Informatics Bulletin, Faculty of Computers and Artificial Intelligence, Helwan University
Published Online *Vol 7 Issue 2, July 2025*
(https://fcihib.journals.ekb.eg)

Unmasking the Digital Deception: A Comprehensive Survey of Large Vision Models (LVMs) for Deepfake Detection

Ahmed Ashraf Bekheet[1], Amr S. Ghoneim[1], Ghada Khoriba[2]
[1]Computer Science, Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt
[2]Center for Informatics Science (CIS), School of Information Technology and Computer Science (ITCS), Nile University, Giza, Egypt
ahmed_ashraaf@fci.helwan.edu.eg, amr.ghoneim@fci.helwan.edu.eg, ghadakhoriba@nu.edu.eg

***Abstract*— Digital videos are among the most prevalent types of multimedia in everyday life. They are extensively shared on social media channels like Facebook, Instagram, WhatsApp, and YouTube via the Internet. The rapid advancements in artificial intelligence (AI), machine learning (ML), and deep learning (DL) have led to the development of sophisticated techniques and tools for multimedia manipulation. These technological innovations have facilitated the creation of falsified digital images and videos. Consequently, detecting these manipulated digital media has become a critical concern, necessitating a thorough examination of current forgery detection methodologies. Our extensive survey categorizes these methodologies across three visual, audio, and multimodal audio-visual domains. The survey broadly examines deepfake detection strategies, with a particular emphasis on applying recent deep learning techniques, specifically large vision models (LVMs). It includes an in-depth comparative analysis of various deep learning approaches, focusing on LVMs, and demonstrates their superior performance relative to earlier techniques. Multiple metrics and datasets support this analysis. Additionally, it offers new solutions and guides future research in multimodal deepfake detection by exploring new dimensions of video manipulation, such as text overlays and motion dynamics. It also highlights the growing importance of expanding the role of LVMs and underscores the importance of developing comprehensive and diverse datasets to enhance the robustness and validation of detection techniques.**

***Index Terms*— Deepfakes, Audio-Visual Deepfake Detection, Large Vision Models (LVMs), Convolutional Neural Networks (CNNs), Transformers, Vision Transformer (ViT), Voice Conversion, Mel Frequency Cepstral Coefficients (MFCC)**

I. Introduction

Digital video is characterized by a series of images that have been captured by a digital camera sometimes in conjunction with an audio track and possible additional data dimensions. Users are increasingly using multimedia content in their daily lives with a strong prevalence of digital video. Surveillance cameras, now commonly deployed throughout offices and homes and in public areas, have become a valuable enhancement to safety and security efforts. Video has also become a common source of evidence in legal cases in a in a number of countries. However, the availability of editing software that incorporates advanced algorithms and signal processing available on most modern smartphone devices has made it possible for individuals to also create false (or modified) digital images and video content for YouTube and social media purposes. Many of these popular editing applications are based on older practices involving algorithm creation by hand and analyzing signals on a case-by-case basis. More recently, with the evolution of new neural network architectures, comes the introduction of better methods known as ”deepfake tools,” that rely on deep learning to perform more complex and advanced changes to original digital media.

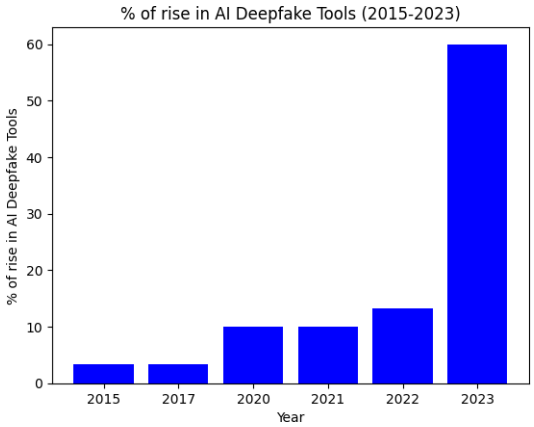

Fig. 1: A chart depicting the annual percentage distribution of Artificial Intelligence (AI) deepfake tools creation.

As shown in Figure 1, the research by humanorai [1], which examines over 30 deepfake tools, illustrates a significant evolution from 2015 to 2023. Initially, there was minimal development, with only 3.3% of tools emerging between 2015 and 2017. However, by 2023, there was a notable increase, with 60% of these tools appearing during that period. This trend highlights that most known deepfake tools were developed during 2023.

The proliferation of various deepfake generation techniques has resulted in significant societal challenges, notably the dissemination of misinformation. This phenomenon distorts reality and undermines the dissemination of accurate

information.

In this survey, we conduct a comprehensive analysis of deepfake detection techniques across visual, audio, and multimodal audio-visual domains, with a particular focus on the advanced capabilities of Large Vision Models (LVMs). The literature reviewed for this study was sourced from prominent researches including Google Scholar, IEEE Xplore, and PubMed, covering studies published from 2019 to 2024. These models exhibit significant advantages over traditional deep learning methods, primarily due to their improved scalability, ability to recognize complex visual patterns, and superior performance with extensive datasets. While previous approaches often faced challenges in addressing the intricacies of sophisticated deepfake techniques, LVMs utilize advanced neural architectures that enhance their capacity to distinguish between genuine and synthetic media. This study highlights the research gap in current detection methodologies, underscoring the need for a deeper exploration of how LVMs can be harnessed to improve the effectiveness and accuracy of deepfake detection.

In our survey, we also highlight significant datasets relevant to deepfake detection across various domains, including visual, audio, and multimodal audio-visual categories, such as DeeperForensics-1 [2], ASVspoof 2021 [3], and FakeAVCeleb [4]. These datasets are characterized by their size, diversity, and the types of manipulations involved, providing a solid foundation for training and validating detection models.

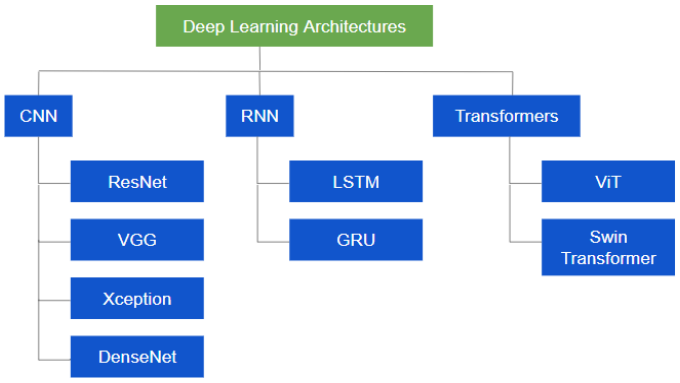

Fig. 2: Various Deep Learning Architecture Types.

Figure 2 illustrates that Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers are prominent deep learning techniques that are being increasingly employed for both the generation and detection of deepfakes.

CNN is a well-founded deep learning architecture that is optimized for the tasks of image classification and image processing. CNN contains multiple layers such as convolutional layers, pooling layers, and fully connected layers. Notable applications of CNNs include architectures such as ResNet [5], VGG [6], Xception [7], and DenseNet [8], each of which has been designed to improve performance and accuracy in various visual recognition tasks.

The main advantage of RNN is its ability to handle sequential or temporal data through the retention of information over time. The LSTM (Long Short-Term Memory) [9] model was developed to mitigate the longstanding vanishing gradient problem in RNN architectures, where memory cells and gating mechanisms allow LSTMs to effectively model dependencies across long sequences. GRU (Gated Recurrent Unit) [10] is a simplified version of an RNN that, like LSTM, employs gating mechanisms.

Transformers have been instrumental in modeling long-range dependencies in sequence data and were originally applied in Natural Language Processing (NLP). Since their inception in NLP, they have been adapted for vision tasks such as Vision Transformers (ViTs) [11], which enable the implementation of self-attention mechanisms to proportionally weight the importance of different spatial locations in input images. The Swin transformer [12], which can be classified as a ViT, is specifically able to model at multiple scales and with a complexity that is linear as an input image grows in size. Cumulatively, these aspects make the Swin transformer very powerful for tasks related to image classification and detection.

Following the discussion of prominent deep learning architectures, we will now explore LVMs and their impact on deepfake detection.

An **LVM** is an end-to-end Artificial Intelligence system created for handling, investigating, and correctly appreciating vision content like images or videos. These models are the visual equivalents of Large Language Models (LLMs) such as BERT [13] and GPT-4 [14], which have been designed to understand human language and generate it. Like the LLMs, LVMs hold many parameters and can therefore learn complex patterns from large-scale datasets. LVM development has greatly felt the impacts of LLM success.

LLMs have demonstrated their capacity to scale by processing larger datasets with complex architecture and training routines. This success has prompted similar mechanisms in the visual domain; LVMs have adapted these to visual data and are informed by deep learning methods developed for the LLMs.

The LVMs exploit transformer models in their architecture to leverage their ability, as initially demonstrated in transformers, to capture long-range patterns in visual data.

Figure 3 highlights several prominent examples of LVMs, such as the Swin Transformer [12], CLIP (Contrastive Language–Image Pretraining) [15], Vision Transformer (ViT) [11], and DALL-E [16].

Our paper focuses on LVMs that are especially effective in deepfake detection such as Vision Transformer (ViT) and Swin Transformer, and organizes the research into three primary categories:

- *Visual-based deepfake detection*
- *Audio-based deepfake detection*
- *Multimodal audio-visual-based deepfake detection*

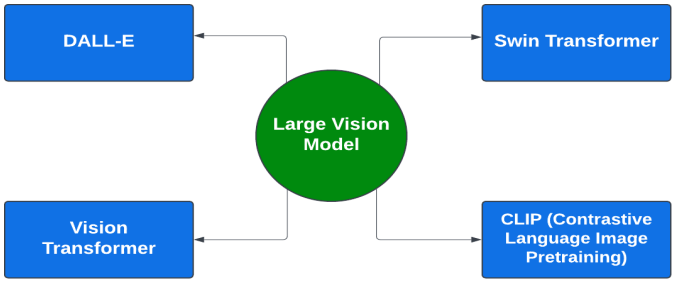

Fig. 3: Prominent examples of LVMs.

The paper's structure is as follows: Section II focuses on deepfake generation and detection based on visual cues. Section III covers the audio-based generation and detection of deepfakes. In Section IV, we take a comprehensive approach by discussing the detection of deepfakes that utilize both audio and visual elements, reflecting the multimodal nature of contemporary deepfake creations. Subsequently, Section V provides an in-depth review of commonly utilized datasets in the field of deepfake detection, showcasing their importance in training and validating detection models. In Section VI, we engage in a comparative analysis of various detection models, highlighting the aspects that demonstrate the superiority of LVMs over traditional deep learning techniques. This section aims to elucidate the reasons behind the enhanced performance of LVMs in deepfake detection. Finally, Section VII concludes the paper by summarizing our findings and proposing future research directions, highlighting the ongoing challenges and opportunities in the evolving landscape of deepfake detection.

II. Visual-based Deepfake Detection

This section reviews various approaches employed in the identification of fake images and videos, focusing particularly on advanced deepfake detection techniques that utilize LVMs. It analyzes the methods for both the generation and detection of deepfake images and videos.

Generative Adversarial Networks (GANs) [17], though not classified as LVMs, remain the leading technique for generating synthetic images and videos after being trained on datasets. In addition to simply swapping faces, GANs can also swap faces and modify facial expressions while synchronizing lip movements with audio. A GAN consists of two major parts: a generator and a discriminator. The generator acts on data much like a decoder by generating synthetic data from random noise or latent representations. The discriminator also functions in the learning process but does not simply generate the audio; it tries to label the data as fake or real. Therefore, the discriminator's role is to evaluate the validity and authenticity of the generated data in differentiating real from fake samples. Through the adversarial process, the generator consistently improves its ability to generate realistic outcomes, while the discriminator simultaneously improves its ability to invalidate fakes from the real data. In reality, GANs are usually used for generation purposes rather than detection.

An effective LVM for visual deepfake detection is the Convolutional Vision Transformer (CViT), which combines CNNs for feature extraction with Vision Transformer [18]. This approach adeptly manages complex data relationships, facilitating the identification of deepfakes. Deressa Wodajo et al. introduced this approach for identifying deepfake videos. The CViT model comprises two main components: feature learning (FL) and Vision Transformer (ViT).

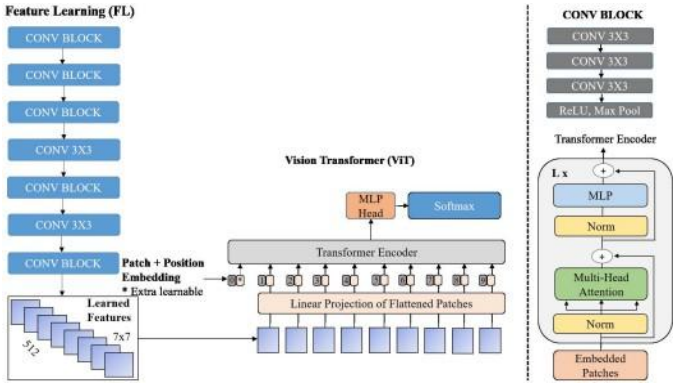

Fig. 4: The Framework of CViT for Visual Deepfake Detection [18].

Like a VGG structure without fully connected layers, the FL component extracts facial features from frames for subsequent analysis rather than classification.

The ViT component processes the feature maps extracted by FL. These maps are divided into seven patches and then embedded into a linear sequence of size 1 x 1024. Position embeddings are added to maintain spatial information. ViT utilizes a transformer with an encoder, an MLP (Multilayer Perceptron) block for feedforward operations, and a softmax layer for classification (real vs. fake). Notably, unlike the original transformer, ViT lacks a decoder.

The authors curated their dataset, dividing 162,174 images into training (112,378), validation (24,898), and test (24,898) sets with a ratio of 70:15:15. Both authentic and fake classes are balanced across all sets. CViT achieved high prediction accuracies: 91% on the FaceForensics++ dataset and 87.25% on the DFDC dataset during validation, with a 91.5% accuracy on the FaceForensics++ dataset during evaluation.

Earlier techniques have utilized deep learning architectures, including CNNs and RNNs, for detecting visual deepfakes.

For example:

- *The CNN-based model SCnet, introduced in [19], effectively captures forensic features using a gradually hierarchical stack of convolutional layers.*
- *LSTM networks were incorporated into hybrid CNN-LSTM models that utilized optical flow features for improved accuracy in video deepfake detection [20].*
- *The RCN model [21] uses CNNs and GRU cells to track temporal changes in video frames, leading to better recognition of altered faces.*

III. Audio-based Deepfake Detection

In audio synthesis, we begin by enumerating the deepfake techniques used to produce fake recordings.

1) Voice Conversion: Using artificial intelligence, this

method makes one person's voice sound like another. Gender-specific speech transformations are possible, for instance, without the need for the target speaker's voice data, thanks to AI models that can change a male voice to sound female or vice versa [22].

2) Text-to-Speech (TTS): This technique uses sophisticated deep learning models to produce spoken words from textual input. For instance, Google's WaveNet exhibits notable progress in TTS technology by generating remarkably lifelike speech synthesis from text [23].
3) Speech Synthesis with Emotion: Using this technique, artificial speech with particular emotional overtones is produced. To meet the contextual requirements of the message, it is conceivable, for example, to produce synthetic speech that can convey emotions like joy, despair, or rage [24].

Deepfake audio detection techniques fall into three main categories: waveform-based, image-based, and feature-based techniques.

Feature-Based Approaches: They rely on extracting features from audio transformations. These are used to extract both long-term and short-term features, such as Constant Q Cepstral Coefficients (CQCC) or Mel Frequency Cepstra coefficients (MFCC), which is shown in Figure 5. Hassan et al. [25] combined RNNs and the short-term spectrum characteristics in for detection.

Image-Based Approaches: these use computer vision algorithms to analyze audio spectrograms or mel-spectrograms as shown in Figure 5. Bartusiak et al. applied CNNs and transformer to normalized grayscale spectrograms for detection [26].

Waveform-Based Methodologies: These directly pass the raw audio waveform to deep neural networks. One hypothesis, also holds among researchers is shallow networks detect small artifacts and deep networks capture high-level features [27]. TSSDNet, where the ResNet-like and Inception blocks were presented alternatively [28].

After discussing a variety of audio deepfake generation and detection methods, let's delve into a recent method for audio deepfake detection that makes use of LVMs. We will specifically examine the SpotNet methodology [29].

The SpotNet framework, as shown in Figure 6, is a two- fold method for detecting synthetic audio that targets logical attacks like text-to-speech (TTS) and voice conversion (VC).

In the first fold, the raw audio signals are preprocessed to extract front-end spoofing features (FSF) that capture important temporal and spectral features, such as the spectral envelope, spectral contrast, and Mel-spectrogram. These features are intended to efficiently detect spoofing efforts that modify voice signal characteristics.

In the second fold, these FSF maps serve as input to the Logical Spoofing Transformer Encoder (LSTE), a model that leverages token embedding and transformer encoder blocks to extract deep attentive features from the speech data. These features are subsequently passed through a multi-layer spoofing classifier, which includes five dense layers, batch normalization, and dropout layers, improving the model's capacity to differentiate between authentic and fake audio.

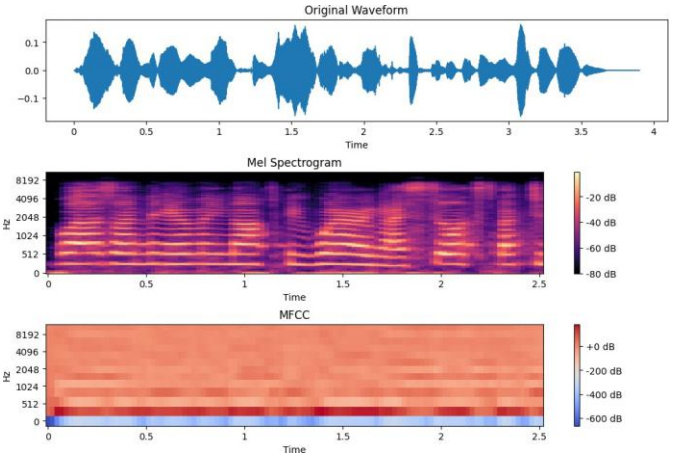

Fig. 5: Example of an audio signal alongside its corresponding Mel-spectrogram and MFCC.

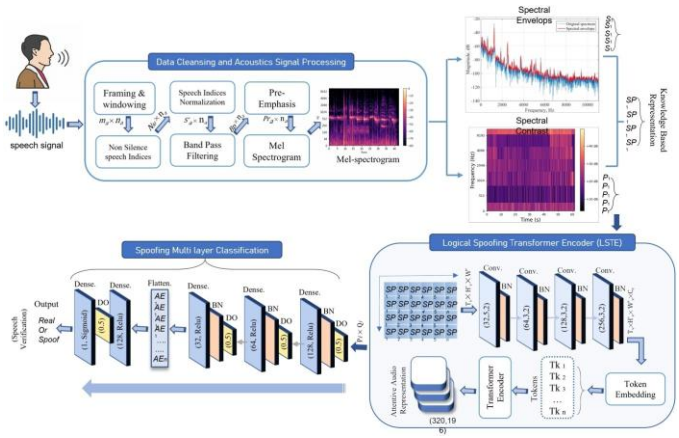

Fig. 6: The Framework of SpotNet for Audio Deepfake Detection [29].

The SpotNet architecture also includes several convolutional layers for further feature extraction and employs various preprocessing techniques such as windowing, non-silence index retrieval, normalization, bandpass filtering, and pre-emphasis filtering. This comprehensive preprocessing and feature extraction process guarantees that the model captures the most pertinent information for spoofing detection.

SpotNet was assessed using the ASVspoof2019-LA dataset [30], achieving an Accuracy of 93.91% and an Area Under the Curve (AUC) of 94.11%. The model demonstrated robustness against various spoofing algorithms, outperforming several state-of-the-art methods.

IV. Audio-Visual-Based Deepfake Detection

In the field of deepfake detection, a recent advancement is the investigation of multimodal methods that can utilize visual and audio characteristics in video data. Recently, in the area of video-based deepfake detection, there have been some

contributions that have looked at both visual and audio aspects of videos in a well-rounded fashion. This is an important advancement in the literature because it utilizes the potential of synergy, and applies multiple data modalities for improved detection of forgeries. LVMs have been incorporated into multimodal audio-visual deepfake detection using multiple modalities with both audio-visual streams. Therefore, in this section, we examine an large vision approach, as a novel and relevant way that considers audio and visual features for video-based deepfake detection. AVFakeNet introduced by Ilyas et al. [31], is an end-to-end model that also employs the Dense Swin Transformer for audio-visual deepfake detection.

This multimodal method consists of a dual-stream model, that incorporate separate models for audio and visual using Mel-Spectrograms for audio, and video frames for visual. The dense Swin transformer handles the classification of features in both audio and visual streams. The audio model uses Mel-Spectrogram features, while the visual model uses facial features identified in each video frame, in each case to perform audio-visual deepfake detection.

In terms of training, the AVFakeNet model uses the Celeb-DF dataset [32] to train the model on video, and ASVSpoof-2019 LA dataset [30] to train the audio model. Celeb-DF provides 590 real videos of 59 celebrities and 5,639 deepfake videos. The ASVSpoof-2019 LA dataset collects speech data of 107 individuals and includes video and audio data.

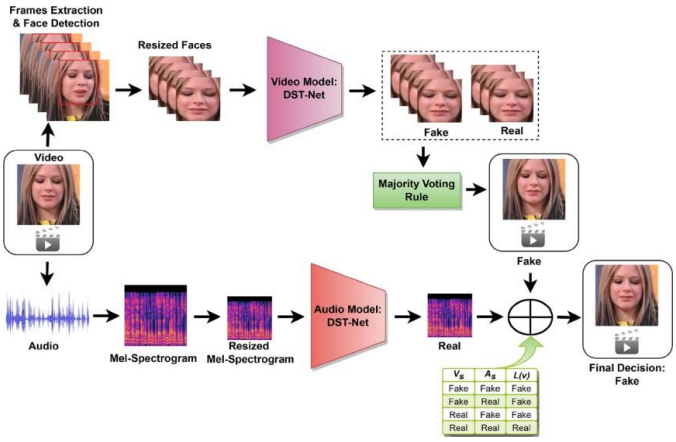

Fig. 7: Workflow of AVFakeNet for Multimodal Audio-Visual Deepfake Detection [31]

Performance evaluation was conducted on the FakeAVCeleb dataset [33], comprising four subsets: Real Audio Real Video (RaRv), Fake Audio Fake Video (FaFv), Real Audio Fake Video (RaFv), and Fake Audio Real Video (FaRv). The AVFakeNet model was assessed in three stages: visual stream only, audio stream only, and both audio and visual streams combined.

The outcomes provide evidence of the effectiveness of various configurations of the AVFakeNet model. The video-only model had a mean accuracy of 88.11%. The audio-only model performed significantly better with a mean accuracy of 96.62%. Combining both models into a multimodal architecture achieved a mean accuracy of 93.00%.

Additionally, there are other multimodal deepfake detection techniques that employ earlier deep learning architectures, such as EmoForen [34] and MDS [35].

V. Dataset Review

In this section, we will comprehensively review prominent datasets utilized in images, videos, and audio deepfake detection.

TABLE I
PROMINENT DATASETS FOR DETECTING DEEPFAKES IN IMAGES AND VIDEOS

Dataset Name	Modality	Real Size	Fake Size
CASIA1 [36]	Images	750	975
CASIA2 [36]	Images	7491	5123
CoMoFoD [37]	Images	4800	4800
DeeperForensics-1.0 [2]	Videos	50,000	10,000
Celeb-DF [32]	Videos	408	795
UADFDV [38]	Videos	49	49
FaceForensics [39]	Videos	1000	1000
FaceForensics++ [40]	Videos	1000	4000

Prominent datasets created to assess the efficacy of deepfake detection models across a variety of visual media modifications are listed in Table I. These datasets cover a variety of synthetic and modified material categories and include both images and videos. The variety of dataset sizes, manipulation techniques, and realism levels offers a thorough testing ground for detection models and enables researchers to compare how well they perform over a broad range of deepfake scenarios. Here is a detailed examination of each dataset's worth and features:

- ***CASIA1*** *and* ***CASIA2*** *are among the earliest datasets focused on image tampering. CASIA1 contains a limited number of tampered images, while CASIA2 offers a larger set. Both datasets provide images with various manipulations, primarily splicing and copy-move forgeries, making them essential for training and evaluating models on traditional image manipulation techniques. They are widely used for initial model testing, especially to assess a model's ability to detect localized, often subtle, visual changes.*
- ***CoMoFoD*** *dataset includes a balanced set of 4,800 real and manipulated images, providing complex and diverse modifications like splicing, retouching, and cloning. CoMoFoD is useful for testing detection models on intricate visual details, challenging the model to distinguish minute and sophisticated alterations. It is particularly relevant for research focused on high-fidelity forgeries and testing a model's sensitivity to different manipulation types within a consistent dataset structure.*
- *Known for its large scale and high-quality deepfake content,* ***DeeperForensics-1.0*** *includes 50,000 real videos and 10,000 manipulated ones. This dataset emphasizes realistic facial manipulations under various lighting, camera angles, and occlusions, which simulates real-world settings. It is invaluable for models focused on robustness, as the dataset's diversity in conditions tests a*

model's adaptability to environmental variations commonly found in real applications.

- *With a focus on deepfake videos created using advanced facial manipulation techniques,* ***Celeb-DF*** *contains 408 real and 795 fake videos of celebrities. It includes more natural facial expressions, lip-syncing, and movement, adding to the realism. This dataset is frequently used as a benchmark for models targeting video-based deepfakes, especially for training on realistic video manipulations where smooth facial transitions and lip synchronization are crucial.*
- *Although smaller in size,* ***UADFDV*** *offers a balanced set of real and manipulated videos with carefully controlled synthetic samples. It is often used for testing a model's performance in detecting novel and controlled tampering techniques in videos, which helps researchers explore model generalization on less common manipulation methods. UADFDV's manageable size also makes it suitable for testing models with limited computational resources.*
- ***FaceForensics*** *and* ***FaceForensics++*** *datasets have become standard benchmarks in video deepfake detection research. FaceForensics contains 1,000 real and 1,000 fake videos, while FaceForensics++ expands on this with 4,000 fake videos using multiple generation techniques, including DeepFakes, Face2Face, and Neural Textures. They offer incremental levels of quality (raw, lightly compressed, and heavily compressed) to test models on different compression artifacts. FaceForensics++ has particularly influenced deepfake research by providing both high-quality and degraded video versions, allowing models to be evaluated on their resilience to video quality loss and compression – conditions common on social media platforms.*

A number of well-known datasets for assessing and enhancing audio deepfake detection systems are listed in Table II. These datasets exhibit significant variations in structure, manipulation types, and intended applications, forming a comprehensive foundation for training and assessing models against a wide array of altered and synthetic audio. Below is a detailed overview of each dataset, highlighting their unique characteristics and contributions to the field of audio deepfake detection.

- ***WaveFake*** *Dataset contains 117,985 audio clips generated through various text-to-speech (TTS) methods, showcasing a rich variety of TTS generation techniques. It is entirely synthetic, which enables researchers to focus on detecting fine-grained artifacts associated with different synthesis methods, crucial for improving detection algorithms.*
- ***FoR*** *Dataset includes 195,000 utterances, consisting of both real human and computer-generated speech. It offers multiple versions of each audio sample, such as original, balanced, shortened, and re-recorded files. This diversity allows models to handle both synthetic and replayed manipulations effectively.*

TABLE II
PROMINENT AUDIO DEEPFAKE DETECTION DATASETS

Dataset Name	Summary
WaveFake dataset [41]	117,985 clips from various text-to-speech methods.
FoR dataset [42]	Versions include original, balanced, shortened, and re-recorded files.
ASVspoof 2015 [43]	Text-to-speech and voice conversion samples, split into training, development, and evaluation sets.
ASVspoof 2017 [44]	Replay sessions from different configurations with 42 speakers.
ASVspoof 2019 [30]	Logical and physical access replay attacks.
ASVspoof 2021 [3]	Includes LA and PA scenarios, plus a speech deepfake dataset.

- ***ASVspoof 2015*** *Dataset: This foundational dataset comprises a total of 2,500 samples, with 16,651 genuine samples and 246,500 spoofed samples. It provides insights into early voice conversion and synthesis methods, making it a crucial resource for training models to detect synthetic voices generated by those techniques.*
- ***ASVspoof 2017*** *Dataset: Focusing on replay attacks, this dataset features approximately 2,880 samples recorded in various acoustic environments and involves 42 speakers. It emphasizes the detection of replayed audio, where recorded real voices are played back to bypass authentication systems, enhancing its relevance in real-world scenarios.*
- ***ASVspoof 2019*** *Dataset expands upon its predecessors with a total of 5,000 samples (2,500 for each category: logical access and physical access). This dataset introduces diverse challenges by combining synthesized and converted voices with replay attacks, addressing both digital and physical spoofing scenarios.*
- ***ASVspoof 2021*** *Dataset: The most recent dataset in the series includes 145,669 genuine samples and 1,420,604 fake samples, providing a robust resource for training and evaluating models. It features dedicated segments for logical access and physical access, along with a speech deepfake segment that reflects current audio manipulation trends and techniques.*

In summary, each dataset offers unique advantages, whether through the type of manipulation, the environmental conditions captured, or the methods of audio generation. The extensive range of samples available across these datasets enables researchers to develop and refine robust detection models capable of addressing a wide array of real-world audio spoofing challenges, ultimately contributing to advancements in the field of audio deepfake detection.

TABLE III
PROMINENT AUDIO-VISUAL DEEPFAKE DETECTION DATASETS

Dataset Name	Total Size	Real Size	Fake Size
FakeAVCeleb [4]	25,000+	570	25,000+
DFDC [45]	128,154	23,654	104,500
DeepFake-TIMIT [46]	620	0	620
LAV-DF [47]	136,304	36,431	99,873
AV-Deepfake1M [48]	1,146,760	286,721	860,039
PolyGlotFake [49]	15,238	766	14,472

The datasets utilized for detecting multimodal audio-visual deepfakes, as outlined in Table III, include various deepfake video datasets that feature manipulated visual and audio streams. These datasets play a crucial role in the development of effective algorithms for audio-visual deepfake detection, providing the necessary data to train and evaluate such systems.

Each dataset contributes distinct features and challenges, enhancing the research landscape in combating audio-visual deepfakes.

- ***FakeAVCeleb*** *dataset consists of over 25,000 audio-visual clips featuring both real and manipulated content, with approximately 570 real and over 25,000 fake samples. This dataset focuses on showcasing advanced deepfake techniques, specifically emphasizing the manipulation of audio and visual streams. The variety of manipulation methods represented enables researchers to create models capable of effectively detecting a broad spectrum of deepfake scenarios, thereby serving as a crucial resource for enhancing multimodal detection systems. However, it does encounter the challenge of imbalanced data distribution, with real samples being significantly outnumbered.*
- ***DFDC*** *dataset includes 128,154 clips, comprising 23,654 real and 104,500 fake instances. It captures a comprehensive range of facial manipulation techniques, allowing researchers to study various types of alterations in depth. The balanced distribution of real and fake samples aids in reducing bias during model training. This dataset is instrumental in the evaluation and refinement of detection algorithms, addressing the challenges posed by different manipulation styles in real-world applications.*
- ***Celeb-DF*** *dataset contains 620 high-resolution videos, focusing entirely on manipulated celebrity faces, with all samples being fake. This dataset is characterized by its high quality and realistic alterations, providing a challenging environment for detection models. By featuring a variety of backgrounds and lighting conditions, Celeb-DF serves as a critical benchmark for developing advanced detection systems capable of identifying subtle manipulations, thus pushing the boundaries of current deepfake detection methodologies.*
- ***LAV-DF*** *dataset offers a total of 136,304 audio-visual clips, including 36,431 real and 99,873 fake samples. This dataset highlights the importance of replay attacks and synthetic manipulations, providing diverse examples for training detection algorithms. Its structured approach to varying audio-visual conditions equips researchers to create more robust models capable of discerning manipulated content across different scenarios.*
- ***AV-Deepfake1M*** *dataset features an extensive collection of 1,146,760 clips, including 286,721 real and 860,039 fake samples. This dataset aims to provide a large-scale resource for deepfake detection, allowing for the exploration of different manipulation techniques. Its significant size enables researchers to develop and validate their detection models on a wide range of data, addressing the challenges presented by high variability in deepfake characteristics.*
- *Finally,* ***PolyGlotFake*** *dataset comprises 15,238 audio-visual clips, with 766 real and 14,472 fake samples. This dataset is designed to explore the nuances of language and cultural differences in deepfake detection, offering a diverse range of manipulated audio-visual content. By incorporating different languages and accents, it helps researchers understand how language variations can affect detection performance, making it an essential tool for developing globally applicable deepfake detection systems. However, it also faces the challenge of imbalanced data distribution.*

VI. COMPARATIVE RESULTS AND ANALYSIS

This section offers an extensive comparative analysis of diverse deepfake detection methods within visual, audio, and audio-visual domains, including both traditional deep learning models and LVMs.

Table IV offers a comparative examination of key visual deepfake detection methods applied across various datasets, categorized by technique type, best results achieved, and modality. This table underscores the advancements in detection methods from early CNN and RNN approaches to models incorporating LVMs.

- ***Traditional CNN and RNN-Based Techniques:*** *Techniques like RCN (CNN + GRU) and standalone CNNs achieve good performance on datasets such as UADFV and Celeb-DF, but struggle with generalization on diverse datasets like Face-Forensics++, affecting their robustness despite computational efficiency.*
- ***MobileNet and Random Forest:***
  This lightweight combination is suitable for real-time applications but has lower accuracy (90.2%) compared to transformer-based models, illustrating the trade-off between efficiency and accuracy.
- ***Yolo-face, Bi-LSTM, and EfficientNet:***
  Integrating Bi-LSTM with EfficientNet provides reasonable accuracy (85.12%) on composite datasets but incurs higher computational costs and may require significant data preprocessing for better generalization.
- ***CNN with Vision Transformer (CViT):***
  This early LVM integration improves accuracy to 91.5% on datasets like Face-Forensics++ and DFDC, effectively

capturing global context, but faces challenges in real-time application due to processing demands.

TABLE IV
COMPARATIVE ANALYSIS OF METHODS FOR DETECTING VISUAL DEEPFAKES

Ref.	Year	Technique	Dataset	Best Result	Modality
[21]	2019	RCN (CNN and GRU)	Face-Forensics++	AUC: 97.5%	Videos
[50]	2020	MobileNet + Random Forest	DFDC	Acc: 90.2%	Videos
[51]	2020	CNN	UADFV, Face-Forensics++, Celeb-DF	Acc: UADFV (98.73%), FaceForensics++ (91.32%), CelebDF (98.85%)	Videos, Images
[52]	2021	Yolo-face, Bi-LSTM, EfficientNet	A combination of Face-Forensics++ and Celeb-DF	Acc: 85.12%	Videos
[18]	2021	CNN + ViT (CViT)	Face-Forensics++ and DFDC	Acc: 91.5%	Images
[53]	2022	CNN, Transformer Encoder	Face-Forensics	AUC: 99.93%, Acc: 99.67%	Images
[54]	2023	Local and Global Feature Maps	Face-Forensics++, Celeb-DF, DFDC	AUC: 97.66%	Videos
[55]	2023	Encoder-Decoder Transformer	Celeb-DF, Face-Forensics++	AUC: 99.0%	Videos

- ***CNN with Transformer Encoder:***
  Achieving the highest performance (AUC: 99.93%, Accuracy: 99.67%), this model utilizes a Transformer Encoder for enhanced feature extraction but may require optimization for scalability.
- ***Encoder-Decoder Transformer:***
  This architecture (AUC: 99.0%) efficiently handles temporal inconsistencies in deepfake videos but may be limited by high computational requirements for real-time use.

Table V provides a comparative review of key methods in audio deepfake detection, organized by technique, feature type, best performance, and dataset. This overview illustrates the progression from earlier CNN-based approaches to more complex architectures integrating Transformer models, highlighting the diverse strategies and challenges within audio deepfake detection.

- ***Traditional CNN-Based Techniques:***
  Early methods like Light CNN and ResNet-34 achieved Equal Error Rates (EER) between 0.04 and 0.05 on ASVspoof 2019 but struggle with complex audio forgeries, impacting their robustness across diverse datasets.

TABLE V
COMPARATIVE ANALYSIS OF METHODS FOR DETECTING AUDIO DEEPFAKES

Ref.	Technique	Features	Best Result	Dataset
[56]	Light CNN	Genuineness Features	EER = 0.04, t-DCF = 0.102	ASVspoof 2019
[57]	ResNet-34	Log Mel-spectrogram	EER = 0.05	ASVspoof 2019
[58]	ResNet-189	Log Mel-spectrogram	EER = 0.06, t-DCF = 0.157	ASVspoof 2019
[59]	Squeeze-Excitation	Low-level acoustic and whole utterance	EER = 0.59, t-DCF = 0.016	ASVspoof 2019
[60]	Transformer Encoder	LFCC	EER = 0.087, t-DCF = 37.67	ASVspoof 2019, ASVspoof 2021
[61]	Light CNN, Trans-former	Genuenization features	EER = 0.018, t-DCF = 0.102	ASVSpoof 2019
[62]	ResNet 189 with Trans-former Encoder	Direct feature extraction via deep learning	EER = 0.03	FoR, ASVSpoof 2019
[29]	CNNs, Trans-former	Direct feature extraction via deep learning	Acc = 93.91%, AUC = 94.11%	ASVSpoof 2019

- ***Enhanced CNN Variants: ResNet-189 and Squeeze-Excitation Networks:***
  ResNet-189 shows improved feature extraction with a slightly higher EER on ASVspoof 2019, while Squeeze-Excitation networks capture essential features but have limited generalization, indicating a need for further enhancements.
- ***Transformer-Enhanced Models: Transformer Encoder and Light CNN + Transformer:***
  Integrating Transformer Encoders with CNN models improves accuracy, with the Transformer Encoder achieving an EER of 0.087 and Light CNN + Transformer reaching 0.018 on ASVspoof 2019, though requiring higher computational resources.
- ***Advanced CNN-Transformer Hybrids: ResNet-189 with Transformer Encoder:***
  This combination delivers high performance (EER = 0.03) on datasets like FoR and ASVspoof 2019 but may be computationally intensive, limiting real-time applications without optimization.
- ***Comprehensive CNN and Transformer Models: SpoTNet (CNNs + Transformer):***
  SpoTNet achieves 93.91% accuracy and 94.11% AUC on ASVspoof 2019, balancing spatial and temporal feature extraction, but its computational cost may hinder scalability in practical applications.

In summary, both traditional techniques in visual and audio deepfake detection offer computational efficiency and simplicity for real-time applications on less complex datasets. These benefiting from their straightforward architecture, which aids in implementation.

TABLE VI
COMPARATIVE ANALYSIS OF METHODS FOR DETECTING AUDIO-VISUAL DEEPFAKES

Model	Year	DFDC		FakeAVCeleb	
		AUC (%)	ACC (%)	AUC (%)	ACC (%)
EmoForen [34]	2020	84.4	80.6	79.8	78.1
MDS [35]	2021	87.80	86.51	81.80	82.65
JointAV [63]	2021	82.5	83.3	90.2	91.9
AVFakeNet [31]	2022	82.8	86.2	78.4	83.4
BA-TFD [64]	2022	79.1	84.6	80.8	84.9
AVoiD-DF [65]	2023	91.4	94.8	83.7	89.2
MIS-AVoiDD [66]	2023	—	—	97.3	96.2
AVA-CL [67]	2024	84.20	88.64	86.55	89.47
AVT2-DWF [68]	2024	87.57	88.32	88.02	89.20

- *Pros:*
  - *Resource-efficient for real-time applications.*
  - *Simple architectures facilitate implementation.*
  - *Effective for basic manipulations in both audio and visual domains.*
- *Cons:*
  - *Limited generalization on complex datasets.*
  - *Reduced effectiveness for advanced manipulations across diverse datasets.*
  - *Lower accuracy when compared to more advanced models, such as transformer-based architectures.*

Conversely, LVMs, which are augmented by transformers, significantly enhance detection accuracy by effectively capturing complex patterns and dependencies. Despite their strong performance on advanced datasets, these models are often hindered by high computational demands, which can challenge their real-time usability.

- Pros:
  - *High accuracy for complex manipulations in both audio and visual tasks.*
  - *Strong capability in capturing global context and intricate features.*
  - *Robust performance on more challenging datasets.*
- Cons:
  - *High resource requirements limit usability for real-time applications.*
  - *Complexity in implementation and optimization can be a barrier.*
  - *May necessitate resource optimization for practical scalability in deployment.*

Table VI presents a comparative overview of audio-visual deepfake detection techniques evaluated on prominent datasets such as the DeepFake Detection Challenge (DFDC) and FakeAVCeleb. This table highlights performance metrics including AUC and Accuracy (ACC) to illustrate advancements in detection methodologies, particularly with the integration of transformers in multimodal contexts.

- ***Traditional CNN-Based Multimodal Approaches (EmoForen and MDS):***

  Early multimodal methods, such as the MDS that combines CNNs for image data and RNNs for sequential audio data. While computationally efficient, these approaches often struggle to capture complex intermodal relationships, resulting in limited generalizability on challenging datasets like DFDC, with their accuracy generally falling short of transformer-based approaches due to a reduced capacity to process high-level contextual information.
  - *Transformer-Augmented Models (MIS-AVoiDD and AVA-CL):*

    The incorporation of transformers in models like MIS- AVoiDD and AVA-CL enhances performance by capturing long-range dependencies in audio and visual streams. These models achieve higher accuracy on DFDC, effectively leveraging transformers for robust feature fusion, though their computational demands may limit real-time applicability unless optimized.
  - *Advanced Cross-Attention Mechanisms (AVT2- DWF):*

    AVT2-DWF employs cross-attention mechanisms for deeper audio-visual feature fusion, achieving high accuracy and AUC on datasets like FakeAVCeleb. This model highlights the importance of cross-attention for analyzing temporal and spatial inconsistencies but faces scalability challenges due to its complex architecture, necessitating a balance between accuracy and processing speed.

In summary, traditional multimodal approaches based on CNNs, such as MDS and EmoForen, deliver efficient computation but face challenges in capturing intricate

intermodal dynamics, resulting in limited adaptability to more complex datasets like DFDC.

- Pros:
  - Highly efficient for real-time processing.
  - Straightforward architecture supports easy implementation.
  - Capable of addressing basic deepfake manipulations effectively.
- Cons:
  - Struggles to generalize on more sophisticated datasets.
  - Less proficient in modeling complex intermodal relationships.
  - Performance lags behind that of transformer-based models.

In contrast, LVMs, which incorporate transformers and attention mechanisms, such as MIS-AVoiDD and AVA-CL, significantly improve detection accuracy by capturing long-range dependencies in audio and visual streams, particularly on challenging datasets like DFDC. However, their high computational demands can restrict real-time use. Models like AVT2-DWF further enhance audio-visual feature integration and achieve notable accuracy and AUC metrics on datasets like FakeAVCeleb, but their complex architecture presents scalability issues.

- Pros:
  - Superior accuracy for intricate audio-visual manipulations.
  - Strong capabilities for effective feature integration using transformers.
  - High performance in identifying temporal and spatial anomalies.
  - Efficiently integrates audio and visual information in complex scenarios.
- Cons:
  - Elevated computational needs can restrict real-time deployment.
  - Complexity in architecture may pose scalability challenges.
  - Implementation often necessitates careful optimization for efficiency.
  - Balancing detection accuracy with processing latency can be problematic.

Despite advancements in LVMs for deepfake detection, several areas require remaining research to maximize their potential. Improving generalization across diverse datasets is essential, as current models often perform well on specific datasets but struggle with others. Optimizing LVMs for realtime performance is also crucial due to their high computational demands.

Incorporating temporal dynamics into LVMs can enhance the detection of manipulated videos by analyzing the evolution of audio and visual features over time. While attention mechanisms aid in multimodal data fusion, further refinement is needed to improve the integration of audio and visual streams for detecting subtle manipulations.

Scalability poses another challenge, underscoring the need for remaining research on lighter model versions and techniques like pruning to enhance resource efficiency. Additionally, LVMs must improve robustness against evolving evasion tactics from deepfake creators, necessitating research into adversarial training methods.

By addressing these remaining research challenges, LVMs can significantly advance deepfake detection.

VII. CONCLUSION AND FUTURE WORK

This paper provides a comprehensive analysis of various deepfake detection techniques across visual, audio, and multimodal audio-visual domains, placing a significant emphasis on the latest advancements in deep learning, particularly with respect to LVMs. The findings of this study underscore the transformative potential of LVMs in enhancing deepfake detection capabilities, demonstrating their ability to efficiently process complex visual patterns and effectively manage large- scale datasets. Furthermore, this work includes a thorough survey of existing datasets commonly utilized in deepfake detection, which serves to contextualize the comparative analysis presented. Our comparisons highlight the notable superiority of LVMs over traditional deep learning methods, showcasing their advanced architectural features that enable them to outperform older techniques in distinguishing between real and synthetic media.

Exploring the following research avenues holds promise for significant advancements in deepfake detection, helping researchers develop more effective and reliable solutions:

1) **Expanding the Role of LVMs:** LVMs, with their ability to capture intricate visual patterns and contextual information, have significantly improved detection accuracy. Further advancing the use of LVMs in deepfake detection and developing more specialized architectures could lead to even more effective and adaptable detection systems.
2) **Datasets and Benchmarks:** Advance the creation of comprehensive, diverse, and publicly accessible datasets for assessing deepfake detection systems, particularly since many existing multimodal datasets suffer from imbalanced data distribution.
3) **Expanding Detection Scope:** Improve detection methods by incorporating the analysis of text overlays, motion patterns, and video context to identify more comprehensive manipulations.

REFERENCES

[1] HumanOrAI.IO, “Deepfake tools statistics,” https://humanorai.io/deepfake-tools-statistics, 2024.

[2] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy, “Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection,” arXiv preprint, 2020.

[3] He´ctor Delgado, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Xuechen Liu, Andreas Nautsch, Jose Patino, Md Sahidullah, Massimiliano Todisco, Xin Wang, Junichi Yamagishi, and ASVspoof Consortium, “Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan,” http://www.asvspoof.org/, 2021.

[4] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo, "Fakeavceleb: A novel audio-video multimodal deepfake dataset," arXiv preprint arXiv:2108.05080v4, Mar 2022.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[6] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[7] Franc¸ois Chollet, "Xception: Deep learning with depthwise separable convolutions," arXiv preprint arXiv:1610.02357, 2017.

[8] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger, "Densely connected convolutional networks," arXiv preprint arXiv:1608.06993, 2017.

[9] Sepp Hochreiter and Ju¨rgen Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, November 1997.

[10] Kyunghyun Cho, Bart Van Merrie¨nboer, Cem Gulcehre, Dzmitry Bahdanau, Dmitry Serdyuk, Yi Chu, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[14] OpenAI, "Gpt-4 technical report," Tech. Rep., OpenAI, 2023.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pam Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in Proceedings of the 38th International Conference on Machine Learning (ICML), 2021, pp. 8748–8763.

[16] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever, "Zero-shot text-to-image generation," 2021.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, and Bing Xu, "Generative adversarial networks," Advances in Neural Information Processing Systems, vol. 3, no. 11, 2014.

[18] Deressa Wodajo and Solomon Atnafu, "Deepfake video detection using convolutional vision transformer," arXiv preprint arXiv:2102.11126, 2021.

[19] Zhiqing Guo, Lipin Hu, Ming Xia, and Gaobo Yang, "Blind detection of glow-based facial forgery," Multimedia Tools and Applications, vol. 80, no. 5, pp. 7687–7710, February 2021.

[20] Pallabi Saikia, Dhwani Dholaria, Priyanka Yadav, Vaidehi Patel, and Mohendra Roy, "A hybrid cnn-lstm model for video deepfake detection by leveraging optical flow features," arXiv preprint arXiv:2208.00788, 2022.

[21] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," arXiv preprint arXiv:1905.00582, 2019.

[22] Berrak Sisman, Haizhou Zhang, Xue Li, and Junichi Yamagishi, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," in IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020, vol. 29, pp. 132–157, IEEE.

[23] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.

[24] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Robust speaker-adaptive hmm-based text-to-speech synthesis," IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 6, pp. 1208–1230, 2009.

[25] Faizan Hassan and Ahsan Javed, "Voice spoofing countermeasure for synthetic speech detection," in Proceedings of the International Conference on Artificial Intelligence, Settat, Morocco, April 2021, pp. 209–212.

[26] Eric R Bartusiak and Edward J Delp, "Frequency domain-based detection of generated audio," in Proceedings of the IS&T International Symposium on Electronic Imaging: Media Watermarking, Security, and Forensics, Burlingame, CA, January 2021, pp. 273–1–273–7.

[27] Jayant Khochare, Chintan Joshi, Bhushan Yenarkar, Swapnil Suratkar, and Farhaan Kazi, "A deep learning framework for audio deepfake detection," Arabian Journal for Science and Engineering, vol. 47, pp. 3447–3458, November 2021.

[28] Guoqing Hua, Andrew Beng Jin Teoh, and Haizhou Zhang, "Towards end-to-end synthetic speech detection," IEEE Signal Processing Letters, vol. 28, pp. 1265–1269, June 2021.

[29] Awais Khan and Khalid Mahmood Malik, "SpoTNet: A Spoofing-Aware Transformer Network for Effective Synthetic Speech Detection," in Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation (MAD '23), Thessaloniki, Greece, June 2023, ACM, Association for Computing Machinery.

[30] Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi Kinnunen, Ville Vestman, Massimiliano Todisco, Hector Delgado, Md Sahidullah, Junichi Yamagishi, and Kong Aik Lee, "Asvspoof 2019: Spoofing countermeasures for the detection of synthesized, converted, and replayed speech," 2019.

[31] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik, "Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection," Applied Soft Computing, vol. 136, pp. 110124, 2023.

[32] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3207–3216.

[33] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "Fakeavceleb: a novel audio-video multimodal deepfake dataset," 2021.

[34] Tanya Mittal, Ujjwal Bhattacharya, Rishabh Chandra, Aniket Bera, and Dinesh Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in MM, 2020, pp. 2823–2832.

[35] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian, "Not made for each other-audio-visual dissonance-based deepfake detection and localization," in Proceedings of the 28th ACM International Conference on Multimedia. ACM, 2020, pp. 439–447.

[36] J. Dong, W. Wang, and T. Tan, "Casia image tampering detection evaluation database," in 2013 IEEE China summit and international conference on signal and information processing. IEEE, 2013, pp. 422–426.

[37] Various, "Columbia image splicing detection evaluation dataset - list of photographers," 2004, Accessed 16 Mar 2022.

[38] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai created fake videos by detecting eye blinking," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018.

[39] Andreas Ro¨ssler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces," arXiv preprint arXiv:1803.09179, 2018.

[40] Andreas Ro¨ssler and et al, "Faceforensics++: Learning to detect manipulated facial images," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[41] Joel Frank and Lea Scho¨nherr, "Wavefake: A data set to facilitate audio deepfake detection," arXiv preprint, November 2021.

[42] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," in Proceedings of the 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), October 2019, pp. 1–10.

[43] Zhizheng Wu, Junichi Yamagishi, Tomi Kinnunen, Cemal Hanilc¸i, Mohammed Sahidullah, Aleksandr Sizov, Nicholas Evans, Massimiliano Todisco, and He´ctor Delgado, "Asvspoof: The automatic speaker verification spoofing and countermeasures challenge," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 4, pp. 588–604, 2017.

[44] Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Kong Aik Lee, Md Sahidullah, Massimiliano Todisco, and He´ctor Delgado, "Asvspoof 2017: Automatic speaker verification spoofing and countermeasures

challenge evaluation plan," http://www.asvspoof.org/index2017.html, 2018.
[45] 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.
[46] Pavel Korshunov and Sebastien Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," arXiv preprint arXiv:1812.08685, vol. 2, pp. 3, 8, 2018.
[47] Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat, "Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization," arXiv preprint arXiv:2204.06228v2, May 2023.
[48] Z. Cai, S. Ghosh, A. P. Adatia, M. Hayat, A. Dhall, and K. Stefanov, "Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset," arXiv preprint arXiv:2311.15308, 2023.
[49] Y. Hou, H. Fu, C. Chen, Z. Li, H. Zhang, and J. Zhao, "Polyglotfake: A novel multilingual and multimodal deepfake dataset," arXiv preprint arXiv:2405.08838, 2024.
[50] Maulik Patel, Anil Gupta, Shubham Tanwar, and Mohammad Obaidat, "Trans-df: a transfer learning-based end-to-end deepfake detector," in 2020 IEEE 5th international conference on computing communication and automation (ICCCA). IEEE, 2020, pp. 796–801.
[51] D Xie et al., "Deepfake detection on publicly available datasets using modified alexnet," in 2020 IEEE symposium series on computational intelligence (SSCI), 2020.
[52] Aya Ismail, Marwa Elpeltagy, Mervat Zaki, and Kamal A. ElDahshan, "Deepfake video detection: Yolo-face convolution recurrent approach," PeerJ Comput Sci, vol. 7, pp. e730, 2021.
[53] Ziyu Xue, Qingtong Liu, Haichao Shi, Ruoyu Zou, and Xiuhua Jiang, "A transformer-based deepfake-detection method for facial organs," Electronics, vol. 11, no. 24, 2022.
[54] Tong Wang, Hao Cheng, Ka Pui Chow, and Li Nie, "Deep convolutional pooling transformer for deepfake detection," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 19, no. 6, pp. 1–20, 2023.
[55] Faizan Khalid, Muhammad Haris Akbar, Saifullah Gul, Sajid Tanwar, and Mohammad Obaidat, "Swynt: Swin y-net transformers for deepfake detection," in 2023 International Conference on Robotics and Automation in Industry (ICRAI), 2023, pp. 1–6.
[56] Z Wu, RK Das, J Yang, and H Li, "Light convolutional neural network with feature genuinization," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2021, pp. 6349–6353.
[57] PR Aravind, U Nechiyil, and N Paramparambath, "Audio spoofing verification using deep convolutional neural networks by transfer learning," arXiv preprint arXiv:2008.03464, 2020.
[58] M Alzantot, Z Wang, and MB Srivastava, "Deep residual neural networks for audio spoofing detection," arXiv preprint arXiv:1907.00501, 2019.
[59] C-I Lai, N Chen, J Villalba, and N Dehak, "Assert: Antispoofing with squeeze-excitation and residual networks," arXiv preprint arXiv:1904.01120, 2019.
[60] T-P Doan, L Nguyen-Vu, S Jung, and K Hong, "Bts-e: Audio deepfake detection using breathing-talking-silence encoder," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Jun 2023, pp. 1–5.
[61] Z. Wu, R. Kumar Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," 2020, arXiv preprint.
[62] Z. Zhang, X. Yi, and X. Zhao, "Fake speech detection using residual network with transformer encoder," in Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, June 2021, pp. 13–22.
[63] Y. Zhou and S.-N. Lim, "Joint audio-visual deepfake detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14800–14809.
[64] Zhaoyang Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat, "Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization," 2022.
[65] Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren, "Avoid-df: Audio-visual joint learning for detecting deepfake," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 2015–2029, 2023.
[66] V. Sree Katamneni and A. Rattani, "Mis-avoidd: Modality invariant and specific representation for audio-visual deepfake detection," arXiv e-prints, p. arXiv–2310, 2023.
[67] Yibo Zhang, Weiguo Lin, and Junfeng Xu, "Joint audio-visual attention with contrastive learning for more general deepfake detection," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 20, no. 5, pp. 137, January 2024.
[68] Rui Wang, Dengpan Ye, Long Tang, Yunming Zhang, and Jiacheng Deng, "Avt2-dwf: Improving deepfake detection with audio-visual fusion and dynamic weighting strategies," arXiv preprint arXiv:2403.14974, March 2024, Submitted on 22 Mar 2024.