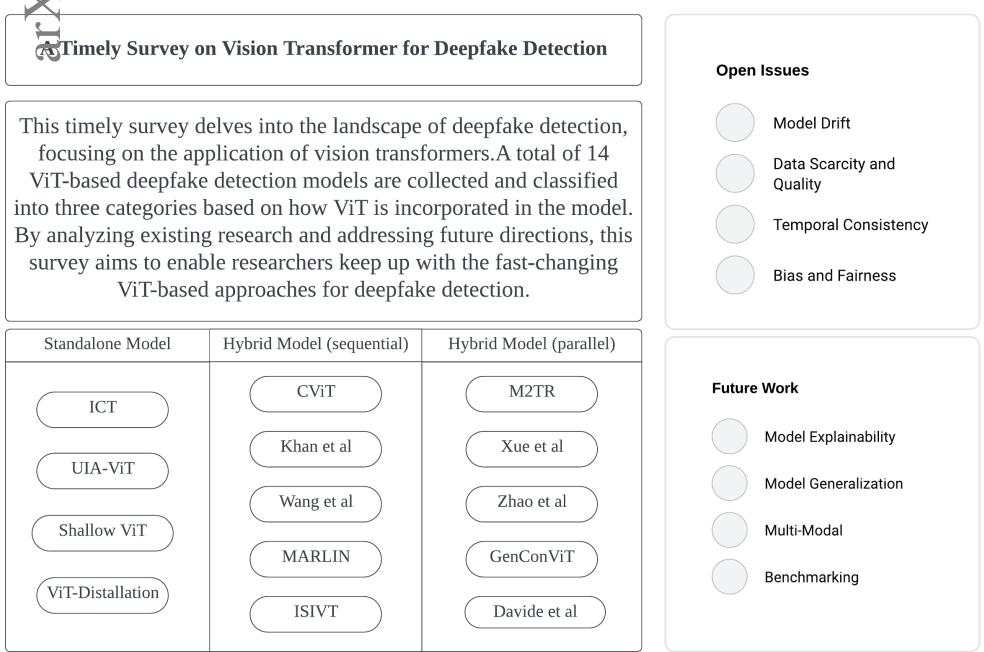


Graphical Abstract

A Timely Survey on Vision Transformer for Deepfake Detection

Zhikan Wang^{1,2}, Zhongyao Cheng¹, Jiajie Xiong^{1,2}, Xun Xu¹, Tianrui Li³, Bharadwaj Veeravalli², Xulei Yang¹



Highlights

A Timely Survey on Vision Transformer for Deepfake Detection

Zhikan Wang^{1,2}, Zhongyao Cheng¹, Jiajie Xiong^{1,2}, Xun Xu¹, Tianrui Li³, Bharadwaj Veeravalli², Xulei Yang¹

- This survey identifies ViT-based deepfake detection as a cutting-edge research direction, highlighting the need for a timely literature review to keep up with rapidly evolving ViT-based approaches.
- This survey provides an up-to-date overview (as of February 28, 2024) of a total of 14 ViT-based deepfake detection models categorized into standalone, sequential, and parallel architectures.
- This survey also explores unresolved issues in deepfake detection and suggests potential directions for future research.

A Timely Survey on Vision Transformer for Deepfake Detection

Zhikan Wang^{1,2}, Zhongyao Cheng¹, Jiajie Xiong^{1,2}, Xun Xu¹, Tianrui Li³, Bharadwaj Veeravalli², Xulei Yang¹

¹*Institute for Infocomm Research (I²R), A*STAR, Singapore*

²*National University of Singapore (NUS), Singapore*

³*Southwest Jiaotong University (SWJTU), China*

Abstract

In recent years, the rapid advancement of deepfake technology has revolutionized content creation, lowering forgery costs while elevating quality. However, this progress brings forth pressing concerns such as infringements on individual rights, national security threats, and risks to public safety. To counter these challenges, various detection methodologies have emerged, with Vision Transformer (ViT)-based approaches showcasing superior performance in generality and efficiency. This survey presents a timely overview of ViT-based deepfake detection models, categorized into standalone, sequential, and parallel architectures. Furthermore, it succinctly delineates the structure and characteristics of each model. By analyzing existing research and addressing future directions, this survey aims to equip researchers with a nuanced understanding of ViT's pivotal role in deepfake detection, serving as a valuable reference for both academic and practical pursuits in this domain.

Keywords: Deepfakes, Vision Transformer, Deep Learning, Face Manipulation.

1. Introduction

Deepfakes involve the generation of images or videos that convincingly depict events or individuals who may not exist or have never engaged in the portrayed activities. These sophisticated manipulations of multimedia content, often driven by deepfake generative models, have the potential to mislead, manipulate, and pose significant threats to a variety of areas, in-

cluding politics [1], media [2], and personal privacy [3]. Driven by advances in deep learning, deepfake has emerged as a more powerful tool for creating highly realistic synthetic media, raising significant concerns about the potential misuse and manipulation of visual content, which underscores the importance of the development of efficient detection mechanisms.

To address the problems posed by deepfake, various detection methods have been proposed. Traditional techniques for detecting deepfakes utilize image and video analysis methods, such as local feature extraction and motion analysis, to identify manipulated content. More recently, a variety of deep learning methodologies, including long short-term memory (LSTM), recurrent neural network (RNN), and hybrid approaches, have been introduced for the detection of manipulated images and videos [4]. Within the diverse array of existing methodologies, approaches built on Vision Transformer (ViT) have emerged as state-of-the-art in current research.

This timely survey delves into the landscape of deepfake detection, focusing on the application of vision transformers (ViTs) [5] – a novel neural network architecture that have demonstrated remarkable success in computer vision tasks. ViTs depart from the traditional convolutional neural networks (CNNs) by building upon self-attention mechanisms, allowing them to capture global dependencies in data sequences [6]. The comprehensive modeling of global context by ViTs enhances its efficacy in capturing overarching features and relationships within images. This capability is proven advantageous for discerning subtle details and characteristics within an image. Consequently, in the domain of deepfake detection, the utilization of ViTs presents a distinctive avenue.

The motivations behind this survey are multifaceted. Firstly, deepfake detection is crucial for safeguarding the integrity of visual content and mitigating the potential harm caused by deceptive multimedia. Secondly, the unique characteristics of ViTs, with their attention-based architectures, offer a promising paradigm for addressing the challenges inherent in deepfake detection. Finally, ViT-based deepfake detection is a cutting-edge research direction, with many researchers working in parallel. There is a need for a timely literature review to keep up with the fast-changing ViT-based approaches to enable researchers to understand ViT's features better and develop robust strategies. This survey aims to provide a comprehensive overview of existing ViT-based deepfake detection models, their strengths, limitations, open issues, and future directions.

The rest of the article is organized as follows: In Section 2, we provide a

basic introduction pertaining to the generation and detection of deepfakes, along with a navigation of related surveys. Section 3 is devoted to ViT-based approaches for detecting deepfakes. Furthermore, Section 4 delves into open issues and potential directions for future research. The survey concludes with a summary in Section 5.

2. Related Works

2.1. Deepfake Generation

Deepfakes are generated through the manipulation of pre-existing videos and images, resulting in the creation of content that exhibits a convincing semblance of authenticity despite being entirely synthetic. Deepfake generation can be categorized into four main groups: identity swap, face reenactment, attribute manipulation, and entire face synthesis [7]. Identity swap encompasses the substitution of an individual’s facial features in an image or video with those of another person. Face reenactment entails the modification of the facial expressions exhibited by an individual within an image or video. In the domain of attribute manipulation deepfakes, there is the manipulation of specific facial attributes, including skin tone, age, gender, and the incorporation or removal of elements like glasses. The synthesis of entire faces involves the generation of facial samples portraying individuals that do not exist in reality [8].

2.2. Traditional Methods for Deepfake Detection

Traditional approaches to tackling deepfake-related challenges often refer to methods and algorithms developed before the advent of advanced deep learning models. These methods typically involve manual inspection, forensic analysis [9], or rule-based approaches [10] to identify anomalies and inconsistencies within multimedia content. Various classic pattern recognition methods have been utilized for deepfake detection, encompassing logistic regression, probabilistic linear discriminant analysis, random forest, gradient boosting decision tree, extreme learning machine, k-nearest neighbor, support vector machine, Gaussian mixture model, etc [11]. Although these approaches may provide some level of effectiveness, their limitations become more evident as deepfake technologies become increasingly complex and sophisticated. Consequently, contemporary efforts have shifted towards harnessing advanced deep learning techniques, such as deep neural networks and ViTs, to improve the accuracy and efficiency of deepfake detection.

2.3. Deep Learning for Deepfake Detection

Deep learning, a machine learning technique based on neural networks, has gained significant attention in recent years due to its remarkable achievements in diverse fields. It has also shown promising results in deepfake detection. CNN [12], RNN [13], LSTM [14], and ViT are the most widely used deep learning techniques in deepfake detection. CNNs excel in capturing spatial dependencies within images, making them particularly effective in discerning facial manipulations and other visual anomalies in frame-level detection. In video-level detection, RNNs and LSTMs are well-suited for the sequential video frames. Their recurrent architectures enable the modeling of temporal dependencies and can capture subtle changes over time (e.g., facial expressions, eye blinks, lip movement). ViTs enhance the understanding of the overall structure of an image, facilitating the identification of inconsistencies or anomalies that may be indicative of manipulation.

2.4. ViT-based Deepfake

Transformers, primarily applied to high-level vision tasks, face challenges in low-level vision tasks like image super-resolution due to the complexity of detailed image generation [15]. ViTs primarily focus on learning global representations and might lack the fine-grained modeling required for generating realistic and highly detailed facial features necessary for deepfakes, their application in generating detailed features for tasks like deepfake generation is limited. While originally designed for vision-related tasks like image recognition, ViTs offer unique advantages in analyzing and understanding the intricate details of deepfake images and videos. By leveraging the power of transformers and attention mechanisms, ViTs provide a new avenue for enhancing deepfake detection algorithms and advancing the field of multimedia forensics. In this article, we explore the application of ViTs in deepfake detection, highlighting their potential, advantages, and challenges in combatting the proliferation of manipulated media content. The details of ViT-based deepfake detection techniques will be discussed in Section 3.

2.5. Surveys on Deepfake

In the years since deepfake appeared, a number of surveys have been conducted on deepfake in the literature. The authors of [7] reviewed the theoretical concepts, foundation, and classification of the deepfake for both generation and detection. A similar study, ref. [16], discussed deep fake generation, detection, datasets, challenges, and research directions. In [4] and [17], an

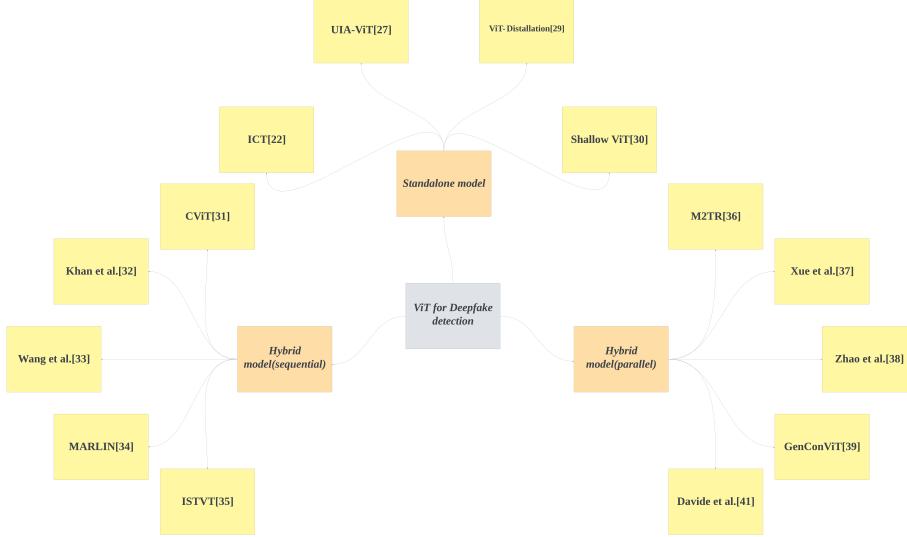


Figure 1: An overview of the main models discussed in this survey.

in-depth exploration of the application of deep learning in deepfake is given. Yi et al. in [11] provided a more nuanced perspective of investigation, focusing on the audio deepfake detection. [18] provided a comprehensive overview of deepfake and underlined publicly available deepfake generation tools and datasets for benchmarking. In addition, [19], [20], [21] have also investigated deepfake to varying degrees. In spite of the numerous surveys conducted on the subject of deepfake, there is a notable scarcity of surveys that address the deepfake technology associated with ViTs. Our survey serves as a valuable supplement in this specific direction, providing researchers with a more focused and nuanced reference in the realm of deepfake technology, particularly pertaining to ViTs.

3. Vision Transformer in Deepfake Detection

When it comes to the task of deepfake detection using ViTs, various methods can be categorized into two broad categories: ViT as standalone models and ViT combined with other techniques as hybrid models. Hybrid models can be further classified into sequential models and parallel models based on how the ViT is combined. The classification of the main models is illustrated in Fig. 1, with additional details about each model are summarized

Model	Dataset	Performance	Source Code (accessed on 8 January 2024)	Year	
Standalone model					
ICT [22]	FaceForensics++ [23], Deeper [24], Celeb-DF-v1 [25], Celeb-DF-v2 [25], DFD [26]	AUC= 90.22 AUC= 93.57 AUC= 81.43 AUC= 85.71 AUC=84.13	https://github.com/LightDXY/	2022	
	FaceForensics++(HQ) [23], DFD [26], Celeb-DF-v2 [25], Celeb-DF-v1 [25], DFDC [28]	AUC=99.33 AUC*=94.68 AUC*=82.41 AUC*=86.59 AUC*=75.80			
	ViT-Distillation [29]	DFD [28]	AUC=97.80, f1=91.9	https://github.com/smu-ivpl/DeepfakeDetection	2021
	RFF [30], Shallow ViT [30]	ACC=92.15, AUC=92.00 ACC=88.52, AUC=88.00	NA	2022	
	DRFFD [30]	ACC=88.52, AUC=88.00	NA	2022	
Hybrid model (Sequential)					
CViT [31]	FaceForensics++(HQ) [23], DFD [28]	ACC=93.75 ACC=91.5, AUC=91	https://github.com/erprogs/CViT	2021	
	FaceForensics++(HQ) [23], DFD [26], DFDC [28]	ACC=99.79 ACC=99.28 ACC=91.69			
Khan et al. [32]	FaceForensics++ [23], DFDC [28], Celeb-DF [25], DF-1.0 [24]	ACC=92.11, AUC=97.66 ACC*=65.76, AUC*=73.68 ACC*=63.27, AUC*=72.43 ACC*=62.46, AUC*=78.19	https://github.com/sohailahmedkhan/	2021	
	MARLIN [34]	FaceForensics++ [23]	ACC=90.71,AUC=93.77		2023
ISTVT [35]	FaceForensics++(HQ) [23]	ACC=99.0	NA	2023	
	FaceForensics++(LQ) [23]	ACC=96.15			
	Celeb-DF [25], DFDC [28]	ACC=99.8 ACC=92.1			
Hybrid model (Parallel)					
M2TR [36]	FaceForensics++(LQ) [23], FaceForensics++(HQ) [23], FaceForensics++(RAW) [23], Celeb-DF [25], SR-DF [36]	ACC=92.89, AUC=95.31 ACC=97.93, AUC=99.51 ACC=99.50, AUC=99.92 AUC=95.5 AUC=86.7	https://github.com/wangjk666/	2022	
	FaceForensics++(LQ) [23], FaceForensics++(HQ) [23], FaceForensics++(RAW) [23], DFD [26], DFDC [28], Celeb-DF [25]	ACC=94.14, AUC=96.43 ACC=98.12, AUC=99.67 ACC=99.67, AUC=99.93 AUC*=94.32 AUC*=75.93 AUC*=82.43			
	Xue et al. [37]	ACC=95.33 ACC=83.53 ACC=99.18 ACC=97.67			
	Zhao et al. [38]	ACC=98.50 ,AUC=99.90 ACC=97.00,AUC=99.6 ACC=98.28 ACC=90.94			
GenConViT [39]	DFDC [28], FaceForensics++ [23], TIMIT [40], Celeb-DF-v2 [25]	ACC=98.50 ,AUC=99.90 ACC=97.00,AUC=99.6 ACC=98.28 ACC=90.94	https://github.com/erprogs/genconvit	2023	
	David et al. [41]	AUC=95.10			

Table 1: Summary of ViT based deepfake detection methods.* means cross-datasets evaluation

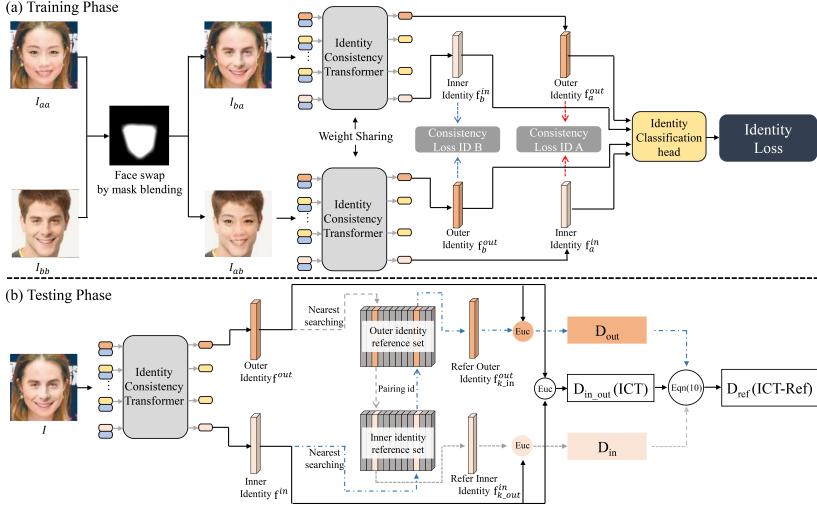


Figure 2: **Overview architecture of ICT.** (a) Training phase and (b) Testing phase. (The figure is taken from [22])

in Table. 1.

3.1. Standalone Models

In this approach, a ViT model is trained specifically for deepfake detection. The ViT is trained on a large dataset of real and deepfake images to learn the distinguishing patterns and features that can differentiate between authentic and manipulated content. This method leverages the self-attention mechanism of ViTs to capture spatial relationships within the image, enabling the detection of anomalies or inconsistencies introduced by deepfake techniques.

Dong et al. in [22] proposed Identity Consistency Transformer (ICT), a novel face forgery detection method that focuses on high-level semantics, specifically identity information, and detecting a suspect face by finding identity inconsistency in inner and outer face regions. Fig. 2 presents the training and testing phases of ICT. While testing, identity reference sets are used to enhance the identity detection of certain celebrities. Trained on the MS-Celeb-1M [42] dataset, ICT achieved 98.56%, 93.17%, 96.41%, 94.43%, and 99.25% AUC on FF++ [23], DFD [26], Celeb-DF-v1 [25], Celeb-DF-v2 [25], and Deeper [24] datasets, respectively. It achieves state-of-the-art performance not only across different datasets but also across various types of

image degradation forms found in real-world applications including deepfake videos. ICT presents several merits: (i) It does not rely on any particular facial forgery method, rather assumes the existence of identity inconsistency, showcasing robust generalization capabilities. (ii) It can utilize publicly accessible genuine facial images as a reference set, thereby improving detection performance, particularly in scenarios involving facial forgeries of celebrities. (iii) It exhibits resilience against diverse forms of image degradation, including scaling, noise, and video encoding, making it well-suited for real-world applications. Despite its many advantages, the method has one major drawback: the method mainly detects fake faces with inconsistent identities, and may not be able to detect facial reproduction results with consistent identities.

ICT demonstrates that intra-frame inconsistency is very effective for deepfake detection, but requires additional pixel-level forged location annotations. To acquire such annotations, some existing methods generate large-scale synthesized data with location annotations, which are only composed of real images and cannot capture the properties of forgery regions. Others generate forgery location labels by subtracting paired real and fake images, yet such paired data is difficult to collect and the generated label is usually discontinuous. To address these issues, Zhuang et al. proposed a novel Unsupervised Inconsistency-Aware method based on Vision Transformer (UIA-ViT) in [27]. As shown in Fig. 3, UIA-ViT has two key components: UPCL (Unsupervised Patch Consistency Learning) and PCWA (Progressive Consistency Weighted Assemble). The process consists of four main steps: (i) Feature Extraction: Employing ViT for facial image feature extraction, the method decomposes the image into a series of blocks and utilizes self-attention mechanisms to learn relationships among these blocks. (ii) Unsupervised Forgery Localization Estimation: Utilizing Multivariate Gaussian (MVG) estimation to represent features of genuine and forged image blocks, and generating pseudo-labels. By comparing the Mahalanobis distance between the features of image blocks and genuine or forged distributions, this approach approximates the location of the forged region without requiring pixel-level annotations. (iii) Block Consistency Learning: Leveraging the self-attention mechanism in ViT to acquire consistency relationships among features of different blocks. Through the design of a consistency loss function, this method constrains attention maps across different layers to capture internal inconsistencies in forged images. (iv) Progressive Consistency Weighted Combination: Generating consistency-aware features using ViT’s classification embeddings and

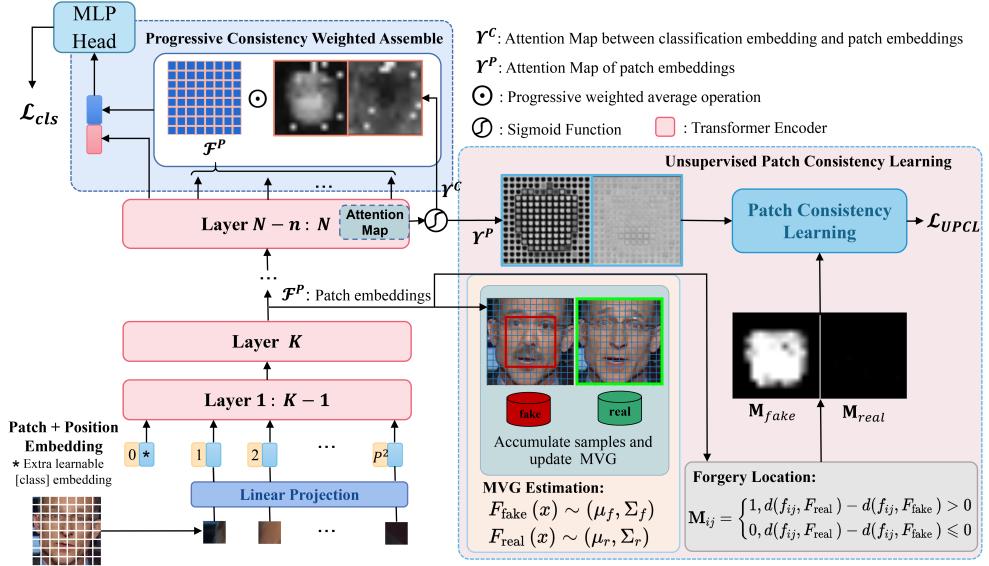


Figure 3: Overview architecture of UIA-ViT. (The figure is taken from [27])

block embeddings. By introducing a variable weighting function, this approach progressively combines classification embeddings and block embeddings, thereby gradually enhancing the final detection performance. Similar to ICT, UIA-ViT evaluates detection performance across datasets. Trained on FF++, UIA-ViT gained 99.33%, 94.68%, 82.41%, 86.59%, 75.80% AUC on FF++ [23], DFD [26], Celeb-DF-v2 [25], Celeb-DF-v1 [25], DFDC [28], respectively. The strengths of UIA-ViT lie in its utilization of the self-attention mechanism of the Vision Transformer. This enables effective capture of both local and global information within images. Additionally, it can learn inconsistencies in facial images without needing pixel-level annotations.

In [29], an approach combining ViT and distillation learning is presented. Compared to a conventional ViT model, it incorporates not only class token into the input feature vectors for training the network to distinguish between real and forged videos but also introduces distillation tokens for learning knowledge from a teacher network (EfficientNet [43]), which aims to enhance the network’s generalization capabilities. The model attained 97.8% AUC and 91.9 f1 score on DFDC [28], which outperformed the SOTA model. In contrast to adding modules to ViT, Shaheen Usmani et al. [30] proposed a

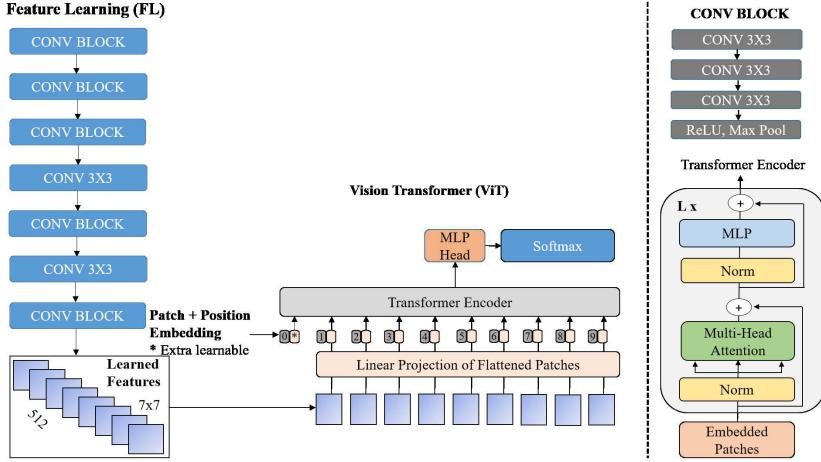


Figure 4: **Overview architecture of CVIT** (The figure is taken from [31])

shallow ViT for deepfake detection. The model has 16.48 times fewer parameters and approximately 2.97 times fewer FLOPS than the baseline ViT. The model performs well even with insufficient training data and computational resources and provides better results than existing deep forgery detection methods (92.15% ACC and 0.92 AUC on RFF dataset, 88.52% ACC and 0.88 AUC on RFFD dataset). Due to its lightweight and efficiency, the shallow ViT is particularly suitable for deep forgery detection in real-time scenarios such as social media.

3.2. Hybrid Models (Sequential Structure)

Hybrid models combine ViTs with other architectures, such as Convolutional Neural Networks (CNNs), to benefit from the strengths of both approaches. The hybrid models may use the initial layers of a CNN for low-level feature extraction, followed by a ViT for higher-level feature learning and detection. This combination allows the model to capture both local details and global context effectively, enhancing the overall detection performance.

In [31], Deressa Wodajo et al. made an initial attempt to combine ViT with CNNs. The proposed Convolutional Vision Transformer (CVIT) is presented in Fig. 4. It combines the learning capabilities of Convolutional Neural Networks (CNN) and ViTs. CNN excels in learning local features of images, while ViTs can learn both local and global features. This combined capability enables the model to associate with each pixel of the image and comprehend the relationships among non-local features. The CVIT model was

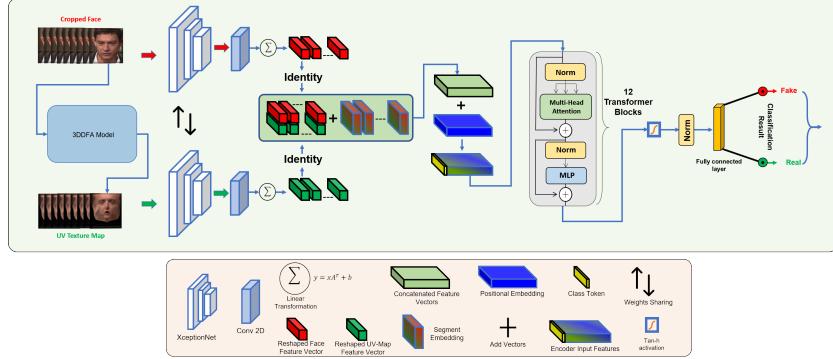


Figure 5: **Overview architecture of Khan’s model.** (The figure is taken from [32])

trained on a diverse collection of facial images that were extracted from the DFDC dataset and performed well on the DFDC, UADFV, and FaceForensics++ datasets. Nevertheless, the model exhibits suboptimal performance on FF++ FaceShifter dataset. This can be attributed to the inherent difficulty in learning visual artifacts, and it is likely that the proposed CVIT did not effectively capture and understand these artifacts. Khan et al. [32] proposed a novel video transformer with incremental learning for detecting deepfake videos. As shown in Fig. 5, the proposed model utilizes both the original facial image and the UV texture map for feature extraction. The UV texture map serves the purpose of preventing information loss when the facial orientation is not frontal. Aligning facial images to the UV map provides information about poses, blinking, and mouth movements. Extracting features from both the facial image and its corresponding UV texture map offers a more comprehensive and precise approach to capturing and utilizing facial features. In addition, an incremental learning strategy was adopted to fine-tune the proposed model on new datasets without sacrificing its performance on previous datasets. Following incremental learning on seven distinct datasets, the model exhibited outstanding performance with accuracy scores of 99.79%, 99.28%, and 91.69% on the FF++ [23], DFD [26], and DFDC [28] datasets, respectively. It is worth noting that they conducted corresponding ablation experiments. The experimental results indicate that under identical conditions, the hybrid model outperformed the standalone ViT where the former achieved 99.28% AUC, 98.92% F1-Score, and 99.28% ACC on FF++, while the latter only attained 77.10% AUC, 68.71% F1-Score, and 73.26% ACC.

Wang et al. [33] not only combined ViT with CNN, but also made some improvements to ViT on this basis. A CNN with a kernel size identical to the feature map dimensions considers convolutional computations involving all feature blocks and their relationships. However, such operations reduce the feature map to one dimension, significantly discarding crucial features and diminishing model performance. The utilization of a pooling transformer is employed to adjust the feature dimensions for image analysis. The proposed Deep Convolutional Pooling Transformer initially undergoes feature extraction through CNN convolution. The extracted features are then input into depthwise separable convolution to obtain Q, K, V, which are subsequently fed into the pooling transformer. The model was trained in a self-made keyframe dataset based on FF++ [23], cross-dataset evaluated on DFDC [28], Celeb-DF [25], DF-1.0 [24]. Researchers discovered that deep ViT models may encounter the issue of attention collapse, where attention maps become excessively similar, leading to a noticeable performance decline with increasing model depth. To address this, re-attention mechanism is employed to preserve attention map diversity.

In [35], a video-level deepfake detection model was proposed. ISTVT (Interpretable Spatial-Temporal Video Transformer), which consists of a novel decomposed spatial-temporal self-attention and a self-subtract mechanism to capture spatial artifacts and temporal inconsistency, demonstrated strong performance on multiple datasets. MARLIN [34] itself is not related to ViT, but their proposed feature extractor MARLIN combined with ViT achieved SOTA performance in FF++ [23], which is much better than MARLIN combined with CNNs.

3.3. Hybrid Models (Parallel Structure)

Some existing methods employ attention mechanisms to fuse information from both ViTs and other models. This approach combines the strengths of ViTs in capturing global relationships with other models that excel at detecting local artifacts or inconsistencies. Attention-based fusion methods dynamically weigh the contributions of different models or features based on their relevance, enhancing the overall deepfake detection performance.

Davide et al. [41] presented a deepfake detection model combining EfficientNet and ViTs. In contrast to the current state-of-the-art methodologies, they do not employ distillation or ensemble techniques. The top-performing model attained an AUC of 0.951 and an F1 score of 88.0%, which is in

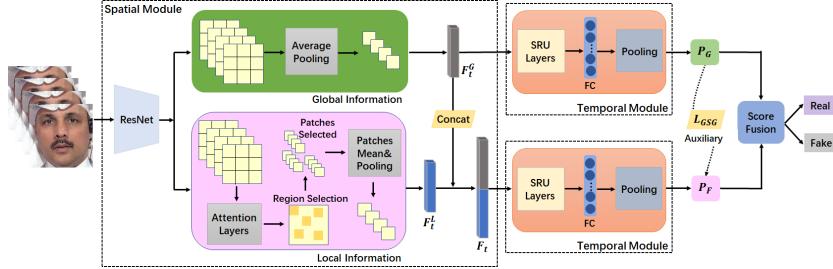


Figure 6: Overview architecture of Zhao' model. (The figure is taken from [44])

close proximity to the current state-of-the-art performance on the DFDC [28] dataset.

In [44], Zhao et al. proposed a novel model consist of a spatial module and a temporal module as shown in Fig. 6. The spatial module comprises global information flow and local information flow. The input frame sequences undergo feature extraction through these two streams, and the extracted features, after fusion, are then input into the temporal module to capture temporal detection information. The attention layer of the ViT is applied to the local part of the spatial module. They also designed a novel regularization loss called the Global Stream Guidance (GSG) loss, used to guide the selection of local information and the extraction of temporal information in the fusion stream. This loss facilitates the comprehensive utilization of the inherent complementary advantages of both global and local information. Intra-dataset and cross-dataset evaluations were conducted on FF++ [23], DFDC [28], and Celeb-DF [25], resulting in SOTA performance. Furthermore, they conducted ablation studies on the sampling frame numbers and model complexity, arriving at a conclusion similar to [33]: for deep ViT models, deeper and larger does not necessarily lead to better performance.

Xue et al. [37] introduced a DeepFake detection method specifically tailored for subtle facial expression manipulation, facial detail alterations, and blurred images. The proposed method framework includes an organ-selection module, a facial-region interception module, an organ-level transformer, and a classifier. The organ-level transformer identifies deepfake features through organs, utilizing an organ-selection module to reduce the weight of compromised, damaged, and low-quality organs, thereby enhancing accuracy. Simultaneously, a full-face ViT is employed to assist in detecting partial information. The accuracy experiences significant improvement when an in-

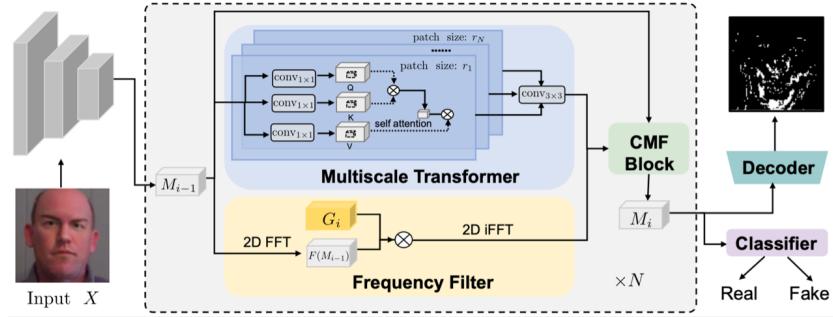


Figure 7: **Overview architecture of M2TR.** (The figure is taken from [36])

creased number of key organ features are used. This method has achieved outstanding performance on numerous datasets, particularly reaching SOTA results on datasets with low-resolution data for its organ detection doesn't rely on high-resolution data.

Generative Convolutional Vision Transformer (GenConViT) was proposed in [39]. GenConViT comprises two independently trained networks and four main modules: Autoencoder (AE), Variational Autoencoder (VAE), ConvNeXt layer, and Swin Transformer. The first network employs AE to generate the latent feature (LF) space of input images, maximizing the probability of category predictions, i.e., determining whether the given input is a deepfake. The second network uses VAE to reconstruct images while simultaneously maximizing category prediction probability and minimizing the loss distance between sample images and reconstructed images. The AE and VAE models extract LF from input video frames, capturing hidden patterns and correlations in the learned visual artifacts of deepfake manipulation. GenConViT achieved classification accuracies of 98.5%, 98.28%, 97%, and 90.94% on DFDC [28], TIMIT [40], FF++ [23], and Celeb-DF [25], respectively, demonstrating its robust performance.

In [36], the authors proposed a Multi-modal Multi-scale Transformer (M2TR), which operates on patches of different sizes to detect local inconsistencies in images at different spatial levels. As presented in Fig. 7, the approach aims to capture the subtle manipulation artifacts at different scales by using transformer models. Apart from RGB information, M2TR further learns to detect forgery artifacts in the frequency domain through a carefully designed cross modality fusion block. The combination of multiscale transformer and frequency filter enables M2TR to extract effective

features from highly compressed images. This capability results in outstanding performance across three different resolution datasets in FF++, particularly excelling in low-resolution datasets compared to other methods. Ablation experiments in the paper indicate that the multiscale transformer is the most crucial component of M2TR. The absence of the multiscale transformer module leads to a significant performance drop across various datasets, particularly in low-quality datasets, with a decrease in accuracy approaching 6%.

4. Benchmark

As shown in Table 1, these models have demonstrated exceptional proficiency on their designated datasets. Nonetheless, given the significant variation in the feature distribution across these datasets, it would be imprudent to directly compare their accuracy metrics to ascertain their relative superiority or inferiority. Therefore, we have decided to reproduce and test models with open-source code on the same datasets. The datasets we decide to use are FaceForensics++(FF++)[23] and Celeb-DF[25].

FaceForensics++ [23] is a comprehensive benchmark dataset designed for the evaluation of digital face manipulation detection algorithms. It extends its predecessor, FaceForensics, by incorporating a richer and more diverse set of manipulated facial videos. The dataset encompasses a wide array of manipulations, including but not limited to deepfake generation, face swapping, facial reenactment, and more traditional computer graphics-based modifications. We mainly selected 4 different fake videos (Deepfakes, Face2Face, FaceSwap, NeuralTextures), and the original video for training and testing. In addition, FF++ has three different compression qualities, we chose raw to test the performance of the model, and low quality to test the robustness of the model to compressed video.

Celeb-DF[25] is a large-scale, challenging dataset specifically designed for the detection of deepfake videos, with a focus on featuring celebrities. This dataset stands out due to its high-quality deepfake video generation, which significantly reduces common artifacts associated with deepfake content, such as unnatural blinking patterns, facial distortions, and poor lip-syncing. By prioritizing the realism of the deepfakes, Celeb-DF provides a rigorous benchmark for deepfake detection algorithms, aiming to mirror the sophistication and quality of deepfakes encountered in real-world scenarios.

It is worth mentioning that although each model uses the same dataset, in the data preprocessing section, we process the input data of each model according to the respective original paper in order to maximise the reproduction of the best performance of the model. For all these model, we didn't use fine tuning, so the performance could not be great as it in their original implement. But in this way, we can avoid cheery picking and gain a relatively more equitable outcome.

The results are listed in table 2. The CViT model seems to struggle relative to the others, which could indicate potential areas for improvement, such as feature extraction or model architecture adjustments. Actually GENCViT is an improved model based on CViT, it achieve much better performance, particularly on the raw FF++ data, which suggests that it may be better at handling less-processed data. M2TR model consistently shows high performance across all three datasets in both accuracy (ACC) and area under the receiver operating characteristic curve (AUC) metrics, indicating robustness and generalizability. Most of the models show large performance degradation when faced with compressed video. Most of the models show large performance degradation when faced with compressed video. However, khan et al [32] and M2TR [36] still maintain good performance, the former by UV texture map and the latter by frequency domain feature extraction.

Model	FF++(LOW)	FF++(RAW)	Celeb-DF
David et al [41]	50.71% ACC 0.763 AUC	81.85% ACC 0.902 AUC	81.27% ACC 0.870 AUC
Khan et al [32]	77.14% ACC 0.723 AUC	86.44% ACC 0.907 AUC	73.78% ACC 0.843 AUC
CViT [31]	58.78% ACC 0.647 AUC	69.16% ACC 0.675 AUC	70.27% ACC 0.679 AUC
GENCViT [39]	48.56% ACC 0.884 AUC	97.68% ACC 0.997 AUC	90.95% ACC 0.981 AUC
M2TR [36]	87.19% ACC 0.904 AUC	95.82% ACC 0.987 AUC	98.46% ACC 0.999 AUC

Table 2: Benchmark over 3 datasets

5. Open Issues & Future Work

In this section, we investigate open challenges and future research directions regarding the application of vision transformers in deepfake detection.

5.1. Open Issues

5.1.1. Model Drift

Existing deepfake detection systems face challenges in keeping pace with the continuous evolution of deepfake techniques. The frozen nature of these systems upon deployment results in a growing gap between their capabilities and the rapidly advancing methods employed by deepfake generators. There is a pressing need to develop adaptive detection mechanisms that can dynamically evolve to counter emerging deepfake strategies, ensuring ongoing effectiveness in the ever-changing landscape of deepfake.

5.1.2. Data Scarcity and Quality

ViT-based deepfake models require extensive training data, particularly high-quality data of the target individual’s face or voice. The limited availability of high-quality data can hinder the training process. For instance, generating deepfakes of lesser-known individuals or private individuals may be challenging due to the scarcity of suitable training data. Additionally, the quality and diversity of training data significantly impact the realism and diversity of generated content. A lack of diverse training data can lead to biased or unrealistic deepfakes.

5.1.3. Temporal Consistency

Maintaining temporal consistency, especially in video deepfakes, remains a challenge. Transformers generate content frame-by-frame, and ensuring that each frame seamlessly transitions from the previous one is complex. Any temporal inconsistency, such as abrupt facial movements or unnatural lip synchronization, can be a telltale sign of a deepfake. Techniques for improving temporal coherence, such as recurrent neural networks (RNNs) or Long Short-Term Memory (LSTM) structures, are still being researched and integrated with transformers to address this issue.

5.1.4. Bias and Fairness

Transformer models, including those used in deepfakes, can inherit biases from their training data. This bias can manifest in the generated content and

reinforce societal prejudices. Techniques for bias mitigation, such as fairness-aware training and data preprocessing, are being researched. Achieving fairness and mitigating bias in transformer-based deepfakes requires a nuanced understanding of the underlying biases and careful model design to promote equitable content generation.

5.2. Future work

5.2.1. Improving Model Explainability

Typically, in the existing literature, the majority of deep-learning-based deepfake detection techniques fail to provide a comprehensive explanation for the final outcome of their detection process. Enhancing interpretability is crucial for researchers to understand how ViT models make decisions in deepfake detection and adjust the model according to the outcome.

5.2.2. Improving Model Generalization

ViT-based deepfake detection models may struggle to generalize well across diverse deepfake datasets. It can be seen from Table 1 that most of the existing methods' performances decrease remarkably when evaluated cross datasets. New techniques such as incremental learning and transfer learning will be researched to enhance generalization and robustness of models.

5.2.3. Enhancing Multi-Modal Approaches

Detecting deepfakes is mainly based on visual information. A few models also use temporal and frequency domain information. More robust deepfake detection should consider multiple modalities, investigating the fusion of audio, text, or contextual information with visual cues. There is significant research scope for developing novel architectures for multimodal deep fake detection.

5.2.4. Benchmarking and Evaluation Standards

In terms of deepfake research, many of the methods are not published in open source code and hard to reproduced. Direct comparison of the performance of different models is inaccurate due to differences in experimental environments and model deployment. The promotion of reproducible results should be encouraged through the provision of extensive datasets, experimental configurations, and open-source codes. Establish comprehensive benchmarks and standards for evaluating deepfake detection is essential for future development.

6. Conclusion

This survey has provided an in-depth exploration of the technical intricacies and advancements within the realm of transformer-based deepfake detection technology. Our exploration of this intricate landscape has unveiled the remarkable capabilities of ViT-based models in identifying forged content across diverse modalities. It is clear that transformers have brought about a paradigm shift in the field, showcasing cutting-edge performance and unparalleled versatility. The survey also delves into open issues and future directions, providing a roadmap for further research. Additionally, we present detailed information on the datasets employed, performance metrics, and access links to the source codes of each discussed model. It is hoped that this survey will inspire researchers who are interested in further integrating ViT into deepfake detection, providing them with valuable insights and convenience in their endeavors.

References

- [1] D. Fallis, The epistemic threat of deepfakes, *Philos. Technol.* 2021, 34, 623–643 (2021).
- [2] N. Giansiracusa, How algorithms create and prevent fake news: Exploring the impacts of social media, deepfakes, gpt-3, and more, Apress: Berkeley, CA, USA; ISBN 978-1-4842-7154-4 (2021).
- [3] J. Hancock, J.T.; Bailenson, The social impact of deepfakes, *Cyberpsychol. Behav. Soc. Netw.* 2021, 24, 149–152 (2021).
- [4] A. Almars, Deepfakes detection techniques using deep learning: A survey., *Journal of Computer and Communications*, 9, 20-35. (2021).
- [5] A. Dosovitskiy, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929v2 (2021).
- [6] B. Chen, et al., Glit: Neural architecture search for global and local image transformer, IEEE/CVF Int. Conf. Comput. Vis., pp. 12–21 (2021).
- [7] Z. Akhtar, Deepfakes generation and detection: A short survey, *Journal of Imaging* (2023).

- [8] F. Juefei-Xu, et al., Countering malicious deepfakes: Survey, battle-ground, and horizon, arXiv:2103.00218 (2022). [arXiv:2103.00218](https://arxiv.org/abs/2103.00218)
- [9] M. T. Jafar, et al., Forensics and analysis of deepfake videos, in: 2020 11th ICICS, 2020, pp. 053–058. doi:[10.1109/ICICS49469.2020.939493](https://doi.org/10.1109/ICICS49469.2020.939493).
- [10] S. Agarwal, et al., Protecting world leaders against deep fakes, in: CVPR Workshops, 2019.
URL <https://api.semanticscholar.org/CorpusID:195732375>
- [11] J. Yi, et al., Audio deepfake detection: A survey, arXiv:2308.14970 (2023). [arXiv:2308.14970](https://arxiv.org/abs/2308.14970)
- [12] S. Tariq, et al., Detecting both machine and human created fake face images in the wild, in: Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, MPS ’18, Association for Computing Machinery, 2018, p. 81–87. doi:[10.1145/3267357.3267367](https://doi.org/10.1145/3267357.3267367).
URL <https://doi.org/10.1145/3267357.3267367>
- [13] E. Sabir, et al., Recurrent convolutional strategies for face manipulation detection in videos, Interfaces (GUI) 3 (1) (2019) 80–87.
- [14] D.-C. Stanciu, B. Ionescu, Deepfake video detection with facial features and long-short term memory deep networks, in: 2021 ISSCS, 2021, pp. 1–4. doi:[10.1109/ISSCS52333.2021.9497385](https://doi.org/10.1109/ISSCS52333.2021.9497385).
- [15] O.-J. K. Sonain Jamil, Md. Jalil Piran, A comprehensive survey of transformers for computer vision, drones (2023).
- [16] A. Naitali, et al., Deepfake attacks: Generation, detection, datasets, challenges, and research directions, Computers (2023).
- [17] T. T. Nguyen, et al., Deep learning for deepfakes creation and detection: A survey, Computer Vision and Image Understanding (2022).
- [18] J. W. Seow, et al., A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities, Neurocomputing 513 (2022) 351–371. doi:<https://doi.org/10.1016/j.neucom.2022.09.135>.

- [19] M. S. Rana, M. N. Nobi, B. Murali, A. H. Sung, Deepfake detection: A systematic literature review, *IEEE Access* 10 (2022) 25494–25513. doi:[10.1109/ACCESS.2022.3154404](https://doi.org/10.1109/ACCESS.2022.3154404).
- [20] A. Malik, M. Kuribayashi, S. M. Abdullahi, A. N. Khan, Deepfake detection for human face images and videos: A survey, *IEEE Access* 10 (2022) 18757–18775. doi:[10.1109/ACCESS.2022.3151186](https://doi.org/10.1109/ACCESS.2022.3151186).
- [21] Y. Patel, et al., Deepfake generation and detection: Case study and challenges, *IEEE Access* 11 (2023) 143296–143323. doi:[10.1109/ACCESS.2023.3342107](https://doi.org/10.1109/ACCESS.2023.3342107).
- [22] X. Dong, et al., Protecting celebrities from deepfake with identity consistency transformer, arXiv preprint arXiv:2203.01318v3 (2022).
- [23] A. Rossler, et al., Faceforensics++: Learning to detect manipulated facial images, in: In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, p. 1–11.
- [24] L. Jiang, et al., Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, arXiv:2001.03024 (2020). arXiv:2001.03024.
- [25] Y. Li, et al., Celeb-df: A large-scale challenging dataset for deepfake forensics, arXiv:1909.12962 (2020). arXiv:1909.12962.
- [26] S. Das, et al., Improving deepfake detection using dynamic face augmentation, arXiv:2102.09603v1 (02 2021).
- [27] W. Zhuang, et al., Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection, arXiv:2210.12752v1 (2022).
- [28] B. Dolhansky, et al., The deepfake detection challenge (dfdc) dataset, arXiv:2006.07397 (2020).
- [29] Y.-J. Heo, et al., Deepfake detection scheme based on vision transformer and distillation, arXiv:2104.01353v1 (2021).
- [30] D. S. Shaheen Usmani, Sunil Kumar, Efficient deepfake detection using shallow vision transformer, *Multimed Tools Appl* (2023).

- [31] D. Wodajo, S. Atnafu, Deepfake video detection using convolutional vision transformer, arXiv:2102.11126 (2021).
- [32] H. D. Sohail Ahmed Khan, Video transformer for deepfake detection with incremental learning, arXiv:2108.05307v1 (2021).
- [33] T. Wang, et al., Deep convolutional pooling transformer for deepfake detection, arXiv:2209.05299v4 (2023).
- [34] Z. Cai, et al., Marlin: Masked autoencoder for facial video representation learning, arXiv:2211.06627 (2023). [arXiv:2211.06627](#).
- [35] C. Zhao, et al., Istvt: Interpretable spatial-temporal video transformer for deepfake detection, IEEE Transactions on Information Forensics and Security 18 (2023) 1335–1348. doi:[10.1109/TIFS.2023.3239223](https://doi.org/10.1109/TIFS.2023.3239223).
- [36] J. Wang, et al., M2tr: Multi-modal multi-scale transformers for deepfake detection, arXiv:2104.09770v3 (2022).
- [37] Z. Xue, et al., A transformer-based deepfake-detection method for facial organs, *electronics* (2022).
- [38] H. Zhao, et al., Protecting celebrities from deepfake with identity consistency transformer, arXiv:2203.01265v1 (2022).
- [39] D. Wodajo, et al., Deepfake video detection using generative convolutional vision transformer, arXiv:2307.07036 (2023). [arXiv:2307.07036](#).
- [40] C. Sanderson, B. C. Lovell, Multi-region probabilistic histograms for robust and scalable identity inference, in: *Advances in Biometrics*, Springer, 2009, pp. 199–208.
- [41] D. A. Cocomini, et al., Combining EfficientNet and Vision Transformers for Video Deepfake Detection, Springer International Publishing, 2022, p. 219–229.
- [42] Y. Guo, et al., Ms-celeb-1m: A dataset and benchmark for large-scale face recognition, arXiv:1607.08221 (2016). [arXiv:1607.08221](#).
- [43] Q. V. L. Mingxing Tan, Efficientnet: Rethinking model scaling for convolutional neural networks, arXiv:1905.11946v5 (2020).

- [44] X. Zhao, Y. Yu, R. Ni, Y. Zhao, Exploring complementarity of global and local spatiotemporal information for fake face video detection, in: ICASSP 2022, 2022, pp. 2884–2888. doi:10.1109/ICASSP43922.2022.9746061.