated audio. Finally, they applied loudness normalization to further enhance the result.

Still, the generation method isn't perfect. To determine points of change, Whisper word transcript timestamps are used to find the start and end points of where to insert the fake segment. As mentioned in the Whisper paper [18], this introduces bias since the Whisper resolution is at 0.02 seconds. Their timestamp prediction predicts time relative to the current audio segment, quantizing all times to the nearest 20 milliseconds, which matches the native time resolution of Whisper models. We've found a strong tendency for the fake segment to be divisible by 0.02 in terms of its length and start and end points. In the training set, there is a 97.6% chance it will start or end on a multiple of 0.02 and an 86.73% chance its length will be a multiple of 0.02.

This quantization also influenced the transition between real and fake segments. This shows in each fake segment (audio and visual) starting and ending with a silence of about 0.01s in length. This is observable in Figure 1, as the audio signal always flattens at the beginning and end of the fake segments.

4.4 Task: 1 Video-Level Deepfake Detection

The goal of the first task of the 1M-Deepfakes Detection Challenge[1] was to distinguish between real, unmodified videos and fake tampered videos. The training was restricted to only having access to video-level labels, meaning no segment-level labels could be used. In theory, this should pose a greater challenge, as the model or algorithm should use the entire video for training to even capture the fake segments.

However, we observed that the real and modified videos can be distinguished based on the encoder version used to encode them, which explains why, during the training of our visual models, we noticed a peculiar anomaly: our visual model showed an unexpected proficiency in differentiating between unmodified real videos and audio-manipulated videos, a task that should have been beyond its capabilities.

Upon examination, we discovered that the counterfeit videos in the generation pipeline had been re-encoded using `ffmpeg`, with a distinct version of the encoder employed in this process. This re-encoding process introduced visual artifacts in the video frames, which our model subsequently learned. Additionally, the videos' metadata bore a distinct mark, confirming the re-encoding.

Using `ffprobe`, we can see that fake videos use encoder version `Lavf58.45.100` while the real ones use `Lavf57.83.100`.

[1] https://deepfakes1m.github.io/

Table 2: Performance of Models and Merging Strategies on Validation Set

Model/Strategy	Subset	mAP	mAR	Score
W-TDL	A	0.87	0.92	0.89
EVA	V	0.58	0.83	0.70
Audio Localization	F	0.58	0.89	0.74
Overlap Merge	F	0.64	0.90	0.77
Basic Merge	F	0.70	0.91	0.80
Lower Visual conf. upon Audio	F	0.80	0.91	0.85

This meant that the real and fake videos could be easily distinguishable and presented problems when using the real video subset for training the visual model on task 2.

4.5 Task 2: DeepFake Temporal Localization

In task 2 the goal was to predict the exact start and end timestamps of fake segments within videos. Here, competitors had access to frame-level labels. The metrics for this task were AP (Average Precision) and AR (Average Recall) at N most confident detections. The final score was then calculated as follows:

$$\text{Score} = \frac{1}{8} \sum_{\text{IoU} \in \{0.5,0.75,0.9,0.95\}} \text{AP@IoU} + \frac{1}{10} \sum_{N \in \{50,30,20,10,5\}} \text{AR@N}.$$

Additionally, the metric implemented in the code has a FPS (frames per second) parameter that is used to convert the timestamps into frame indices, ensuring consistent and precise temporal alignment for accurate IoU (Intersection over Union) calculation. This conversion allows for meaningful evaluation and comparison of proposals and ground-truth segments across videos with different frame rates.

This is especially noticeable with visual model predictions that are limited to a resolution of 25 FPS since that is the frame rate of the videos. Audio, on the other hand, can be more precise. Our audio model operates at 50 FPS. Generally, frame-level labels were provided at a resolution of 100 FPS, but as mentioned in Subsection 4.3, most of them fall on 50 FPS. We chose to evaluate our models at 50 FPS for validation.

One important consideration is the sensitivity of the metric, as most fake segments are short, with an average length of 0.326s, or 16 frames. Therefore, a small one-frame error in the prediction reduces the IoU more quickly compared to the same error in a longer segment. This indicates that precise localization is more crucial for shorter segments because a small deviation in a short segment's predicted start or end frame constitutes a larger proportion of its total duration.

4.6 Training and Data Selection

For our first experiment for task 2, we trained a visual model on frames from real segments and frames from fake segments in visually modified videos. As mentioned in Section 4.4, this caused the model to learn encoder artifacts, and the frame-level predictions became noisy.

Because of that, we used non-modified and tampered frames from visually modified videos instead, with an added 15% of the real videos. This results in 233,257 training videos. At each pass, a random frame is chosen from the video.

For the audio model, the selection was more complex. We used the audio-modified and real videos from the dataset and selected windows of 64 feature vectors or 1.28s at each fake segment. The algorithm considers the start point of the fake segments and then creates a window by taking 64 vectors to the right from the start. If there are not enough vectors to the right, the window is shifted to the left accordingly. For real videos, we chose the middle 64 vectors. This approach greatly contributed to more balanced labels by increasing the percentage of fake frames from 3.7% to 14.4%.

We used High-Performance Computing (HPC) with Pytorch Lightning for training. The loss function was Binary Cross-Entropy, with the audio model including an additional embedding loss weighted at 0.1. AdamW was the optimizer, with learning rates of 2e-5 for the visual model and 1e-4 for the audio model, and CosineAnnealingLR was used as the learning rate scheduler.

The visual model had a batch size of 4, effectively 64 when training on 4 nodes with 4 GPUs using the Distributed Data Parallel (DDP) strategy. The audio model had a batch size of 128 on one GPU. Based on validation loss monitoring, training concluded at epoch 57 for the visual model and epoch 37 for the audio model.

4.7 Segment Extraction

Both of these model output per-frame predictions at their respective FPS. The audio model predictions are made by moving the window through the frames. Finally, the remaining window is computed by moving the window to the end and removing the duplicate predictions that arise after the left boundary.

To extract segments from these per-frame predictions, an algorithm creates a binary mask based on a threshold. It then creates the segments with their confidence being the average of their frame predictions. Post-processing operations remove short segments and low-confidence segments, with all thresholds determined by the validation set performance. Finally, a conversion factor is used to get the predictions in seconds.

5 RESULTS

Several merging strategies were tested, including simply combining both predictions (Basic Merge), using the audio predictions when visual and audio segments overlap (Audio Localization), employing a union approach when both overlap (Overlap Merge), and reducing the confidence of visual segments where audio segments were found (Lower Visual conf. upon Audio). The latter proved to be the best.

We optimized our models on the validation set of the dataset. Performance metrics for each model on different subsets are summarized in Table 2. The subsets are defined as follows: V (videos containing visually modified segments), A (audio segments), and F (full dataset), with the metrics being summarized as Mean Average Precision (mAP) and Mean Average Recall (mAR). We observe that the audio model achieves far greater results at precise temporal localization partly because it can produce scores at a greater resolution (see Figure 1) and is unaffected by video compression; the most optimal merging strategy confirms this. Additionally, the audio model is significantly faster, processing each sample in approximately 0.2563 seconds compared to the video model's 6.853 seconds

Table 3: Results on Competition Test Set

Method	Mod.	AP@0.5	AP@0.75	AP@0.9	AP@0.95	AR@50	AR@30	AR@20	AR@10	AR@5	Score
BA-TFD+ [4]	AV	44.42	13.64	00.48	00.03	48.86	44.51	40.37	34.67	29.88	0.2715
UMMAFormer [24]	AV	51.64	28.07	07.65	01.58	44.07	43.93	43.45	42.09	40.27	0.3249
EVA&W-TDL	AV	**94.75**	**88.75**	**70.43**	**50.66**	**89.17**	**89.17**	**89.17**	**89.13**	**88.78**	**0.8262**

per sample. This makes the audio model much more suitable for large-scale applications, where speed and efficiency are crucial.

Finally, we evaluate our approach on the competition test set, obtaining a score of 0.8262, which vastly outperforms the current baseline methods tested by the organizers, as shown in Table 3. AP (Average Precision) measures how well the predicted fake segments are localized. For example, at a threshold of 0.75 IoU, which indicates a moderate overlap between the predicted and actual fake segments, our model achieves a precision of 88.75%, demonstrating its effectiveness in accurately pinpointing the location of fake segments. Furthermore, the consistently high AR (Average Recall) demonstrates that our model effectively captures most of the actual fake segments, ensuring thorough detection without generating too many unnecessary proposals. These results could be further improved by using a sliding window approach instead of a moving window and by implementing preprocessing steps to remove noise from the audio, which currently causes some false detections.

6 CONCLUSION

In this paper, we present our solution for the 1M-DeepFakes Detection Challenge, focusing on the task of Temporal Deepfake Localization. Our approach integrates a window-based method for audio deepfake detection (W-TDL) with a visual frame-based model (EVA) to effectively identify and localize manipulated segments in audio and visual data. Through competition and experiments, we demonstrate that our method outperforms existing state-of-the-art techniques on the AV-Deepfake1M dataset.

However, the proposed dataset has some limitations. Despite its size, the generation pipeline is limited to only one visual generation method and an audio stitching method is limited by transcript resolution, providing observable boundaries between fake and real speech. Furthermore, real and fake videos are trivially distinguishable by the encoder versions used in their encoding.

Future work would include creating more sophisticated and diverse generation pipelines that create more seamless transitions between fake and real segments, while also improving our detection methods by implementing audio-visual feature fusion. We will also explore better metrics to more accurately evaluate the performance of these advanced models, focusing on methods that are less dependent on the size of the segments to ensure fairer and more consistent assessments.

ACKNOWLEDGMENTS

The research presented in this paper was supported by the Slovenian Research and Innovation Agency ARIS as part of the research project J2-50065 DeepFake DAD, and ARIS programmes P0-0250 and P2-0214.

REFERENCES

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *CoRR* abs/2006.11477 (2020). arXiv:2006.11477 https://arxiv.org/abs/2006.11477

[2] Marko Brodarič, Vitomir Štruc, and Peter Peer. 2024. Cross-dataset deepfake detection: evaluating the generalization capabilities of modern deepfake detectors. In *Proceedings of the 27th Computer Vision Winter Workshop (CVWW 2024)*. Slovensko društvo za razpoznavanje vzorcev = Slovenian Pattern Recognition Society, 47–56. https://cvww2024.sdrv.si/proceedings/

[3] Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, and Kalin Stefanov. 2023. AV-Deepfake1M: A Large-Scale LLM-Driven Audio-Visual Deepfake Dataset. *arXiv preprint arXiv:2311.15308* (2023).

[4] Zhixi Cai, Shreya Ghosh, Abhinav Dhall, Tom Gedeon, Kalin Stefanov, and Munawar Hayat. 2023. Glitch in the matrix: A large scale benchmark for content driven audio–visual forgery detection and localization. *Computer Vision and Image Understanding* 236 (2023), 103818.

[5] Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Cândido Júnior, Eren Gölge, and Moacir Antonelli Ponti. 2021. YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone. *CoRR* abs/2112.02418 (2021). arXiv:2112.02418 https://arxiv.org/abs/2112.02418

[6] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. *CoRR* abs/1806.05622 (2018). arXiv:1806.05622 http://arxiv.org/abs/1806.05622

[7] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397* (2020).

[8] Luka Dragar, Peter Peer, Vitomir Štruc, and Borut Batagelj. 2023. Beyond detection: visual realism assessment of deepfakes. In *Proceedings of the 32nd International Electrotechnical and Computer Science Conference ERK 2023*. Slovenska sekcija IEEE; Fakulteta za elektrotehniko, Portorož, Slovenija, 363–366. https://erk.fe.uni-lj.si/2023/papers/dragar(beyond_detection_).pdf

[9] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19358–19369.

[10] David Güera and Edward J Delp. 2018. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 1–6.

[11] Marco Huber, Fadi Boutros, Anh Thi Luu, Kiran Raja, Raghavendra Ramachandra, Naser Damer, Pedro C. Neto, Tiago Gonçalves, Ana F. Sequeira, Jaime S. Cardoso, João Tremoço, Miguel Lourenço, Sergio Serra, Eduardo Cermeño, Marija Ivanovska, Borut Batagelj, Andrej Kronovšek, Peter Peer, and Vitomir Štruc. 2022. SYN-MAD 2022: Competition on Face Morphing Attack Detection Based on Privacy-aware Synthetic Training Data. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*. 1–10. https://doi.org/10.1109/IJCB54206.2022.10007950

[12] Sahar Husseini and Jean-Luc Dugelay. 2023. A Comprehensive Framework for Evaluating Deepfake Generators: Dataset, Metrics Performance, and Comparative Analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 372–381.

[13] Marija Ivanovska and Vitomir Štruc. 2024. On the vulnerability of deepfake detectors to attacks generated by denoising diffusion models. In *Proceedings of the 27th Computer Vision Winter Workshop (CVWW 2024)*. CPS; IEEE Computer Society, 1051–1060. https://ieeexplore.ieee.org/document/10495703

[14] Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. *CoRR* abs/2106.06103 (2021). arXiv:2106.06103 https://arxiv.org/abs/2106.06103

[15] Yisroel Mirsky and Wenke Lee. 2021. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)* 54, 1 (2021), 1–41.

[16] Bo Peng, Xianyun Sun, Caiyong Wang, Wei Wang, Jing Dong, Zhenan Sun, Rongyu Zhang, Heng Cong, Lingzhi Fu, Hao Wang, et al. 2023. DFGC-VRA: DeepFake Game Competition on Visual Realism Assessment. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 1–9.

[17] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, Yang Liu, and Jianjun Zhao. 2020. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM international conference on multimedia*. 4318–4327.

[18] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 1182, 27 pages.

[19] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. 2022. Deepfake detection: A systematic literature review. *IEEE access* 10 (2022), 25494–25513.

[20] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. 2023. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14653–14662.

[21] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sébastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-François Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling. 2020. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech and Language* 64 (2020), 101114. https://doi.org/10.1016/j.csl.2020.101114

[22] Yuankun Xie, Haonan Cheng, Yutian Wang, and Long Ye. 2023. An Efficient Temporary Deepfake Location Approach Based Embeddings for Partially Spoofed Audio Detection. arXiv:2309.03036 [cs.SD] https://arxiv.org/abs/2309.03036

[23] Peipeng Yu, Zhihua Xia, Jianwei Fei, and Yujiang Lu. 2021. A survey on deepfake video detection. *Iet Biometrics* 10, 6 (2021), 607–624.

[24] Rui Zhang, Hongxia Wang, Mingshan Du, Hanqing Liu, Yang Zhou, and Qiang Zeng. 2023. UMMAFormer: A Universal Multimodal-adaptive Transformer Framework for Temporal Forgery Localization. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. ACM. https://doi.org/10.1145/3581783.3613767

[25] Tao Zhang. 2022. Deepfake generation and detection, a survey. *Multimedia Tools and Applications* 81, 5 (2022), 6259–6276.