SHORT-PAPER

1M-Deepfakes Detection Challenge

ZHIXI CAI, Monash University, Melbourne, VIC, Australia

ABHINAV DHALL, Flinders University, Adelaide, SA, Australia

SHREYA GHOSH, Curtin University, Perth, WA, Australia

MUNAWAR HAYAT, Qualcomm Incorporated, San Diego, CA, United States

DIMITRIOS KOLLIAS, Queen Mary University of London, London, U.K.

KALIN STEFANOV, Monash University, Melbourne, VIC, Australia

View all

.

1M-Deepfakes Detection Challenge

Zhixi Cai
Monash University
Melbourne, Australia
zhixi.cai@monash.edu

Abhinav Dhall
Flinders University
Adelaide, Australia
abhinav.dhall@flinders.edu.au

Shreya Ghosh
Curtin University
Perth, Australia
shreya.ghosh@curtin.edu.au

Munawar Hayat
Qualcomm
San Diego, United States
hayat@qti.qualcomm.com

Dimitrios Kollias
Queen Mary University of
London
London, United Kingdom
d.kollias@qmul.ac.uk

Kalin Stefanov
Monash University
Melbourne, Australia
kalin.stefanov@monash.edu

Usman Tariq
American University of
Sharjah
Sharjah, United Arab
Emirates
utariq@aus.edu

Abstract

The detection and localization of deepfake content, particularly when small fake segments are seamlessly mixed with real videos, remains a significant challenge in the field of digital media security. Based on the recently released AV-Deepfake1M dataset, which contains more than 1 million manipulated videos across more than 2,000 subjects, we introduce the 1M-Deepfakes Detection Challenge. This challenge is designed to engage the research community in developing advanced methods for detecting and localizing deepfake manipulations within the large-scale high-realistic audio-visual dataset. The participants can access the AV-Deepfake1M dataset and are required to submit their inference results for evaluation across the metrics for detection or localization tasks. The methodologies developed through the challenge will contribute to the development of next-generation deepfake detection and localization systems. Evaluation scripts, baseline models, and accompanying code will be available on https://github.com/ControlNet/AV-Deepfake1M.

CCS Concepts

• **Computing methodologies** → Computer vision; • **Security and privacy** → Social aspects of security and privacy.

Keywords

Datasets, Deepfake, Localization, Detection

1 Introduction

Rapid advances in generative AI have significantly transformed the landscape of digital content creation, making the generation and manipulation of video [13, 43, 48] and audio [18, 19, 40] more accessible and efficient than ever before. Although these technologies have enabled many positive applications in various domains, they have also created deepfakes, highly realistic but artificially manipulated media that can misrepresent people without their consent [29, 58]. Deepfakes pose serious ethical and security concerns, as they can change individuals saying or doing things they never actually did, leading to widespread misinformation, disinformation, and even malicious activities such as online harassment and fraud [2, 36, 38, 44]. The effectiveness of deepfake detection and localization methods is highly dependent on the datasets. Although the past few years have seen an increase in publicly available datasets focused on visual-only [17, 23, 26], audio-only [27, 53], and audio-visual [20] content manipulations, most of these datasets assume that the entire content is either real or fake. This presents a significant limitation as it overlooks the growing trend of embedding small, subtle manipulations within otherwise real content. Such clever manipulations can completely alter the meaning of content, as discussed in [7], yet they are not adequately addressed by existing benchmark datasets and challenges.

To overcome the gap, the AV-Deepfake1M dataset [4] was introduced, providing a large-scale benchmark of audio-visual content-based deepfake videos for the task of temporal deepfake localization. Based on this dataset, the 1M-Deepfakes Detection Challenge not only focuses on the binary classification of deepfake content, but also emphasizes the task of localization, which means identifying the specific timestamps within a video where manipulations have occurred. The challenge is planned to contribute to improve current detection methods and aims to run as an ongoing benchmarking for the next several years, continually introducing new challenges of deepfake technology to keep pace with its rapid evolution.

2 Related Work

Research on deepfake detection has progressed significantly due to the development of key datasets created using various manipulation techniques. Figure 1 outlines these relevant datasets. Korshunov and
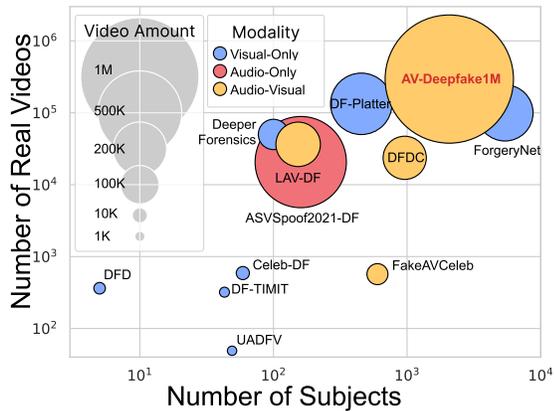
Figure 1: Comparison of related datasets with AV-Deepfake1M. This figure illustrates a comparison of AV-Deepfake1M with other accessible datasets, highlighting the number of subjects and the quantity of real versus fake videos. The figure is reproduced from the AV-Deepfake1M paper.

Marcel [22] were among the first to curate a deepfake dataset, DF-TIMIT, which involved face swapping on the Vid-Timit dataset [37]. Subsequently, other notable datasets such as UADFV [51], Face-Forensics++ [35], and Google DFD [30] were developed. These datasets are relatively small due to the complexity of face manipulation and the limited availability of open-source manipulation techniques [26].

In 2020, Facebook introduced the large-scale DFDC dataset [11] for deepfake classification, comprising 128,154 videos created using various manipulation methods and featuring real videos of 3,000 actors. DFDC has since become a standard benchmark for deepfake detection. Following DFDC, several new datasets were released, including Celeb-DF [26], DeeperForensics [17], and Wild-DeepFake [59], which focus on the binary task of deepfake classification and primarily target visual manipulation detection [9].

In 2021, the OpenForensics dataset [24] was introduced for spatial detection, segmentation, and classification. More recently, FakeAVCeleb [20] was released, addressing both face-swapping and face-reenactment methods with manipulated audio and visual modalities. ForgeryNet [15] is one of the latest additions to deepfake detection datasets, focusing on visual-only identity manipulations and supporting video/image classification as well as spatial and temporal forgery localization tasks.

In 2022, LAV-DF [7] was introduced as the first content-driven deepfake dataset specifically designed for temporal localization. However, LAV-DF suffers from limitations in both quality and scale. Meanwhile, state-of-the-art methods for temporal localization [5, 57] are already demonstrating very strong performance. AV-Deepfake1M [4] addresses these limitations by enhancing the quality, diversity, and scale of the datasets available for temporal deepfake localization.

3 Challenge Description

3.1 Dataset

The challenge is based on AV-Deepfake1M dataset [4]. The AV-Deepfake1M dataset is a large-scale audio-visual deepfake dataset,
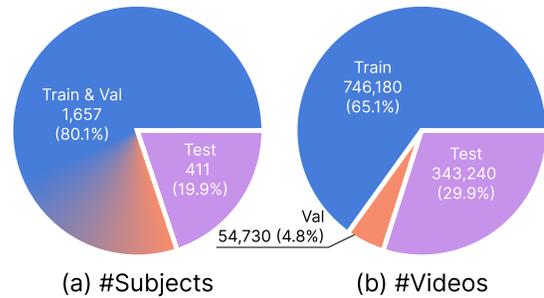

Figure 2: Data partitioning in AV-Deepfake1M. (a) The count of subjects within the *train*, *validation*, and *test* sets. (b) The count of videos present in the *train*, *validation*, and *test* sets. The figure is adapted from the AV-Deepfake1M paper.

containing 1,886 hours of audio-visual data from 2,068 unique subjects with diverse background settings. This dataset is positioned as the most extensive audio-visual benchmarking effort, as shown in Figure 1. The 1M-Deepfakes Detection Challenge involves two main tasks:

Deepfake Detection. This task requires identifying whether a given audio-visual sample of a single subject is deepfake or real.
Deepfake Temporal Localization. This task involves determining the specific time intervals within an audio-visual sample of a single subject that is manipulated or fake.

3.2 Data Partitioning

To ensure a robust benchmark for the submitted models, we divided our dataset into train, validation, and test sets. We randomly allocated 1,657 subjects to the train set and 411 subjects to the test set, ensuring no overlap between these groups. The validation set was randomly selected from the train set based on the video level rather than the identity level. For the test set, we ensured that all audio manipulations were generated using high-quality VITS models. Finally, the distribution of the train, validation, and test sets is 65%, 5%, and 30%, respectively (refer to Figure 2 for more details).

3.3 Evaluation Metrics

3.3.1 Deepfake Detection. Deepfake detection is calculated as video label prediction. For deepfake detection/classification, we use the area under the ROC curve (AUC). The value is in the range of [0.5, 1] where 0.5 denotes random guesses and 1 perfect prediction.

3.3.2 Deepfake Temporal Localization. For deepfake temporal localization, we use average precision (AP) and average recall (AR) in line with previous works [3, 7]. We set the intersection over union (IoU) thresholds at 0.5, 0.75, 0.9, and 0.95 for AP, and [0.5:0.95:0.05] for AR. For the number of proposals for AR, we select 50, 30, 20, 10, and 5. Finally, we average all AP metrics and all AR metrics to compute the overall average values for AP and AR. The average values of these metrics are then used to obtain the final results. The score ranges from 0 to 1, where 0 indicates random guesses and 1 represents perfect predictions. The final score S is calculated as:

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

Figure 3: Temporal deepfake localization benchmark. This figure compares the performance of state-of-the-art methods on the AV-Deepfake1M dataset.

Figure 4: Deepfake Detection Benchmark. Comparison of state-of-the-art method performance on the AV-Deepfake1M dataset across various evaluation protocols.

3.4 Benchmark

The results of the off-the-shelf models for deepfake temporal localization [1, 5, 7, 10, 32, 41, 46, 47, 55, 57] and deepfake detection [1, 6, 8–10, 14, 25, 42, 45, 49, 56] on the AV-Deepfake1M [4] dataset are depicted in Figure 3 and Figure 4.

4 Challenge Participation Details

In the 1M-Deepfakes Detection Challenge 2024, 191 teams signed the EULA and registered for the challenge. The evaluation was carried out on CodaBench [50], facilitating automated participation, submission, and evaluation. By the submission deadline, there were 1034 successful submissions, indicating significant community interest. The number of submissions for Task 1 (detection/classification) and Task 2 (temporal localization) are not balanced. Specifically, the number of submissions to Task 1 is 51% higher than those to Task 2. This imbalance suggests that the community focus is mainly on deepfake detection, thereby overlooking the significant threat posed by deepfakes that are not fully fabricated. The latest work on deepfake detection methodologies further supports this trend [12, 16, 31, 34, 39, 52, 54].

4.1 Challenge Track 1 (Detection/Classification)

In this track, the top 3 teams are Fake-detector (USTC), UQCV (UQ) and FRIdatas (FRI). The Fake-detector (USTC) team mainly

introduced the Audio-Visual Local-Global Interaction Module (AV-LG Module), which consists of Local Intra-Region Self-Attention, Global Inter-Region Self-Attention, and Local-Global Interaction, enhancing the fine-grained perception and cross-modal interaction capabilities of the model on deepfakes detection. The key method for UQCV (UQ) team is weakly supervised learning via pseudo-labeling. Using audio-based and mouth region-based pseudo-labeling, the proposed method gradually filters out high-confidence groups and adds those pseudo-labeled samples to the training set to fine-tune the network until all predictions are high-confidence. Finally, the third-ranked team, FRIdatas (FRI) has developed an audio-visual method to detect fake chunks in the video. A frame-based approach with a vision transformer model has been incorporated in a manner that mitigates the impact of different encoder versions in real and fake videos. For audio, Wav2Vec-XLS-R features with an embedding module have been incorporated with temporal convolution operations.

4.2 Challenge Track 2 (Temporal Localization)

In this track, the top 3 teams are Gradiant (Gradiant) [33], Maya (USTC) and nudt24 (NUDT). The top team, Gradiant, employs the gMLP method for audio features and complements it with a pathway that processes optical flow and RGB features using a UMMAFormer architecture. The outputs from both pathways are merged to boost performance. The Maya (USTC) team utilizes the BYOL-A model

to extract audio features and both TSN and InternVideo models to extract visual features. The BYOL-A audio features and the InternVideo visual features were used for model training in the video-level deepfake detection task, while the BYOL-A audio features and TSN visual features were used for the deepfake temporal localization task. Finally, the third-ranked team nudt24 (NUDT) proposes a Task Hierarchical Emotion-aware for Fake Detection (THE-FD) architecture that initially handles video-level data (Task 1) and subsequently naturally adapts to frame-level data (Task 2). The benefit of this approach is it allows for inheritance and sharing of features while utilizing a shared structure.

5 Research Impact

Existing deepfakes detection datasets (DFDC [11], DF-TIMIT [22], Celeb-DF [26], etc.) have made a seminal contribution to the progress in deepfakes detection. The current set of datasets is limited by their size and the range of tasks. The latter refers to their focus on the binary classification task of deepfakes detection only. For progressing the research in deepfake detection it is important to learn and evaluate the detectors on varied data representing a large number of subjects, situations, and contexts. A comparison of the AV-Deepfake1M with the existing datasets is presented in Figure 1 (for a detailed comparison, please refer to AV-Deepfake1M [4]). The assumption that manipulation constitutes the entirety of a video has been shown to fail based on deepfake videos on social media, where a segment of the video is deepfaked. The 1M-Deepfakes Detection Challenge will accelerate research in deepfakes analysis through the availability of both detection and localization tasks.

6 Conclusion and Future Directions

We introduced the 1M-Deepfakes Detection Challenge 2024, the first challenge addressing content-driven deepfake detection and localization in well-defined conditions. We presented publicly available datasets and evaluation protocols for both tasks and evaluated baseline approaches. The evaluation server will remain accessible to researchers beyond the 1M-Deepfakes Detection Challenge 2024 deadline, contributing to continuing progress on both tasks.

The task of identifying deepfakes goes beyond the typical classification approach. In this work, we explore both classification and localization, however, there are important aspects of deepfake analysis for future work. A major challenge in deepfakes comes from the variability due to different **cultures and languages**. From an observer's perspective, detecting deepfakes in non-native languages is non-trivial [21]. Typically, foreign language videos are accessed and understood through voice-overs and native language subtitles. This makes the task of deepfake detection more challenging, as the semantic context and nuances of a particular culture or language may be missed by an observer. For a deepfake detector, scenarios like this require datasets containing content in different languages and methods that are agnostic to languages.

Beyond Faces: The current literature mainly focuses on the facial region in a video, as most deepfakes involve facial manipulation. The perceived meaning of a video can be altered for an observer with a simple edit to the non-verbal gestures of the subject's face and body. There is a need to explore this line of work. An example is the detection of sign language-based deepfakes [28]. On the other hand, in scenarios involving groups of people, changing the group structure and/or the location of group members can lead to misinterpretation. Understanding group structures and the relationships between group members will be necessary to detect deepfakes in such conditions.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. 1–7. ISSN: 2157-4774.

[2] John Brandon. 2019. There Are Now 15,000 Deepfake Videos on Social Media. Yes, You Should Worry. *Forbes* (Oct. 2019).

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 961–970.

[4] 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:2311.15308 [cs].

[5] 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 (Nov. 2023), 103818.

[6] Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. 2023. MARLIN: Masked Autoencoder for Facial Video Representation LearnINg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Vancouver, BC, Canada, 1493–1504.

[7] Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. 2022. Do You Really Mean That? Content Driven Audio-Visual Deepfake Dataset and Multimodal Method for Temporal Forgery Localization. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. Sydney, Australia, 1–10.

[8] Francois Chollet. 2017. Xception: Deep Learning With Depthwise Separable Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1251–1258.

[9] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. 2020. Not made for each other- Audio-Visual Dissonance-based Deepfake Detection and Localization. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. Association for Computing Machinery, New York, NY, USA, 439–447.

[10] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. 2022. Combining EfficientNet and Vision Transformers for Video Deepfake Detection. In *Image Analysis and Processing – ICIAP 2022 (Lecture Notes in Computer Science)*, Stan Sclaroff, Cosimo Distante, Marco Leo, Giovanni M. Farinella, and Federico Tombari (Eds.). Springer International Publishing, Cham, 219–229.

[11] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The DeepFake Detection Challenge (DFDC) Dataset. arXiv: 2006.07397 [cs].

[12] Chao Feng, Ziyang Chen, and Andrew Owens. 2023. Self-Supervised Video Forensics by Audio-Visual Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10491–10503.

[13] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. 2022. Long Video Generation with Time-Agnostic VQGAN and Time-Sensitive Transformer. In *Proceedings of the European Conference on Computer Vision (ECCV) (Lecture Notes in Computer Science)*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 102–118.

[14] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2021. Lips Don't Lie: A Generalisable and Robust Approach To Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5039–5049.

[15] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. 2021. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4360–4369.

[16] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik. 2023. AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection. *Applied Soft Computing* 136 (March 2023), 110124.

[17] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. 2020. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2889–2898.

[18] Ziyue Jiang, Jinglin Liu, Yi Ren, Jinzheng He, Chen Zhang, Zhenhui Ye, Pengfei Wei, Chunfeng Wang, Xiang Yin, Zejun Ma, and Zhou Zhao. 2023. Mega-TTS 2: Zero-Shot Text-to-Speech with Arbitrary Length Speech Prompts. arXiv:2307.07218 [cs, eess].

[19] Ziyue Jiang, Yi Ren, Zhenhui Ye, Jinglin Liu, Chen Zhang, Qian Yang, Shengpeng Ji, Rongjie Huang, Chunfeng Wang, Xiang Yin, Zejun Ma, and Zhou Zhao. 2023. Mega-TTS: Zero-Shot Text-to-Speech at Scale with Intrinsic Inductive Bias. arXiv:2306.03509 [cs, eess].

[20] Hasam Khalid, Shahroz Tariq, and Simon S. Woo. 2021. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. arXiv: 2108.05080 [cs].

[21] Muhammad Riyyan Khan, Shahzeb Naeem, Usman Tariq, Abhinav Dhall, Malik Nasir Afzal Khan, Fares Al Shargie, and Hasan Al-Nashash. 2023. Exploring Neurophysiological Responses to Cross-Cultural Deepfake Videos. In *Companion Publication of the 25th International Conference on Multimodal Interaction*. 41–45.

[22] Pavel Korshunov and Sebastien Marcel. 2018. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. arXiv:1812.08685 [cs].

[23] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. 2021. KoDF: A Large-Scale Korean DeepFake Detection Dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10744–10753.

[24] Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2021. OpenForensics: Large-Scale Challenging Dataset for Multi-Face Forgery Detection and Segmentation In-the-Wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10117–10127.

[25] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face X-Ray for More General Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5001–5010.

[26] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3207–3216.

[27] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. 2023. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 2507–2522.

[28] Shahzeb Naeem, Muhammad Riyyan Khan, Usman Tariq, Abhinav Dhall, Carlos Ivan Colon, and Hasan Al-Nashash. 2024. Generation and Detection of Sign Language Deepfakes-A Linguistic and Visual Analysis. *arXiv preprint arXiv:2404.01438* (2024).

[29] Kartik Narayan, Harsh Agarwal, Kartik Thakral, Surbhi Mittal, Mayank Vatsa, and Richa Singh. 2023. DF-Platter: Multi-Face Heterogeneous Deepfake Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9739–9748.

[30] Dufour Nick and Jigsaw Andrew. 2019. Contributing Data to Deepfake Detection Research.

[31] Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. 2024. AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27102–27112.

[32] Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *INTERSPEECH 2023*. ISCA, 3222–3226.

[33] Diego Pérez-Vieites, Juan José Moreira-Pérez, Ángel Aragón-Kifute, Raquel Román-Sarmiento, and Rubén Castro-González. 2024. Vigo: Audiovisual Fake Detection and Segment Localization. In *ACM international conference on multimedia*.

[34] Muhammad Anas Raza and Khalid Mahmood Malik. 2023. Multimodaltrace: Deepfake Detection Using Audiovisual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 993–1000.

[35] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1–11.

[36] Ian Sample. 2020. What are deepfakes – and how can you spot them? *The Guardian* (Jan. 2020).

[37] Conrad Sanderson (Ed.). 2002. *The VidTIMIT Database*. IDIAP.

[38] Oscar Schwartz. 2018. You thought fake news was bad? Deep fakes are where truth goes to die. *The Guardian* (Nov. 2018).

[39] Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. 2024. Detecting and Grounding Multi-Modal Media Manipulation and Beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024), 1–18.

[40] Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2023. NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers. arXiv:2304.09116 [cs, eess].

[41] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. 2023. TriDet: Temporal Action Detection With Relative Boundary Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18857–18866.

[42] Kaede Shiohara and Toshihiko Yamasaki. 2022. Detecting Deepfakes With Self-Blended Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18720–18729.

[43] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2022. Make-A-Video: Text-to-Video Generation without Text-Video Data. arXiv:2209.14792 [cs].

[44] Daniel Thomas. 2020. Deepfakes: A threat to democracy or just a bit of fun? *BBC News* (Jan. 2020).

[45] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. 2022. M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22)*. Association for Computing Machinery, New York, NY, USA, 615–623.

[46] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. VideoMAE V2: Scaling Video Masked Autoencoders With Dual Masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14549–14560.

[47] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. 2022. InternVideo: General Video Foundation Models via Generative and Discriminative Learning. arXiv:2212.03191 [cs].

[48] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7623–7633.

[49] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. ISSN: 2379-190X.

[50] Zhen Xu, Sergio Escalera, Adrien Pavao, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns* 3, 7 (2022).

[51] Ke Yang, Peng Qiao, Dongsheng Li, Shaohe Lv, and Yong Dou. 2018. Exploring Temporal Preservation Networks for Precise Temporal Action Localization. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (April 2018).

[52] Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. 2023. AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake. *IEEE Transactions on Information Forensics and Security* 18 (2023), 2015–2029.

[53] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, Shan Liang, Shiming Wang, Shuai Zhang, Xinrui Yan, Le Xu, Zhengqi Wen, Haizhou Li, Zheng Lian, and Bin Liu. 2022. ADD 2022: the First Audio Deep Synthesis Detection Challenge. arXiv:2202.08433 [cs, eess].

[54] Yang Yu, Xiaolong Liu, Rongrong Ni, Siyuan Yang, Yao Zhao, and Alex C. Kot. 2023. PVASS-MDD: Predictive Visual-audio Alignment Self-supervision for Multimodal Deepfake Detection. *IEEE Transactions on Circuits and Systems for Video Technology* (2023), 1–1.

[55] Chen-Lin Zhang, Jianxin Wu, and Yin Li. 2022. ActionFormer: Localizing Moments of Actions with Transformers. In *Proceedings of the European Conference on Computer Vision (ECCV) (Lecture Notes in Computer Science)*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 492–510.

[56] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. arXiv:2306.02858 [cs, eess].

[57] 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)*. Association for Computing Machinery, New York, NY, USA, 8749–8759.

[58] Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. 2021. Face Forensics in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5778–5788.

[59] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. 2020. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. Association for Computing Machinery, New York, NY, USA, 2382–2390.