VCapAV: A Video-Caption Based Audio-Visual Deepfake Detection Dataset

Yuxi Wang[1,*], *Yikang Wang*[1,2,*], *Qishan Zhang*[1], *Hiromitsu Nishizaki*[2], *Ming Li*[1,3,†]

[1]Suzhou Municipal Key Laboratory of Multimodal Intelligent System, Digital Innovation Research Center, Duke Kunshan University, China
[2]Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi, Japan
[3]School of Computer Science, Wuhan University, China

ming.li369@dukekunshan.edu.cn

Abstract

Most existing deepfake datasets focus on speech synthesis or voice cloning, with little attention given to non-speech environmental sounds. Existing audio-focused datasets also lack video content, restricting progress in multimodal detection. To bridge gaps, we introduce VCapAV, a large-scale audio-visual dataset, designed to advance deepfake detection research involving environmental sound manipulations in multimodal scenarios. VCapAV is constructed through an innovative data generation pipeline that synthesizes realistic environmental audio using Text-to-Audio and Video-to-Audio approaches, while deepfake videos are generated through a Text-to-Video model. We establish two baseline detection tasks on this dataset: (i) audio-only deepfake detection, and (ii) visual-only deepfake detection. Experimental results show that existing detection models on the VCapAV dataset compared to standard datasets such as ASVspoof 2019 LA and AV-Deepfake1M. The dataset and baseline codes * are released.

Index Terms: deepfake detection, speaker verification, anti-spoofing, audio-visual dataset

1. Introduction

The rapid advancement of content generation technologies enables the creation of highly realistic synthetic media, including text, video, and audio. As generative models continue to improve, the boundary between real and fake content becomes increasingly blurred, often surpassing human perceptual capabilities [1]. This raises concerns about the misuse of such technologies, such as spreading misinformation or committing fraud, highlighting an urgent need for robust and reliable detection methods.

The effectiveness of DeepFake detection approaches largely relies on high-quality benchmark datasets. In recent years, the number of publicly available datasets has been growing, covering various types of content manipulations, including visual-only [2], audio-only [3], and audio-visual combinations [4–6]. These datasets support the evaluation of detection models across different modalities.

However, most existing audio datasets and detection methods primarily focus on speech content, often overlooking non-speech environmental audio, which is equally critical in multimedia. This limitation results in noticeable vulnerabilities when detecting DeepFakes involving environmental sounds or non-linguistic audio manipulations, which can significantly alter the semantic meaning of content or the acoustic scene of the event.

Recently, at Interspeech 2024, Xie et al. [7] use an AudioLDM [8] based generation pipeline, opening new directions for general audio deepfake techniques. However, our work demonstrates greater sophistication in audio-only deepfake methods and further expands into video deepfake, building a more comprehensive system that encompasses multimodal manipulations.

To address the lack of support for video and multimodal data in existing work, we introduce VCapAV—a dataset consisting of 90,990 clips, totaling 252.75 hours, specifically designed for audio-visual deppfake detection tasks with a focus on environmental sounds. The VCapAV dataset is constructed using an innovative data generation pipeline that integrates state-of-the-art (SOTA) deepfake strategies and cutting-edge audio and video synthesis technologies. This pipeline generates realistic environmental audio and visual content manipulations.

The comprehensive dataset includes visual-only, audio-only, and audio-visual DeepFake data, offering extensive support for evaluating detection models across different modalities. It also enables rigorous evaluation and benchmarking of detection algorithms under diverse deepfake conditions.

2. VCapAV Dataset

The VCapAV dataset is a deepfake audio-visual dataset focusing solely on non-speech content. It contains audio-visual recordings in various background environments, involving five audio deepfake methods and one video deepfake method, with a total duration of 252.75 hours. The synthesized audio data in the VCapAV Dataset is through two approaches: Video-To-Audio (V2A) and Text-To-Audio (TTA), where the text comes from the caption of the video signal. The V2A approach utilizes two models, while the TTA approach employs three models. The synthesized video data in the VCapAV Dataset is generated using the Kling large model †, producing high-quality fake videos with 1080p resolution at 30 frames per second, with a duration of approximately 10 seconds per video.

To ensure the diversity and authenticity of the content, we select a test set of 15,446 video samples from the VGGSound dataset [9]. VGGSound is a large-scale audiovisual dataset containing over 210,000 10-second video clips sourced from YouTube, meticulously labeled with more than 310 sound event classes. The chosen subset comprehensively represents various audiovisual content, which is crucial for generating accurate and meaningful captions that serve as a basis for V2A and TTA synthesis.

Referring to the generation methodology of the AV-Deepfake1M Dataset [4], this dataset incorporates three different modality combinations in the generated fake data:

*These authors contributed equally.
†Corresponding author.
*https://VCapAV.github.io/

†https://klingai.com/

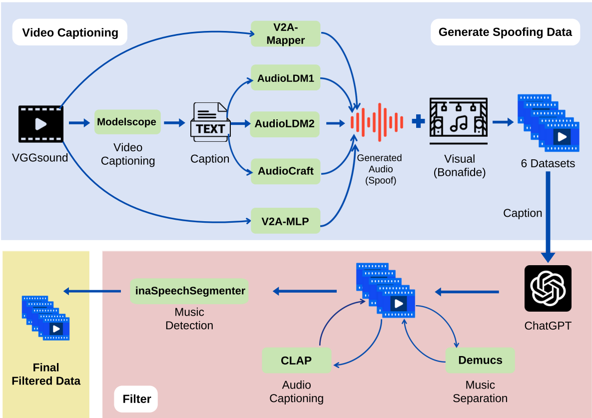

Figure 1: *Overview of the VCapAV dataset creation pipeline.*

- **Fake Audio + Real Visual**: The visual data remains unaltered, while the audio data is manipulated. This category divides into two forging approaches: V2A and TTA. The text is obtained from a video captioning model, ModelScope [10].
- **Real Audio + Fake Visual**: The audio data remains unaltered, while the visual data is manipulated.
- **Fake Audio + Fake Visual**: Both the audio and visual data are manipulated. Kling uses the video captioning results, which we obtained from the videos, as the input to the TTV model.

2.1. Audio Generation Pipeline

2.1.1. V2A Manipulation

This approach involves generating audio directly aligned with the visual content of the video. We explore two methods:

- **V2A-Mapper** [11]: The V2A data, made publicly available by Wang et al., consists of synthetic audio generated by applying the V2A model to the VGGSound test set [9].
- **V2A-MLP** [12]: Yi et al. propose an MLP-based mapper to train the V2A generation model. Their approach reimplements and streamlines a state-of-the-art (SOTA) V2A model with a lightweight architecture, achieving superior performance over the baseline. In addition, they incorporate a scene detector to handle multi-scene videos, effectively addressing scene transitions. Evaluations on VGGSound demonstrate improvements in both audio fidelity and relevance.

2.1.2. TTA Manipulation

This strategy focuses on generating audio based on textual descriptions derived from video content. It consists of two stages:

- **Video Captioning Stage**: Text descriptions are generated from videos in the video captioning stage. We employ the ModelScope framework [10] on the VGGSound dataset [9]. We feed the preprocessed video data into the ModelScope model to generate captions that integrate both visual and auditory cues.
- **Audio Synthesis Stage**: Based on the generated captions, we synthesize audio using the following models:

(i) **AudioLDM1** [8]: This model employs a self-supervised learning framework with Latent Diffusion Models (LDMs). Through contrastive learning, AudioLDM1 aligns textual and auditory modalities within a unified latent space, reducing the reliance on large-scale paired audio-text datasets and generating high-quality audio outputs from text inputs.

(ii) **AudioLDM2** [13]: Building upon AudioLDM1 [8], AudioLDM2 refines the alignment between text and audio modalities by incorporating additional enhancements, ensuring that the generated audio maintains contextual relevance and high fidelity.

(iii) **Audiocraft** [14]: Developed by Facebook Research, Audiocraft is a deep learning-based library for audio processing and generation. We specifically adopt the AudioGen model from this library, which generates audio from textual descriptions, enabling the synthesis of realistic soundscapes aligned with provided prompts.

2.2. Video Generation Pipeline

We generate fake videos using the Kling large model, with descriptive text from real video captioning as the prompts. Compared to other similar video generation models, the Kling large model offers several advantages. It delivers superior visual fidelity and more consistent frame quality, ensuring that even complex scenes are rendered with remarkable detail. Moreover, its efficient prompt-based synthesis allows for faster generation of high-resolution content, making it a compelling choice for applications requiring diverse and high-quality video outputs. The Kling large model produces 10-second videos in 1080p resolution, yielding a total of 242 clips.

Since the captions often contain repetitive wording while conveying the same core meaning, with identical underlying keywords (e.g., “A band performs guitar“ and “A crowd of people play guitar on the stage“ describing the same event), we employ a keyword-based filtering approach. We manually identify frequently occurring keywords from the captions, resulting in a refined set of descriptors corresponding to 242 predefined sound event classes. Using this curated descriptor set, we generate 242 fake videos through Kling API.

2.3. Filtering Pipeline

To ensure the integrity of our dataset, we implement a filtering process that addresses both non-environmental music and duration discrepancies. The following sections detail our approach.

2.3.1. Music Filtering

Identification of Accompaniment Music. During the generation of audio captions, we observe that certain videos incorporate accompaniment music added during post-production rather than capturing authentic environmental sounds. Although these musical segments contribute to the overall aesthetic, they do not align with the environmental audio that constitutes the focus of our study. It is therefore imperative to exclude such videos from our dataset.

Multi-step Filtering Process. We implement a systematic, multi-step filtering process to identify and remove instances of non-environmental music. We consider live music as part of environmental sounds, whereas background accompaniment music is added during post-production and may interfere with deepfake detection. In the first step, we use ChatGPT [‡] to generate a comprehensive script that enumerates a wide range of musical instruments (e.g., piano, violin, various drums). This

[‡]https://chat.openai.com/

script is applied to the video captions produced by the ModelScope model [10]. The descriptions generated by the ModelScope model are primarily focused on the scene and rarely mention background accompaniment music. If a caption contains references to any of these instruments, the corresponding video is marked as potentially containing live music, which is considered part of the environmental context. This initial filtering reduces the dataset to 13,016 samples by eliminating videos in which music is introduced during editing.

Refinement with Demucs and CLAP. Subsequently, we employ the Demucs [15, 16] speech separation model to further refine the dataset. Demucs separates audio tracks into distinct components (e.g., vocals, accompaniment, environmental sounds). We process the audio files using a script that generates textual descriptions via the CLAP [17, 18] Audio-to-Caption model. For example, processing the original audio file `bV7Vr3sUq_s_000078.wav` with CLAP yields the description "A drill is being used and then a person talks," whereas processing the same file with Demucs yields "A song is played and then a melody is played." This discrepancy underscores the efficacy of Demucs in isolating musical elements. We then reapply the filtering step to detect any residual references to musical instruments, noting that videos identified as containing live music are excluded; the remaining cases pertain to accompaniment music likely added during post-production.

Verification via inaSpeechSegmenter. To further validate our filtering process, we introduce an additional verification step using the inaSpeechSegmenter [19], a model designed for precise music detection. This model outputs exact time intervals during which music is present in the audio track. For instance, in an analysis of a video featuring a snowy landscape, inaSpeechSegmenter identifies a segment of music between the 8- and 10-second marks.

Filtering Outcome. By the results from the CLAP-generated captions and the inaSpeechSegmenter detections [19], we achieve a high degree of accuracy in isolating non-environmental music. This workflow ultimately yields 782 audio files to be filtered out, representing potential cases of background music. Following manual review, we further reduce these to 523 cases, thereby refining the final VCapAV dataset from 15,446 samples to 14,923.

2.3.2. Duration Discrepancies in the VGGSound Dataset

Analysis of Video Duration Variability. In integrating the generated audio with original videos from the VGGSound dataset [9], we observe significant discrepancies in audio and video durations that impede synchronization and compromise data integrity. Although the dataset is designed to contain 10-second clips, our analysis reveals variability in video lengths: a mean of 10.94 seconds, a median of 10.08 seconds, a mode of 10.02 seconds, and a range from 2.02 to 15.72 seconds.

Synchronization Strategy. Since the V2A-Mapper [11] generates audio tracks with an exact duration of 10 seconds, these inconsistencies pose synchronization challenges. To address this issue, for videos with a duration of 10 seconds or less, we extract audio segments matching the original video length and align the audio with the video. For videos exceeding 10 seconds, we trim the excess footage to retain only the first 10 seconds, thereby ensuring synchronization between the video and its corresponding audio track. Our analysis indicates that 591 videos have durations shorter than 10 seconds and require no modification, while longer videos have segments beyond 10 seconds trimmed to ensure consistency throughout the dataset.

2.4. Dataset Statistics

Table 1: *Statistics of the VCapAV dataset*

Category	Number of Clips	Total Duration (hours)
Real Video	14,923	41.45
Fake Video	242	0.67
Real Audio	14,923	41.45
Fake Audio	74,615	207.26
Real Video + Real Audio	14,923	41.45
Real Video + Fake Audio	74,615	207.26
Fake Video + Real Audio	242	0.67
Fake Video + Fake Audio	1,210	3.36
Total	**90,990**	**252.75**

Table 1 summarizes the key characteristics of the VCapAV dataset. After applying our multi-stage filtering process, 523 samples are removed from the original 15,446 files, yielding a final dataset of 14,923 original audio-visual clips. Each clip is 10 seconds or less in length with audio uniformly downsampled to 16,000 Hz. The dataset encompasses multiple deepfake types: it includes 14,923 real videos and 242 fake videos (generated via the Kling large model) paired with 14,923 real audio samples and 74,615 fake audio samples (produced using five distinct manipulation methods). This results in 14,923 pairs of real video with real audio, 74,615 pairs of real video with fake audio, 242 pairs of fake video with real audio, and 1,210 (242*5) pairs of fake video with fake audio. Additionally, the dataset covers 242 sound event classes, ensuring a diverse representation of environmental sounds.

As shown, the final dataset retains diversity and complexity to facilitate comprehensive evaluations of audio-visual environmental sound deepfake detection methods. The presence of multiple deepfake techniques also ensures robustness in testing various detection algorithms, ultimately supporting the advancement of deepfake detection research.

3. Benchmarks and Metrics

3.1. Data Partitioning

Table 2: *Data partitioning for audio detection and video detection tasks*

	Composition	Total Size
Audio Train	AudioLDM1, AudioLDM2, V2A-Mapper, Bonafide	44,596 (11,149 * 4)
Dev 1	AudioLDM1, AudioLDM2, V2A-Mapper, Bonafide	15,096 (3,774 * 4)
Dev 2	V2A-MLP, Audiocraft, Bonafide	11,322 (3,774 * 3)
Dev 3	AudioLDM1, AudioLDM2, V2A-Mapper, V2A-MLP, Audiocraft, Bonafide	22,644 (3,774 * 6)
Video Train	Bonafide, Kling	11,192, 181
Dev	Bonafide, Kling	3,731, 61

For each task, the VCapAV dataset is divided into a train set and a development set, without a separate test set. All data are split in a 3:1 ratio to ensure stable training and fair evaluation of the models.

In the audio-only deepfake detection task, different fake generation methods are assigned to different development sets to assess the model's generalization ability to unseen manipulation techniques. Dev 1 set contains the same manipulation techniques as the training set, serving as a benchmark for evaluating model performance under similar data distributions. Dev 2 set consists of manipulation techniques not present in the training set, designed to assess model robustness to new types of fake audio. Dev 3 set integrates all manipulation techniques to evaluate model performance in a setting where multiple fake generation methods co-exist.

For the visual-only deepfake detection task, to ensure a balanced distribution of real and fake data across different tasks, we maintain a consistent ratio of bonafide and fake samples, preventing potential bias in the dataset distribution from affecting model evaluation. In the future, we will add more data in the fake video class and provide an update on the trials.

3.2. Benchmark Methods

In the benchmark evaluation for deepfake detection, we consider the following methods for audio-based and visual-based detection baselines. For the audio-only deepfake detection task, we use three models commonly adopted in audio spoofing detection: LightCNN [20], ResNet18 [21], and AASIST [22]. For the visual-only deepfake detection task, we train a video-level classification model Meso4 [23] with video samples and their corresponding labels.

3.3. Metrics

We evaluate the performance of audio-only and video-only detection models using the following metrics [24, 25]:

- **Equal Error Rate (EER)**: The error rate at the operating point where the false acceptance rate (FAR) equals the false rejection rate (FRR).
- **Accuracy (Acc.)**: The proportion of correctly classified samples over the total number of samples.
- **Area Under the Receiver Operating Characteristic Curve (AUC)**: Measures the ability of the model to discriminate between bonafide and fake samples across different thresholds.

4. Results and Analysis

4.1. Audio-only Deeofake Detection Results

Although the AASIST model demonstrated robust performance across all development subsets, LightCNN and ResNet18 models exhibited significant performance degradation on Dev2 and Dev3 subsets. This result suggests that existing detection models face performance bottlenecks when confronted with unseen types of sophisticated spoofing attacks or more challenging audio synthesis conditions.

Notably, the ResNet18 model trained on the ASVspoof 2019 LA dataset yielded a high EER of 41.18% when tested directly on the VCapAV Dev3 subset, highlighting a severe performance drop under cross-dataset evaluation. This further validates that the VCapAV dataset presents a distributional shift in attack types, synthesis methods, and audio conditions compared to existing standard datasets such as ASVspoof 2019 LA.

The performance disparities across VCapAV subsets and the cross-dataset generalization failure collectively indicate that the VCapAV dataset encompasses spoofing techniques and audio characteristics that are underrepresented in current benchmark datasets. This underscores the potential of VCapAV to expose the robustness limitations of existing detection systems and to serve as a more challenging evaluation platform for future research.

Table 3: *Detection performance on the audio-only task across different development subsets*

Subset	Metrics	Methods		
		LightCNN	**ResNet18**	**AASIST**
Dev 1	EER (%)	0.8435	0.3135	0.02
	minDCF (%)	9.72	5.43	0.01
	AUC (%)	99.96	99.96	100
	Acc. (%)	98.42	99.7	99.92
Dev 2	EER (%)	18.8924	14.6728	2.96
	minDCF (%)	83.29	78.34	0.54
	AUC (%)	90.46	93.82	88.91
	Acc. (%)	73.52	75.17	99.52
Dev 3	EER (%)	9.6714	9.8198	1.78
	minDCF (%)	78.53	78.27	0.36
	AUC (%)	96.06	96.16	99.81
	Acc. (%)	88.24	87.87	96.29

4.2. Visual-only Deepfake Detection Results

Table 4: *Performance of the Meso4 system trained on VCapAV video-only training set on AV-Deepfake1M and VCapAV datasets*

Methods	AV-Deepfake1M		VCapAV	
	AUC (%)	**Acc. (%)**	**AUC (%)**	**Acc. (%)**
Meso4	63.05	49.51	68.15	54.5

We train and evaluate the Meso4 [23] system on visual-only task, achieving an accuracy of 54.50% on the development set. In contrast, the reported accuracy of the Meso4 system on the AV-Deepfake1M dataset [4] is 75%, despite that much more training data were adopted. The significant performance gap suggests that the deepfake techniques or visual feature distributions in the VCapAV dataset differ considerably from those in existing mainstream video deepfake datasets such as AV-Deepfake1M. Also, the former is focusing on the general acoustic events, while the latter is more on the virtual human talking. More works need to be done to enhance the robustness against different scenarios.

5. Conclusion and Future Work

The VCapAV dataset offers a novel benchmark for general audio-visual deppfake detection, addressing gaps in existing datasets by introducing more complex and realistic environmental sound manipulation scenarios. Compared to previous benchmarks, VCapAV potentially covers some new attack techniques and presents greater challenges.

Future work will focus on enhancing audio-visual deepfake detection by exploring multi-modal fusion approaches. Integrating video-based visual cues with audio signals is a promising direction, as visual context often complements audio information, offering additional evidence to improve detection performance. Developing joint audio-visual deepfake detection models capable of leveraging temporal synchronization and cross-modal correlations will likely improve robustness against increasingly sophisticated deepfake attacks.

6. Acknowledgements

This research is funded by Duke Kunshan University's foundation project "Emerging AI Technologies for Natural Language Processing". Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

7. References

[1] T. Zhou, W. Wang, Z. Liang, and J. Shen, "Face Forensics in the Wild," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 5778–5788.

[2] P. Kwon, J. You, G. Nam, S. Park, and G. Chae, "Kodf: A large-scale korean deepfake detection dataset," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 744–10 753.

[3] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan *et al.*, "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9216–9220.

[4] Z. Cai, S. Ghosh, A. P. Adatia, M. Hayat, A. Dhall, T. Gedeon, and K. Stefanov, "AV-Deepfake1M: A Large-Scale LLM-Driven Audio-Visual Deepfake Dataset," in *Proc. 32nd ACM International Conference on Multimedia*, Oct. 2024, pp. 7414–7423.

[5] Z. Cai, S. Ghosh, A. Dhall, T. Gedeon, K. Stefanov, and M. Hayat, "Glitch in the Matrix: A Large Scale Benchmark for Content Driven Audio-Visual Forgery Detection and Localization," *Computer Vision and Image Understanding*, vol. 236, p. 103818, Not Available 2023.

[6] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, "ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech," *Computer Speech & Language*, vol. 64, Nov. 2020.

[7] Z. Xie, B. Li, X. Xu, Z. Liang, K. Yu, and M. Wu, "FakeSound: Deepfake General Audio Detection," in *Proc. INTERSPEECH 2024 – 25th Annual Conference of the International Speech Communication Association*, Kos, Greece, Sep. 2024.

[8] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," *Proceedings of the International Conference on Machine Learning*, 2023.

[9] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[10] T. M. Team, "Modelscope: bring the notion of model-as-a-service to life." https://github.com/modelscope/modelscope, 2023.

[11] H. Wang, J. Ma, S. Pascual, R. Cartwright, and W. Cai, "V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[12] M. Yi and M. Li, "Efficient video to audio mapper with visual scene detection," 2024. [Online]. Available: https://arxiv.org/abs/2409.09823

[13] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.

[14] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[15] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *ICASSP 23*, 2023.

[16] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, 2021.

[17] Y. Wu*, K. Chen*, T. Zhang*, Y. Hui*, T. Berg-Kirkpatrick, and S. Dubnov, "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*, 2023.

[18] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2022.

[19] D. Doukhan, E. Lechapt, M. Evrard, and J. Carrive, "Ina's mirex 2018 music and speech detection system," in *Music Information Retrieval Evaluation eXchange (MIREX 2018)*, 2018.

[20] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC Antispoofing Systems for the ASVspoof2019 Challenge," in *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] 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)*. IEEE, 2022, pp. 6367–6371.

[23] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2018, pp. 1–7.

[24] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[25] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.