Customized Transformer Adapter With Frequency Masking for Deepfake Detection

Zenan Shi, Haipeng Chen, Yixin Jia, Dong Zhang, *Member, IEEE*, Wei Lu, *Member, IEEE*, and Xun Yang

***Abstract*—The rapid advancement of AI-generated content has intensified concerns over deepfakes due to increasingly sophisticated and visually convincing forgeries. To this end, the pre-trained Vision Transformer (ViT) model has become a de facto choice for deepfake detection, thanks to its powerful learning capability. Despite favorable results achieved by existing ViT-based methods, they have inherent limitations that could result in suboptimal performance in scenarios with continuously evolving forgery techniques, such as overfitting to single forgery patterns or placing excessive emphasis on dominant forgery regions. In this paper, we propose CUTA, a simple yet effective deepfake detection paradigm that utilizes ViT adapters as the medium and fully exploits the spatial- and frequency-domain features of given images to overcome the limitations of existing methods. Specifically, CUTA focuses on *frequency domain masking* within the input space, which obscures parts of the high-frequency image to intensify the training challenge while preserving subtle forgery cues in the frequency domain to facilitate comprehensive forgery representations. Furthermore, we propose two task-customized modules within the ViT model, *i.e.*, the *texture enhancement module* and the *multi-scale perceptron module*, to seamlessly integrate local texture and rich contextual features. These two modules ensure an organic interaction between the task-specific forgery patterns and general semantic features within the pre-trained ViT framework. The experimental results on several publicly available benchmarks demonstrate CUTA's superiority in performance, particularly showcasing its significant advantages in both cross-dataset and cross-manipulation scenarios. Code and models are available at https://github.com/Zenanshi92/CUTA**

Received 29 August 2024; revised 21 February 2025 and 21 April 2025; accepted 18 May 2025. Date of publication 29 May 2025; date of current version 18 June 2025. This work was supported in part by the Key Projects of Science and Technology Development Plan of Jilin Province under Grant 20230201088GX and in part by the National Natural Science Foundation of China under Grant 62276112 and Grant 62441237. The associate editor coordinating the review of this article and approving it for publication was Dr. Daniel Moreira. *(Corresponding author: Dong Zhang.)*

Zenan Shi and Haipeng Chen are with the College of Computer Science and Technology, Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China (e-mail: shizn@jlu.edu.cn; chenhp@jlu.edu.cn).

Yixin Jia is with the College of Software, Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China (e-mail: yxjia23@mails.jlu.edu.cn).

Dong Zhang is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China (e-mail: dongz@ust.hk).

Wei Lu is with the School of Computer Science and Engineering, Ministry of Education Key Laboratory of Information Technology, Guangdong Province Key Laboratory of Information Security Technology, Sun Yat-sen University, Guangzhou 510006, China (e-mail: luwei3@mail.sysu.edu.cn).

Xun Yang is with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: xyang21@ustc.edu.cn).

Digital Object Identifier 10.1109/TIFS.2025.3574983

***Index Terms*—Deepfake detection, vision transformer, ViT adapter, frequency domain masking.**

I. Introduction

THE rise of the advanced artificial intelligence generated content approach, exemplified by Sora [2], [3], heralds a new stage in the evolution of deepfake creation [4], [5], [6]. Deepfake technology, facilitated by various off-the-shelf and open-source methods like Face2Face [7], FSGAN [8], and SimSwap [9], allows for identity swaps, expression alterations, and attribute modifications, resulting in visually untraceable videos. Consequently, concerns have emerged regarding the potential misuse of creating fake videos that fabricate people's words and actions [10], [11]. Examples include telecommunications fraud, child abductions, and the use of celebrity faces in live commerce events [12]. These developments pose significant challenges and security risks in various fields, including journalism, politics, and military affairs, endangering the stability of nations and societies [13]. Under these circumstances, there is an urgent imperative to develop powerful deepfake detection techniques capable of precisely and autonomously discerning fake videos to mitigate these pressing threats [14].

Previous methods for deepfake detection [15], [16], [17], [18] traditionally frame the challenge as a standard binary classification task. They employ off-the-shelf backbones, such as convolutional neural networks and Vision Transformer (ViT) architectures [19], [20], to extract global facial features, followed by a binary classifier used to distinguish between authentic and forged faces [21], [22]. It is well known that training a ViT model for deepfake detection requires a large dataset to achieve good generalization performance. To address this, as shown in Figure 1(a), traditional deepfake detection methods initialize the model using ImageNet pre-trained ViT weights from open-source model zoos. The model is then trained on face forgery data. However, these works utilize the pre-trained model by fine-tuning either partial or full model weights of the ViT backbone. Such utilization is straightforward but inefficient. As shown in Figure 1(b), recent research on Efficient Parameter Transfer Learning (EPTL) has shown a more efficient way of utilizing pre-trained ViT models for deepfake detection. The DF-Adapter [23] pioneered the application of adapter techniques in the field of deepfake detection, which quickly adapts a pre-trained ViT with a dedicated dual-level adapter. While attention mechanisms are effective in assigning feature weights based on relevance, showing

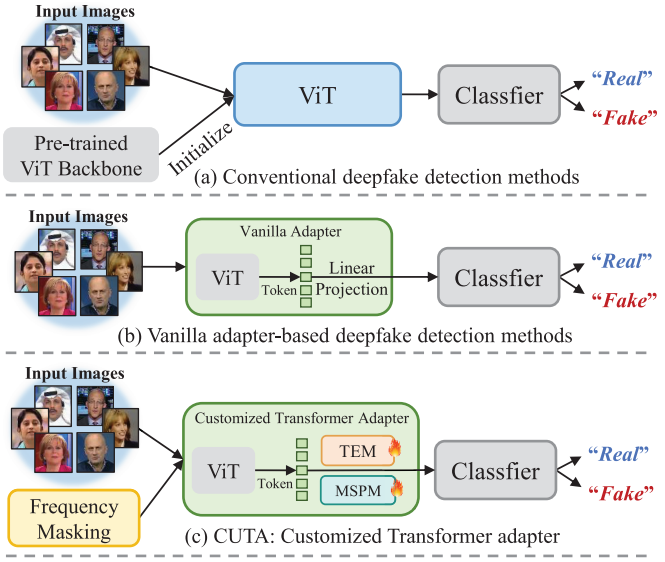


Fig. 1. (a) In the traditional learning paradigm of training a Vision Transformer (ViT) for deepfake detection, a pre-trained ViT model is used for initialization, leveraging the knowledge from the pre-trained datasets. (b) In the cutting-edge Efficient Parameter Transfer Learning paradigm, adapter modules are integrated into a pre-trained ViT model. During training, only the parameters of adapter modules are updated, while the pre-trained parameters remain fixed [1]. The straightforward adoption of the plain ViT-Adapter for deepfake detection fails to extract task-specific forgery patterns. (c) Our proposed *Customized Transformer Adapter* overcomes this limitation by emphasizing frequency domain masking and introducing two task-customized modules, thereby enabling more efficient fine-tuning of the pre-trained ViT and achieving superior performance over previous methods, particularly in terms of generalization ability and robustness.

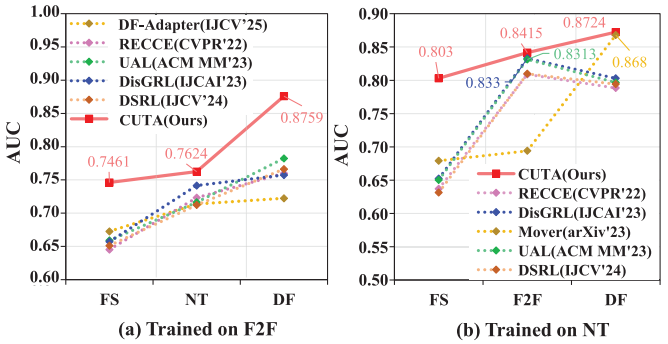


Fig. 2. Result comparisons with the state-of-the-art methods including DF-Adapter [23], RECCE [39], UAL [40], DisGRL [41], Mover [37], and DSRL [42] on the cross-manipulation setting. (a) Trained on Face2Face (F2F). (b) Trained on NeuralTextures (NT). "FS" and "DF" denote FaceSwap and DeepFakes in FaceForensics++ (Low Quality) [15], respectively.

promising results in intra-dataset assessments, they struggle to spot subtle forgery traces amid increasingly realistic forgeries.

To further enhance forgery detection performance, MoE-FFD [24] introduced mixture-of-experts modules that update only the lightweight low-rank adaptation and convpass adapter layers, while keeping the ViT backbone frozen, thus achieving parameter-efficient training. Additionally, FA-ViT [25] utilized a linear adapter to incorporate forgery-related knowledge for deepfake detection. An adapter-based incremental learning scheme [26] was proposed for face forgery detection. This scheme integrates small vanilla trainable adapter modules, which are retrained along with their corresponding classification layers. Although they achieve satisfactory detection accuracy, we identify the limitation of using vanilla adapters, which lack task-specific forgery cues such as locality and texture, making them ineffective in extracting local information. Grounded in these observations, we customize novel adapter modules on the pre-trained ViT backbone, enabling task-agnostic semantics that seamlessly interact with texture details and multi-scale contextual features, thereby achieving more generalized deepfake detection.

To encourage the detection of forgery artifacts that exist in fake faces, contemporary detection techniques rely on semantic visual indicators of deception. These include anomalies in the blending boundaries between real and fake faces [27], [28], consistent and symmetrical skin tones [29], face action dependencies [30], and facial incongruities [31]. Another avenue of research examines the specific forgery pattern, such as noise attributes [32], upsampling artifacts in spectrograms [33], and high-frequency features [34], [35]. Upon meticulously studying the implementation of existing methods, we empirically discerned two enduring issues: 1) Besides the markedly distinct textures and other low-level attributes between real and forged faces, some face swapping techniques also modify universal high-level semantics, such as facial style and shape. Therefore, it is essential to leverage high-level semantics to maintain robustness against low-level feature variations in deepfake detection. 2) Extracting frequency characteristics with hand-crafted filter banks that are content-irrelevant, or focusing solely on specific image regions such as blended boundaries, mouths, and eyes, while disregarding potentially informative areas, renders these methods inherently unable to adapt to the complexities of evolving scenarios. Collectively, these various factors culminate in unsatisfactory accuracy and constrained generalizability.

In this paper, as shown in Figure 1(c), we offer insightful explorations into leveraging ViT adapters as the medium to investigate forgery-specific features in the spatial and frequency domains, and delve into the interplay between them and general semantics. To this end, we propose a novel approach to the ViT adapter named CUTA, which aims to uncover frequency clues and tailor task-oriented adapters while preserving the versatility of the ViT [36]. Unlike Mover [37] and UFID [38], we introduce an additional *frequency domain masking* (FDM) in the input space (*Ref.* Section III-B). This approach complements the RGB image input stream in a supervised fashion by extracting subtle forgery-specific features from high-frequency maps, *e.g.*, noise and boundary details, which are crucial for detecting complex forgery patterns. As shown in Figure 2, considering the cross-manipulation scenario, it's evident from the quantitative results that compared to several state-of-the-art methods, our CUTA enables the model to perform effectively in face reenactment scenarios characterized by subtle forgery patterns, with minimal modifications to general semantics.

Customized features also matter. ViT [36] and its variants have demonstrated remarkable effectiveness across diverse

tasks due to the rich semantic representations learned from pre-trained weights. Based on this background, ViT-Adapter [1] and DF-Adapter [23], [25] have highlighted the effectiveness of high-level semantic features from larger pre-trained ViT models. Instead of the prevailing cross-attention adopted by adapters in existing methods, we advocate for distinct task-specific adapter modules within the ViT block, *i.e.*, a texture enhancement module (*Ref.* Section III-D) and a multi-scale perceptron module (*Ref.* Section III-E) to integrate texture features and multi-scale contextual insights effectively. These two modules facilitate an organic interaction between task-specific forgery patterns and general task-agnostic semantic features. Our method not only enhances intra-dataset detection, but also improves deepfake detection across diverse datasets and manipulation types, demonstrating notable benefits in cross-dataset and cross-manipulation scenarios. To demonstrate the superior performance of the proposed CUTA, extensive experiments are conducted on several publicly available benchmark datasets, including FaceForensics++ (FF++) [15], Celeb-DF [43], WildDeepfake [16], DFDC [44], and DiffSwap [45]. The obtained results validate that CUTA shows a consistent improvement in performance across different baseline models and achieves new state-of-the-art performance in both cross-dataset and cross-manipulation scenarios.

Our major contributions are summarized as follows:

- We propose CUTA, a novel framework for deepfake detection that utilizes ViT adapters to strategically incorporate spatial- and frequency-domain characteristics, achieving superior performance compared to previous methods in terms of generalization ability and robustness.
- We carefully design the FDM in the input space to enhance the challenge of the training process and extract forgery-specific cues. This enables the detector to identify a wider range of manipulation artifacts.
- We specialize in customizing ViT adapters to effectively adapt a pre-trained ViT, enabling general task-agnostic semantic features to organically interact with texture details and multi-scale contextual information, resulting in more comprehensive and adaptable forgery representations.

II. Related Work

A. Deepfake Detection

Deepfake detection aims to predict whether a given image has been artificially manipulated with [4], [5], and [46]. Current methods are chiefly divided into spatial domain-based and frequency domain-based detection techniques. Spatial domain-based methods predominantly focus on capturing low-level visual cues. Early efforts in this domain primarily leverage discrepancies arising from face forgery procedures as visual indicators, including variations in blending boundaries [27], lip movements [31], and eye blinking [47]. Additionally, AUNet [48] comprehensively establishes the relationship between different facial action units to enhance the generalizability of forgery detection. In recent studies, specific forgery patterns, such as various local textures [18], [49], [50], [51], 3D decomposition illumination [52], and patch diffusion [53], have been intensively analyzed to highlight the appearance differences between authentic and forged faces. Furthermore, to uncover more essential forged clues, noise attributes [32], [54], knowledge distillation [55], contrastive pseudo learning [56], [57], intra-frame inconsistency [58], visual semantic relations [59], and reconstruction residual [39] are also utilized [60]. However, these methods tend to capture category-level differences rather than intrinsic discrepancies between authentic and forged images, resulting in unsatisfactory detection accuracy and limited generalization capability.

Regarding frequency domain-based detection techniques, PEL [32] and SPSL [33] explore the local frequency statistics based on sliding window DCT and upsampling artifacts in the phase spectrum to identify forgery indicators, respectively. However, these features are typically content-agnostic and adhere to conventional spatial-frequency fusion principles, resulting in suboptimal interaction between the two domains. SFDG [61] and LVNet [34] address this limitation by employing dynamic graph learning and cross-modal consistency enhancement to extract content-adaptive frequency cues, thereby facilitating the comprehensive fusion of spatial and frequency features. Although current methods generalize well to unseen forgeries, they primarily focus on spatial or frequency artifacts and often overlook general semantic information. This creates an opportunity for improving cross-domain generalization. Our approach aims to bridge this performance gap by merging task-agnostic semantic features from a large pre-trained ViT with detailed forgery patterns from both domains, yielding superior forgery representations with enhanced generalizability.

B. Masked Image Modeling

The Masked Autoencoder (MAE) [62], a prominent masked image modeling technique, has demonstrated remarkable efficacy within the computer vision domain. Leveraging self-supervised pre-training and a pre-trained encoder mechanism, extended versions of MAE quickly emerged to learn robust representations with relatively small modifications [63], [64], [65], thereby achieving significant performance improvements across various downstream tasks [66]. Extensive experiments, as detailed in [67], have underscored the effectiveness of employing masking techniques, showcasing state-of-the-art results in out-of-distribution detection. Furthermore, the application of masked image modeling has also empowered high-capacity models, offering a promising avenue for achieving state-of-the-art generalization performance [68], [69]. Mover [37] was among the first to apply masked autoencoders to deepfake detection, using facial part consistency through masking and recovery processes with reconstruction loss as the primary learning objective. This approach provided valuable insights into detecting unknown forgeries, but it fell short of capturing forgery-specific features. Building on the work of [68], we focus on frequency domain masking in the input space, emphasizing classification losses to distinguish between genuine and forged images. Our method obscures portions of high-frequency image components, intensifying the training challenge while preserving forgery-specific cues.

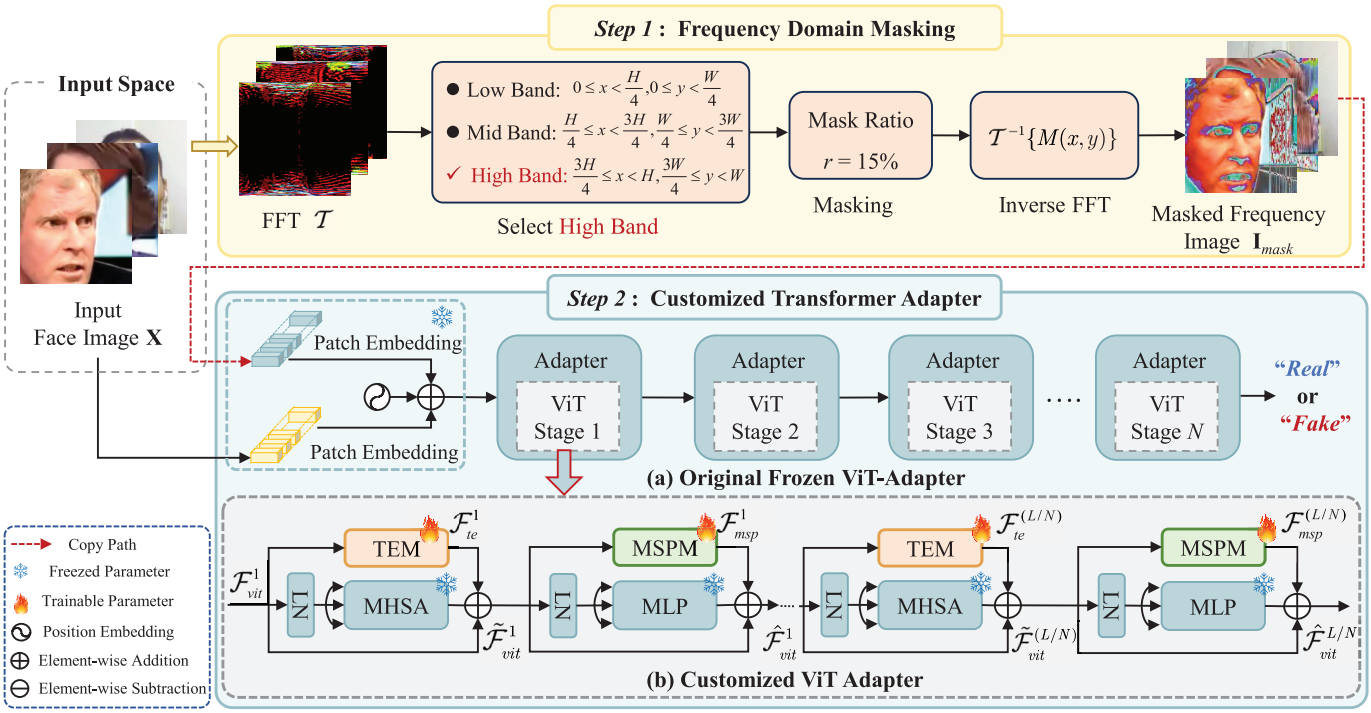


Fig. 3. Overall architecture of our proposed model. *Step 1: Frequency Domain Masking* enhances the challenge of training and uncovering subtle forgery-specific patterns without significant alterations to general semantics. *Step 2: Customized Transformer Adapter* can be divided into two parts. *(a) The original Frozen ViT-Adapter*, featuring its encoder layers partitioned into stages of size N (typically $N = 4$) for feature interaction. *(b) Our Customized ViT Adapter* framework integrates two pivotal components within the ViT block: a texture enhancement module (TEM) and a multi-scale perceptron module (MSPM), which adapt the pre-trained ViT with texture forgery information and multiscale contextual features derived from the single-scale features of ViT. Therefore, our customized ViT adapters enable the creation of comprehensive forgery representations, leading to more generalized deepfake detection.

This approach creates comprehensive forgery representations, thereby facilitating more generalized deepfake detection.

C. Vision Adapter

Adapter modules [70] have garnered increasing attention, as they introduce novel modules within transformer encoders, facilitating task-specific fine-tuning and swift adaptation of pre-trained models to downstream NLP tasks. In the computer vision community, various adapters have been proposed to tailor the plain ViT for tasks such as incremental learning [71] and object detection [72]. However, under standard training settings, their detection performance still lags behind recent models that adeptly integrate image priors. Hence, a robust ViT-Adapter [1] has been devised with task priors and input, serving as a dense prediction task adapter for ViT. Recently, the DF-Adapter [23] pioneered the application of adapter techniques in the field of deepfake detection, which quickly adapts a pre-trained ViT with a dedicated dual-level adapter. While its efficacy enables generalizable deepfake detection, the DF-Adapter presents a negative impact that the current model may struggle to detect forgeries based on face reenactment (*e.g.*, Face2Face and NeuralTextures in the FaceForensics++ dataset [15]). This limitation arises from the fact that these manipulations exhibit only subtle forgery patterns without making significant alterations to general semantics. Therefore, the efficient and comprehensive use of forgery-specific patterns alongside general semantic information remains a key research focus in the field of deepfake detection.

III. Customized Transformer Adapter

A. Overview

An architectural overview of CUTA is illustrated in Figure 3. Generally, CUTA can be divided into two steps. ***Step 1***) Frequency domain masking, which transforms the input face image into the frequency domain, followed by the application of a masking operation. It enhances the training challenges and captures forgery-specific patterns from the high-frequency bands. ***Step 2***) Customized transformer adapter mainly consists of two components. a) The original frozen ViT-Adapter [1], which serves as our baseline and contains a spatial prior module, a spatial feature injector, and a multi-scale feature extractor. While capturing and injecting spatial prior knowledge in the form of convolution and cross-attention enhances detection accuracy for dense prediction tasks, the straightforward adoption of the plain Vit-Adapter for deepfake detection fails to extract task-specific forgery patterns. b) To cultivate the generation of more comprehensive and adaptable forgery representations from spatial and frequency domains, the second part entails the tailor-made ViT-Adapter designed specifically for deepfake detection, which enables the organic interaction between the general task-agnostic semantic features extracted from ViT and the intricate texture forgery details and multi-scale contextual characteristics by the adapters. Details of each part are introduced in the following sections.

B. Frequency Domain Masking (FDM)

Masked image modeling has demonstrated remarkable success in self-supervised representation learning [69]. Notably,

Mover [37] pioneers the adaptation of masked autoencoders [62] for deepfake detection through masking and recovery processes. While Mover provides valuable insights for identifying unknown forgeries, it falls short in capturing forgery-specific features. To this end, inspired by [68], we additionally introduce Frequency Domain Masking (FDM) in the input space to intensify the training challenge while extracting forgery-specific cues, including noise and boundary information, from high-frequency maps. This approach enables enhancing the detector's capability to identify a broader range of artifacts in the frequency domain. Unlike traditional masked image modeling, which primarily relies on reconstruction loss in self-supervised pre-training, our method employs masking in a supervised setting, with a focus on classification loss to distinguish between real and fake images.

As illustrated in Figure 3, our FDM leverages the Fast Fourier Transform (FFT) $\mathcal{T}$ to depict the input face image $\mathbf{X}(u,v) \in \mathbb{R}^{3\times H\times W}$ in terms of its frequency representation $F(x,y)$, which can be expressed as:

$$F(x,y) = \mathcal{T}\{\mathbf{X}(u,v)\}, \tag{1}$$

where x and y denote the frequencies along the width and height of the image, respectively. u and v are spatial coordinates of the input image. The selection of frequency bands is pivotal within the context of FDM, where each band harbors distinct information. These bands are delineated into **Low**, **Mid**, and **High** categories, facilitating the isolation of the impacts of particular frequency components on the holistic image attributes [73], [74]. The demarcation of regions for each frequency band, contingent upon the dimensions $H \times W$ of the FFT of the image, is outlined as follows:

- *Low Frequency Band*:$0 \le x < \frac{H}{4}, 0 \le y < \frac{W}{4}$
- *Mid Frequency Band*:$\frac{H}{4} \le x < \frac{3H}{4}, \frac{W}{4} \le y < \frac{3W}{4}$
- *High Frequency Band*:$\frac{3H}{4} \le x < H, \frac{3W}{4} \le y < W$

Specifically, the *Low Frequency Band* encapsulates the coarse or global features of an image, capturing the fundamental aspects that define its essence. The *Mid Frequency Band* targets medium-frequency components, which frequently embody textures and other intricate details. Finally, the *High Frequency Band* concentrates on high-frequency noise and edge details, which are less dominant but possibly crucial for tasks like deepfake detection. Considering the characteristics of each frequency band and the task at hand, the high-frequency band was selected in this study. Additionally, we empirically demonstrate that frequency masking yields superior performance.

Given the selected frequency band $[x_{start}, x_{end}]\times[y_{start}, y_{end}]$, Q denotes the specific frequencies within $F(x,y)$ to be masked by the mask ratio r and is calculated: $Q = \lceil r \times (x_{end} - x_{start})(y_{end} - y_{start})\rceil$. Frequencies within this region are masked and set to zero, resulting in a masked frequency representation $M(x,y)$:

$$M(x,y) = \begin{cases} 0, & if(x,y) \in Q \\ F(x,y), & otherwise \end{cases}. \tag{2}$$

Specifically, r randomly specifies the portion of pixels in a spatial-wise manner that will be masked. The masked frequency image $\mathbf{I}_{mask}(u,v)$ is derived by performing the inverse FFT $\mathcal{T}^{-1}$ on $M(x,y)$,

$$\mathbf{I}_{mask}(u,v) = \mathcal{T}^{-1}\{M(x,y)\}. \tag{3}$$

Consequently, $\mathbf{I}_{mask}$ and the RGB face image $\mathbf{X}$ are utilized as input for training the ViT-Adapter in the context of universal deepfake detection. It is important to note that masking is exclusively employed during the supervised training phase to help acquire generalizable representations by the detector. No masking is applied during the testing phase.

C. Frozen ViT-Adapter

As illustrated in Figure 3(a), we employ vanilla ViT-Adapter [1] with the frozen ViT backbone as our baseline, referred to as **Frozen ViT-Adapter**. For the original ViT-Adapter, the whole network is composed of N (usually $N = 4$) stages. Each stage contains one stage of pre-trained ViT with frozen parameters during the training and one stage of the adapter with trainable parameters for fast adaptation. In order to obtain a more comprehensive and adaptable forgery representation, in addition to the input RGB image $\mathbf{X} \in \mathbb{R}^{3\times H\times W}$ of the original ViT-Adapter, we additionally introduce the masked frequency image $\mathbf{I}_{mask} \in \mathbb{R}^{3\times H\times W}$ into the patch embedding of frozen ViT-Adapter, where (H, W) is the resolution of the input image. Here, the two images are partitioned into $P \times P$ non-overlapping patches ($P = 16$) and then flattened onto sequential patches $\mathbf{X}_p \in \mathbb{R}^{k\times(P^2C)}$ and $\mathbf{I}_p \in \mathbb{R}^{k\times(P^2C)}$, respectively. Here, C is the number of channels, and $k = HW/P^2$ is the resolution of each image patch. P^2C denotes the resulting number of patches. Subsequently, the two sequential patches are projected into D-dimensional tokens, respectively, reducing the feature resolution to $1/P$ of the original image. Following this, the two kinds of tokens are supplemented with positional embedding, which serve as inputs to the Stage 1, denoted as $\mathcal{F}^i_{vit} \in \mathbb{R}^{\frac{HW}{P^2}\times D}$, traversing through the encoder layer of the Frozen ViT-Adapter.

In particular, we illustrate Stage 1 as an example depicted in Figure 3(b). Given a pre-trained ViT with a total of L blocks (comprising a Multi-Head Self-Attention (MHSA) layer and an MLP layer in each block), we distribute these blocks evenly into N ViT stages, with each stage containing L/N blocks. The corresponding stage of the customized ViT adapter encompasses L/N TEM and MSPM, seamlessly integrated with the ViT block for adaptation. After N iterations of feature interaction, we obtain high-fidelity multi-scale features. Subsequently, we partition and restructure features into three target resolutions: 1/8, 1/16, and 1/32. Finally, we amalgamate the outcome of the ViT block with the 1/32 scale feature map, serving as the input to a standard classification head for generating a final binary predicted probability $\hat{y}$. The cross-entropy loss $\mathcal{L}$ is applied in our proposed CUTA for binary classification learning to supervise predicted probability $\hat{y}$ with binary labels of 0 and 1, *i.e.*, $\mathcal{L} = -\left[y\log\hat{y} + (1-y)\log\hat{y}\right]$, where y is a binary label indicating whether the input image is manipulated.

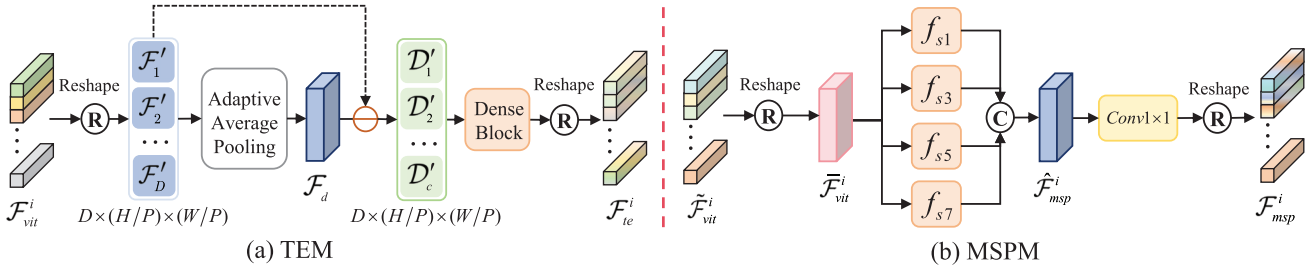


Fig. 4. Illustration of two key components in our customized ViT adapter: (a) Texture Enhancement Module (TEM) and (b) Multi-Scale Perceptron Module (MSPM). These components enable task-agnostic semantic features to effectively interact with texture details and multi-scale contextual information, enhancing the adaptability and comprehensiveness of forgery representations.

D. Texture Enhancement Module (TEM)

Textural information in feature maps, which often contains artifacts introduced by forgery methods, has been proven to be a crucial clue for deepfake detection tasks [18]. The textural information here primarily resides in the high-frequency components of feature maps. Motivated by this observation, we propose a Texture Enhancement Module (TEM), as illustrated in Figure 4(a). Specifically, average pooling is applied to obtain a downsampled feature map that primarily preserves low-frequency components, *i.e.*, large-scale smooth information. By subtracting the pooled map from the input features, low-frequency information is suppressed while high-frequency components are retained, preserving more textural details associated with forgery artifacts. As a result, the introduction of the TEM facilitates the interaction between general task-agnostic semantic features and fine-grained texture details, leading to more comprehensive and discriminative forgery representations.

For the i-th block, a TEM is initially embedded in parallel with MHSA, aiming to inject explicit, discrepancy-specific texture priors, which are usually subtle and occur in local regions. In each TEM, depicted in Figure 4(a), the input features $\mathcal{F}^i_{vit} \in \mathbb{R}^{\frac{HW}{P^2} \times D}$ are first permuted and reshaped to the size $D \times (H/P) \times (W/P)$ as $\mathcal{F}'$. Subsequently, an adaptive average pooling is applied to derive the non-textural feature map $\mathcal{F}_d$. Following this, differentiated feature maps containing texture information are residually extracted to capture the discrepancy information. Building upon [75], 1-D convolutions, sharing parameters, are additionally introduced to delineate the global features of the entire map. Specifically, a dense block comprising 1×1 convolution and 3×3 convolution is devised and applied to the differentiated feature maps $\mathcal{D}' \in \mathbb{R}^{c \times h \times w}$. Ultimately, the output feature maps $\tilde{\mathcal{F}}^i_{vit}$ are obtained through a reshape function to restore the input features size. The above adaptation process of TEM can be expressed in a high-level sketch as follows:

$$\begin{cases} \mathcal{F}_d \leftarrow Avp(Re(\mathcal{F}^i_{vit})) \\ \mathcal{D}' \leftarrow \mathcal{F}' - \mathcal{F}_d \\ DB \leftarrow Conv_{3\times3}(\text{ReLU}(BN(Conv_{1\times1}(\mathcal{D}')))) \\ \mathcal{F}^i_{te} \leftarrow Re(DB(\mathcal{D}')) \end{cases}, \tag{4}$$

where $Re(\cdot)$, $Avp(\cdot)$, and $DB(\cdot)$ represent the reshape function, adaptive pooling operation, and dense block, respectively. $Conv_{3\times3}$ and $Conv_{1\times1}$ denote convolutional operations with kernel sizes of 3×3 and 1×1, respectively. $\mathcal{F}^i_{te}$ represents the adapted low-level texture features from the TEM in i-th block of ViT, which can be further fused with the original output of MHSA layer and the input as follows:

$$\tilde{\mathcal{F}}^i_{vit} = \mathcal{F}^i_{te} + MHSA(LN(\mathcal{F}^i_{vit})) + \mathcal{F}^i_{vit}, \tag{5}$$

where LN and $MHSA$ denote the LayerNorm layer and MHSA layer, respectively.

E. Multi-Scale Perceptron Module (MSPM)

The Multi-Scale Perceptron Module (MSPM) is designed to provide multi-scale invariance and robustness by leveraging hierarchical features from multiple resolutions, which is essential for detecting forgeries of varying sizes. Forged regions in deepfakes often exhibit scale-dependent inconsistencies. To address this issue, we incorporate the MSPM after each MHSA layer and in parallel with the MLP layer within each ViT block. This configuration facilitates the extraction of rich hierarchical forgery cues, as depicted in Figure 3(b). Specifically, drawing inspiration from [23] and [27], MSPM is conceived as an ASPP structure [76] aimed at adapting the pre-trained ViT with additional contextual forgery features, such as blending boundary, and global inconsistencies, improving robustness in deepfake detection. As shown in Figure 4(b), it primarily comprises four separable convolutional layers with varying kernel sizes, capturing both fine-grained details and coarse structures. To dynamically aggregate multiscale features in the adaptation, an additional concatenation operation is introduced after the four separable convolutional layers. This multi-scale feature aggregation allows the model to remain invariant to scale variations and adapt to different manipulation types. Moreover, akin to the TEM, to align with the dimensions of the input features $\tilde{\mathcal{F}}^i_{vit}$, the reshape and convolution functions are integrated between the commencement and culmination of the MSPM for shape transformation. The above adaptation process of MSPM can be expressed in a high-level sketch as:

$$\begin{cases} \bar{\mathcal{F}}^i_{vit} \leftarrow Re(\tilde{\mathcal{F}}^i_{vit}) \\ \hat{\mathcal{F}}^i_{msp} \leftarrow Cat(f_{s1}(\bar{\mathcal{F}}^i_{vit}), f_{s3}(\bar{\mathcal{F}}^i_{vit}), f_{s5}(\bar{\mathcal{F}}^i_{vit}), f_{s7}(\bar{\mathcal{F}}^i_{vit})) \\ \mathcal{F}^i_{msp} \leftarrow Re(Conv_{1\times1}(\hat{\mathcal{F}}^i_{msp})) \end{cases}, \tag{6}$$

where $Cat(\cdot)$ denotes feature concatenation along the channel dimension. f_{si} denotes the separable convolution with a kernel size of $i \times i$. $\mathcal{F}^i_{msp}$ represents adapted multi-scale features from

MSPM in i-th block of ViT, which can be further fused with the original output of MLP layer and the input as follows:

$$\hat{\mathcal{F}}_{vit}^{i} = \mathcal{F}_{msp}^{i} + MLP(LN(\tilde{\mathcal{F}}_{vit}^{i})) + \tilde{\mathcal{F}}_{vit}^{i}, \tag{7}$$

where MLP denotes the MLP layer.

IV. Experiments

A. Experimental Settings

1) Datasets: To facilitate a fair comparison with state-of-the-art methods, we conducted experiments on FaceForensics++ (FF++) [15], Celeb-DF [43], WildDeepfake [16], DFDC [44], and DiffSwap [45]. FaceForensics++ is a widely utilized dataset that offers videos at two different compression tiers (low quality (LQ) and high quality (HQ)), and includes 4,000 manipulated videos generated by four representative techniques: DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). Celeb-DF [43] comprises 590 authentic videos and 5,639 forged videos synthesized by an improved manipulation algorithm. WildDeepfake [16] is a real-world dataset comprising 707 deepfake videos, embodying diverse facial expressions and forgery types encountered in real-life settings, making it particularly challenging. DFDC [44] is a notable large-scale deepfake detection dataset unveiled by Facebook, encompassing over 100,000 video clips. The diversity of manipulations and perturbations in DFDC [44] poses significant challenges for existing detectors. DiffSwap [45] is a newly released dataset from 2024, containing 30,000 high-quality face swaps generated using the diffusion-based DiffSwap method [82]. This dataset facilitates the evaluation of cross-dataset generalization.

2) Implementation Details: Following the official FF++ dataset [15] protocol, we used 720 videos for training, 140 for validation, and the remaining 140 for testing. We utilized RetinaFace [83] to delineate facial regions from video sequences in the FF++ [15], Celeb-DF [43], and DFDC [44] datasets. For the FF++, Celeb-DF and DFDC datasets, we adopted a conservative crop which enlarges the facial region by a factor of 1.3 around the center of the identified face. Then, the input face images are resized to 224×224 and augmented by random cropping. We maintained the facial image size of 224×224 in the WildDeepfake [16] dataset. In contrast, the facial images in the DiffSwap [45] dataset are cropped from 256×256 to 224×224. All experiments are conducted on PyTorch and configured on the platform with NVIDIA GeForce RTX 3090 GPUs. We selected the ViT-Base [36] pretrained on ImageNet-21K and froze all ViT blocks as our backbone. The Adam [84] optimizer is employed with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e{-}8$. The initial learning rate is set to $1e{-}5$ and decayed by 50% every five epochs.

3) Evaluation Metrics: Following previous works [37], [61], we employed the commonly used evaluation metrics, including Accuracy (Acc), Area Under the Receiver Operating Characteristic Curve (AUC), and Equal Error Rate (EER).

B. Quantitative Results

1) Intra-Dataset Evaluation Results: To demonstrate the effectiveness of our proposed CUTA, we compare it with current state-of-the-art methods on the FF++ (HQ and LQ), WildDeepfake, and Celeb-DF datasets. As shown in Table I, our approach consistently achieves admirable performance across all metrics. Specifically, in terms of Acc, our method achieves 98.21% on FF++ (HQ) and 85.62% on WildDeepfake, showcasing a notable improvement over DSRL [42], *i.e.*, a 0.58% performance gain on FF++ (HQ) and 0.99% on WildDeepfake. Different from DSRL which models the commonalities of genuine faces in both spatial and frequency domains, our model employs frequency domain masking to capture forgery-specific features. Furthermore, we boost the information interaction between spatial and frequency domains via ViT adapters, rather than dynamic graph learning in SFDG [61]. Although our CUTA achieves satisfactory detection accuracy on FF++ (LQ), its performance is still inferior to that of FA-ViT [25]. This may be attributed to the fact that FA-ViT requires an additional fine-grained adaptive learning strategy to guide models in capturing detailed information, making it more sensitive to patterns within the dataset. We will explore alternative design strategies to further improve performance in the intra-dataset setting. In contrast to RECCE [39] and DisGRL [41], which amplify authentic compact visual patterns across various granularity spaces via reconstruction learning, our approach attains notably superior AUC values on both HQ and LQ settings of FF++. This comprehensive analysis unequivocally highlights the efficacy of CUTA through superior outcomes.

To further demonstrate the importance of statistical significance in performance evaluation, we have conducted five independent training runs for both our proposed method and the second-best competitor DSRL [42] under the same experimental settings. The results, presented in the last two rows of Table I, include the mean values and corresponding standard deviations. The findings consistently demonstrate that CUTA outperforms DSRL across all evaluated scenarios. Specifically, CUTA achieves higher average accuracy in every test scenario, as indicated by the mean values, while also exhibiting lower standard deviations, indicating greater stability. For instance, with respect to the Acc metric on FF++ (HQ), CUTA achieves $97.65\% \pm 0.80\%$, compared to $96.84\% \pm 0.89\%$ for DSRL. Similarly, in terms of AUC on WildDeepfake, CUTA attains $92.74\% \pm 1.21\%$, significantly surpassing the $91.10\% \pm 1.35\%$ obtained by DSRL. These results collectively highlight the effectiveness, stability, and robustness of our CUTA, positioning it as a highly competitive solution for deepfake detection.

The DFDC [44] dataset is widely regarded as one of the most challenging benchmarks for face forgery detection due to the high visual quality and diversity of manipulated videos it contains. Moreover, since few existing studies report its performance on this dataset, we conducted an intra-testing performance evaluation using Acc, AUC, and LogLoss metrics, and compared our results with re-implemented results introduced in RECCE [39] and the state-of-the-art SFDG [61]. As shown in Table II, our method outperforms SFDG by 1.76% and 1.17% in terms of Acc and AUC, respectively, and achieves a 0.023 reduction in

TABLE I

INTRA-DATASET RESULT COMPARISONS IN TERMS OF ACC (%) AND AUC (%). "HQ" AND "LQ" DENOTE THE HIGH-QUALITY VERSION AND THE LOW-QUALITY VERSION OF THE USED DATASET, RESPECTIVELY. "–" REPRESENTS THAT THERE ARE NO EXPERIMENTAL RESULTS REPORTED FOR THAT DATASET IN THE ORIGINAL PAPER. BEST RESULTS ARE BOLDED, AND SECOND-BEST UNDERLINED. THE UPPER PART PRESENTS THE BEST PERFORMANCE HIGHLIGHTED IN BOLD BLACK. THE LAST TWO ROWS SHOW THE MEAN VALUES AND CORRESPONDING STANDARD DEVIATIONS OBTAINED FROM FIVE INDEPENDENT TRAINING RUNS

Methods	Pub./Year	FF++ (HQ)		FF++ (LQ)		WildDeepfake		Celeb-DF	
		Acc ↑	AUC ↑	Acc ↑	AUC ↑	Acc ↑	AUC ↑	Acc ↑	AUC ↑
Xception [77]	CVPR'17	95.73	96.30	86.86	89.30	77.25	86.76	97.90	99.73
Ef-b4 [78]	ICML'19	96.63	99.18	86.67	88.20	82.33	90.12	98.19	99.83
Add-Net [16]	ACM MM'20	96.78	97.74	87.50	91.01	76.25	86.17	96.93	99.55
F^3-Net [17]	ECCV'20	97.52	98.10	90.43	93.30	80.66	87.53	95.95	98.93
MADD [18]	CVPR'21	97.60	99.29	88.69	90.40	82.62	90.71	97.92	99.94
RFM [79]	CVPR'21	95.69	98.79	87.06	89.93	77.38	83.92	97.96	99.94
PEL [32]	AAAI'22	97.63	99.32	90.52	94.28	84.14	91.62	–	–
RECCE [39]	CVPR'22	97.06	99.32	91.03	95.02	83.25	92.02	98.59	99.94
ITA-SIA [80]	ECCV'22	97.64	99.35	90.23	93.45	83.95	91.34	98.48	99.96
DisGRL [41]	IJCAI'23	97.69	99.48	91.27	95.19	84.53	93.27	98.71	99.91
UAL [40]	ACM MM'23	98.13	99.56	91.35	95.42	84.16	92.56	98.43	99.78
SFDG [61]	CVPR'23	98.19	99.53	92.28	95.98	84.41	92.57	99.22	99.96
FA-ViT [25]	TCSVT'24	97.86	99.60	**92.71**	**96.64**	–	–	–	–
ATSC [81]	TIFS'24	97.90	99.52	91.96	94.54	–	–	–	–
DSRL [42]	IJCV'24	97.63	99.44	92.31	96.12	84.63	92.11	99.24	99.96
CUTA (Ours)	Submission	**98.21**	**99.63**	92.35	96.51	**85.62**	**93.54**	**99.34**	**99.97**
DSRL [42]	IJCV'24	96.84±0.89	99.16±0.33	91.85±0.49	95.43±0.79	83.12±1.72	91.10±1.35	98.62±0.73	99.56±0.41
CUTA (Ours)	Submission	**97.65±0.80**	**99.48±0.31**	**92.15±0.32**	**96.21±0.64**	**84.42±1.53**	**92.74±1.21**	**98.96±0.66**	**99.77±0.22**

TABLE II

EXPERIMENT RESULTS OF INTRA-TESTING ON THE DFDC [44] BENCHMARK DATASET. ACC (%), AUC (%), AND LOGLOSS ARE USED AS THE EVALUATION METRICS. HERE, SMALLER LOGLOSS REPRESENTS A BETTER PERFORMANCE

Methods	Acc ↑	AUC ↑	LogLoss ↓
Xception [77]	79.35	89.50	0.492
Ef-b4 [78]	76.45	89.98	0.524
RFM [79]	80.83	89.75	0.581
Add-Net [16]	78.71	89.85	0.507
F^3-Net [17]	76.17	88.39	0.520
MADD [18]	76.81	90.32	0.529
RECCE [39]	81.20	91.33	0.434
SFDG [61]	86.99	94.44	0.379
CUTA (Ours)	**88.75**	**95.61**	**0.356**

TABLE III

RESULT COMPARISONS ON CROSS-DATASET GENERALIZATION ON CELEB-DF [43], WILDDEEPFAKE [16], AND DFDC [44] WHEN TRAINED ON FF++ (LQ) [15], EVALUATED USING AUC (%) AND EER

Methods	Training	Celeb-DF		WildDeepfake		DFDC	
		AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓
Xception [77]	FF++ (LQ)	61.80	0.417	62.72	0.407	63.61	0.406
Ef-b4 [78]		68.75	0.360	65.30	0.388	67.75	0.369
RFM [79]		65.63	0.385	57.75	0.455	66.01	0.391
Add-Net [16]		57.83	0.444	54.21	0.462	51.60	0.548
F^3-Net [17]		67.95	0.368	60.49	0.434	57.87	0.442
PEL [32]		69.18	0.357	67.39	0.383	63.31	0.404
MADD [18]		68.64	0.371	65.65	0.397	63.02	0.410
RECCE [39]		68.71	0.357	64.31	0.405	69.06	0.361
DisGRL [41]		70.03	0.342	66.73	0.392	70.89	0.343
CADDM [85]		71.01	0.350	65.40	0.395	70.81	0.355
UCF [86]		71.13	0.349	71.05	**0.332**	69.79	0.359
SFDG [61]		75.83	0.303	69.27	0.377	73.64	0.337
DSRL [42]		71.50	0.348	71.46	0.343	72.16	0.345
CUTA (Ours)		**76.45**	**0.291**	**71.51**	0.340	**76.03**	**0.304**

LogLoss. These experimental results demonstrate the satisfactory performance of our method under extreme scene variations.

2) Cross-Dataset Evaluation Results: To explore the generalization ability on unseen datasets, we focus on more challenging cross-dataset experiments by training and testing on different datasets from diverse domains. Following SFDG [61], we refer to its reimplementation of state-of-the-art models for fair and unbiased comparison on FF++ (LQ), testing them on Celeb-DF, WildDeepfake, and DFDC datasets. Table III showcases the comparative results in terms of AUC and EER metrics. It unequivocally demonstrates that our CUTA achieves a certain improvement in generalization ability by taking good advantage of customized task-specific ViT adapters with frequency domain masking. Notably, the AUC score of our CUTA on Celeb-DF (↑ 4.95%) and DFDC (↑ 3.87%) datasets is significantly enhanced when compared with DSRL [42].

To assess the performance of our method against new deepfake detection techniques, we trained the models on the FF++ (HQ) [15] and tested them on Celeb-DF [43], WildDeepfake [16], and DiffSwap [45]. Since the DiffSwap dataset is relatively new, we chose to compare with models that provided open-source code. As shown in Table IV, we observe that our CUTA generally outperforms all listed approaches on unseen test data, often by a clear margin. Moreover, our method demonstrates substantial improvements on the latest

TABLE IV

RESULT COMPARISONS ON CROSS-DATASET GENERALIZATION ON CELEB-DF [43], WILDDEEPFAKE [16], AND DIFFSWAP [45] WHEN TRAINED ON FF++ (HQ) [15]. * REPRESENTS THE RESULTS REPRODUCED USING OPEN-SOURCE CODE OR MODEL

Methods	Training	Celeb-DF		WildDeepfake		DiffSwap	
		AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓
Xception [77]	FF++ (HQ)	66.14	0.393	66.58	0.404	75.22	0.328
Ef-b4* [78]		73.68	0.332	71.08	0.373	78.21	0.295
MADD* [18]		70.54	0.356	70.24	0.365	79.98	0.276
LTW* [86]		77.26	0.294	67.48	0.391	77.98	0.292
RECCE* [39]		70.62	0.351	67.98	0.391	77.85	0.292
DisGRL* [41]		74.14	0.341	69.57	0.373	79.32	0.273
DSRL* [42]		78.46	0.295	74.42	0.324	82.41	0.264
CUTA (Ours)		**79.42**	**0.281**	**78.14**	**0.312**	**84.87**	**0.238**

diffusion-based face swapping dataset, DiffSwap, highlighting its effectiveness against the most recent forgery techniques. For example, when tested on the challenging WildDeepfake and DiffSwap datasets, our CUTA improves the AUC scores of the second-best method by 3.72% and 2.46%, respectively.

These improvements mainly benefit from the customized ViT adapters, which learn the general and comprehensive representation. In addition, FDM enhances the detector's capability to identify a broader range of artifacts against distributional shifts. Overall, our CUTA can cultivate expansive and adaptable representations, and it boasts an exceptional generalization capability for deepfake detection.

3) Cross-Manipulation Evaluation Results: To further demonstrate the generalization across different manipulated types, we conduct fine-grained cross-manipulation experiments on the FF++ (LQ) dataset. Specifically, the model is trained on one manipulated type and evaluated on the remaining three. As shown in Table V, CUTA surpasses existing methods in the majority of scenarios. It is pertinent to acknowledge that DCL [87] demonstrates superior performance compared to our method when training on F2F and testing on DF. Similarly, DCL [87] exhibits better performance than CUTA when training on NT and testing on DF. Nevertheless, it is crucial to underscore that our proposed CUTA retains competitive performance in terms of average performance. Specifically, compared to the suboptimal UAL [40], our CUTA yields gains of 7.58%, 1.02%, and 15.29% in AUC when training on NT and testing on DF, F2F, and FS, respectively. These overwhelming results suggest that our method effectively excavates forgery-specific clues from spatial and frequency domains via customizing ViT adapters with frequency domain masking, thereby improving the cross-manipulation performance.

4) Multi-Source Manipulation Evaluation Results: We also conduct multi-source manipulation evaluation on FF++ (HQ) in terms of Acc and AUC. The protocols and results are taken from LTW [88] and DisGRL [41]. In this section, we provide additional quantitative results on multi-source manipulation evaluation to better demonstrate the practicality of our method. Specifically, we compare our approach with baseline methods on the GID benchmarks using high-quality (HQ) images. As shown in Table VI, we can observe that our CUTA obtains state-of-the-art performance on all protocols in terms of Acc and AUC. For instance, our CUTA surpasses MoE-FFD [24] and DisGRL [41] on the GID-FS by 5.0% and 1.8% in AUC. This illustrates that our proposed CUTA offers reliable generalization across various conditions, further demonstrating

TABLE V

CROSS-MANIPULATION EVALUATION IN TERMS OF AUC (%) ON DIFFERENT FORGERY TYPES OF FF++ [15]

Methods	Training	F2F	FS	NT	Avg.
Xception [77]	DF	66.21	68.67	66.79	67.22
Ef-b4 [78]		65.46	70.56	63.53	66.52
RFM [79]		65.18	72.69	63.44	67.10
Add-Net [16]		68.67	68.61	68.36	68.55
MADD [18]		66.41	67.33	66.01	66.58
DCL [87]		77.13	61.01	75.01	71.05
RECCE [39]		70.66	74.29	67.34	70.76
DisGRL [41]		71.76	75.21	68.74	71.90
DSRL [42]		68.49	72.57	68.01	69.69
UAL [40]		77.92	72.89	70.25	73.69
CUTA (Ours)		**87.01**	**78.82**	**76.53**	**80.79**
Methods	Training	DF	FS	NT	Avg.
Xception [77]	F2F	72.93	64.26	70.48	69.22
Ef-b4 [78]		70.45	61.78	68.66	66.96
RFM [79]		67.80	64.67	64.55	65.67
Add-Net [16]		70.24	59.54	69.74	66.51
MADD [18]		73.04	65.10	71.88	70.01
DCL [87]		**91.91**	59.58	66.67	72.72
RECCE [39]		75.99	64.53	72.32	70.95
DisGRL [41]		75.73	65.71	74.15	71.86
DSRL [42]		76.63	65.11	71.23	70.99
UAL [40]		78.23	65.89	71.64	71.92
CUTA (Ours)		87.59	**74.61**	**76.24**	**79.48**
Methods	Training	DF	F2F	NT	Avg.
Xception [77]	FS	79.54	62.88	56.46	66.29
Ef-b4 [78]		81.70	58.85	53.15	64.57
RFM [79]		81.34	61.53	55.02	65.96
Add-Net [16]		72.82	59.50	53.10	61.81
MADD [18]		82.33	61.65	54.79	66.26
DCL [87]		74.80	69.75	52.60	65.72
RECCE [39]		82.39	64.44	56.70	67.84
DisGRL [41]		82.73	64.85	56.96	68.18
DSRL [42]		83.21	64.53	58.44	68.73
UAL [40]		83.80	64.80	57.59	68.73
CUTA (Ours)		**89.54**	**78.45**	**69.73**	**79.24**
Methods	Training	DF	F2F	FS	Avg.
Xception [77]	NT	74.50	78.23	60.19	70.97
Ef-b4 [78]		75.49	76.91	62.30	71.57
RFM [79]		75.39	72.24	62.83	70.15
Add-Net [16]		77.55	75.42	54.30	69.09
MADD [18]		74.56	80.61	60.90	72.02
DCL [87]		**91.23**	52.13	79.31	74.22
RECCE [39]		78.83	80.89	63.70	74.47
DisGRL [41]		80.29	83.30	65.23	76.27
DSRL [42]		79.42	80.98	63.13	74.51
UAL [40]		79.66	83.13	65.01	75.93
CUTA (Ours)		87.24	**84.15**	**80.30**	**83.90**

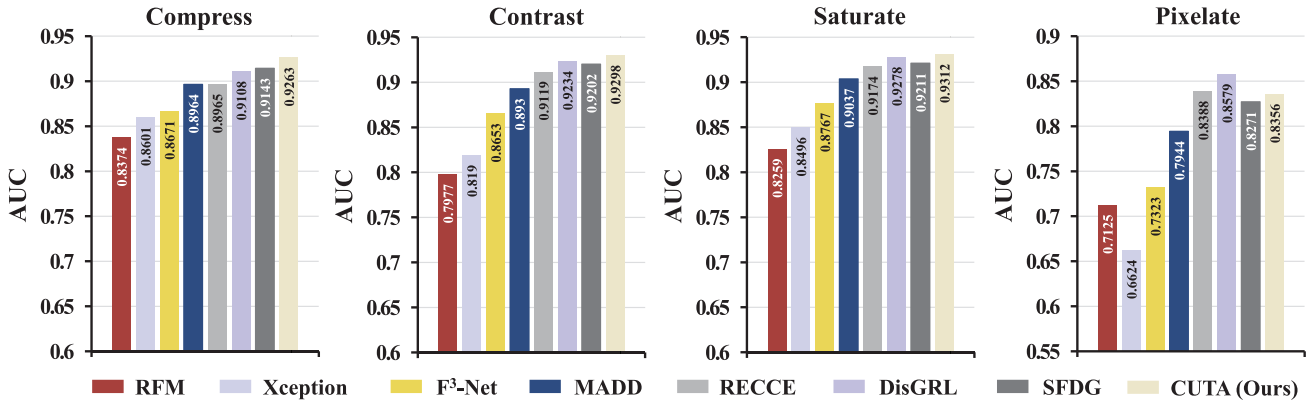


Fig. 5. Robustness evaluation in terms of AUC (%) on the WildDeepfake [16] dataset.

TABLE VI

PERFORMANCE ON MULTI-SOURCE MANIPULATION EVALUATION IN TERMS OF ACC (%) AND AUC (%). GID-DF INDICATES TRAINING ON THE REMAINING THREE FACIAL MANIPULATION TECHNIQUES OF FF++ [15] (HQ) AND TESTING ON DEEPFAKES CLASS. THE SAME FOR THE OTHER SETTINGS

Methods	GID-DF		GID-F2F		GID-FS		GID-NT	
	Acc↑	AUC↑	Acc↑	AUC↑	Acc↑	AUC↑	Acc↑	AUC↑
MLDG [89]	84.2	91.8	63.4	77.1	52.7	60.9	62.1	78.0
LTW [88]	85.6	92.7	65.6	80.2	54.9	64.0	65.3	77.3
MoE-FFD [24]	–	94.7	–	87.7	–	64.7	–	75.9
DisGRL [41]	88.3	95.7	74.7	89.4	58.4	67.9	67.6	84.6
CUTA (Ours)	**89.2**	**96.8**	**76.2**	**91.5**	**61.2**	**69.7**	**69.8**	**86.4**

TABLE VII

COMPARISONS WITH ADAPTATION METHODS UNDER THE CROSS-MANIPULATION EVALUATION IN TERMS OF AUC (%)

Methods	Training	DF	FS
Linear-probing [90]	DF	91.41	67.20
VPT [91]		99.37	77.93
ViT-Adapter [1]		99.19	73.16
DF-Adapter [23]		99.57	76.85
CUTA (Ours)		**99.72**	**78.82**
Linear-probing [90]	FS	75.49	80.41
VPT [91]		86.30	97.25
ViT-Adapter [1]		87.65	98.55
DF-Adapter [23]		88.57	99.04
CUTA (Ours)		**89.54**	**99.48**

the robustness of our framework. Thus, our method is able to handle these variations by utilizing customized transformer adapters with frequency domain masking, thereby improving performance in multi-source manipulation scenarios.

5) Comparisons With Different Adaptation Methods: To accentuate the advantages of CUTA over other prevailing adaptation methods, we compared CUTA with commonly employed fine-tuning methods such as linear probing [90], and three novel parameter-efficient adaptation techniques: VPT [91], ViT-Adapter [1], and DF-Adapter [23]. As illustrated in Table VII, CUTA exhibits significant superiority over the linear probing method. While linear probing fine-tunes only the final layer of the network for efficient parameter adaptation, our proposed task-customized ViT adapter interacts with features from intermediate layers, enabling a broader adaptation. Moreover, CUTA outperforms both VPT and ViT-Adapter, owing to the inclusion of TEM and MSPM. These enhancements enable the adaptation process to effectively integrate intricate local texture details and rich contextual hierarchical features. Notably, compared to the DF-Adapter, which employs a two-layer adapter structure, our proposed method demonstrates a significant 1.97% enhancement in AUC when trained on DF and evaluated on FS. This result highlights the effectiveness of the specifically designed FDM in capturing forgery-specific cues while simultaneously masking high-frequency components of the image, thereby contributing to the observed performance gains.

6) Robustness Analysis: Figure 5 presents the robustness comparisons of different deepfake detection methods [17], [18], [39], [41], [61], [77], [79] under several typical perturbations suggested by SFDG [61] and DisGRL [41], *i.e.*, Compression, Contrast, Saturate, and Pixelate. These results intuitively demonstrate that our CUTA exhibits minimal impact on compression, contrast, and saturate operations, thereby showcasing stronger robustness. However, our CUTA displays greater sensitivity to pixelate operation, potentially due to the utilization of texture enhancement modules and multi-scale perception modules aimed at uncovering subtle manipulation traces. Furthermore, the robustness of CUTA is demonstrated by its consistent superiority over previous methods against various perturbations, outperforming the state-of-the-art DisGRL method by 1.55% in image compression and 0.34% in saturate.

C. Ablation Study

1) Effectiveness of Each Component: As depicted in Table VIII, a series of ablation experiments were conducted on the FF++ (LQ) dataset to validate the effectiveness of each component within our framework. Specifically, the following variants were designed: (a) utilization of a pure ViT-Adapter [1] with a frozen backbone, serving as our baseline model, (b) the baseline model with the proposed TEM, (c) the baseline

TABLE VIII

ABLATION STUDIES ON FF++ [15] IN TERMS OF ACC (%) AND AUC (%)

ID	B	TEM	MSPM	FDM	Acc ↑	AUC ↑
(a)	✓				87.59	89.86
(b)	✓	✓			89.12	92.42
(c)	✓		✓		90.28	93.55
(d)	✓	✓	✓		91.45	95.62
Ours	✓	✓	✓	✓	**92.35**	**96.51**

TABLE IX

THE CROSS-DATASET DETECTION PERFORMANCE (AUC (%)) FOR 15% MASKING OF DIFFERENT FREQUENCY BANDS ON CELEB-DF [43], WILDDEEPFAKE [16], AND DFDC [44]

Masking Frequency Bands	Celeb-DF	WildDeepfake	DFDC
Low Frequency Band	74.01	69.27	74.28
Mid Frequency Band	75.62	70.89	73.54
High Frequency Band	**76.45**	71.51	**76.03**
All Frequency Band	76.12	**71.67**	75.21

model with the proposed MSPM, (d) the proposed method without FDM. Comparing (a) and (b), we observe that TEM brings a considerable improvement in Acc and AUC metrics. By comparing (a) and (c), we demonstrate that adding MSPM leads to a 2.69% increase in Acc and a 3.69% increase in AUC, which can be attributed to the MSPM's ability to extract rich contextual forgery clues. When adding FDM, as shown in (d) and Ours, we observe significant improvements in both Acc and AUC metrics, which can be attributed to the extraction of forgery-specific features (*i.e.*, noise and boundary information) from high-frequency images, allowing for better differentiation between real and forged facial images. Finally, by integrating all components, we achieved optimal performance, with Acc and AUC reaching 92.35% and 96.51%, respectively.

2) Frequency Bands for Masking: In Table IX, we further present the AUC scores of the model when randomly masked across different frequency bands (Low, Mid, High, and All) to verify the effect of these bands on the overall contribution. Notably, the model achieves its highest AUC when high frequency bands are masked, attaining 76.45% on Celeb-DF [43] and 76.03% on DFDC [44]. This underscores the significance of high-frequency noise and edge details, as represented by the high-frequency band, in the deepfake detection task. However, on WildDeepfake [16], masking all frequencies yields the best performance (71.67%), albeit with slightly lower performance when only high frequencies are masked (71.51%). We attribute this disparity to the distinct manipulation methods employed, which may produce varying manipulation traces. Overall, we opt to uniformly employ the high-frequency random masking method in FDM to maximize the effectiveness of the model's deepfake detection performance.

3) Ratio of Frequency Masking: In Table X, we further present the performance of our CUTA on the Celeb-DF [43], WildDeepfake [16], and DFDC [44] datasets across varying masking ratios (0%, 15%, 30%, 50%, and 70%). Notably, at a masking ratio of 15%, the AUC reaches 76.45%, 71.51%, and 76.03% on Celeb-DF, WildDeepfake, and DFDC, respectively. However, as the masking ratio increases, a significant decline in the model's performance is observed. For instance, on Celeb-DF, at the masking ratio of 70%, the AUC decreases by 2.98%. Analysis of this trend reveals the sensitivity of our proposed method to the masking ratio, with excessively high ratios impairing the model's ability to detect subtle manipulation features. Consequently, based on these findings, we opt for 15% as the default masking ratio for frequency masking in our experiments, ensuring optimal balance between performance and masking intensity.

TABLE X

THE CROSS-DATASET DETECTION PERFORMANCE (AUC (%)) OF DIFFERENT MASKING RATIOS (0%, 15%, 30%, 50%, 70%) ON CELEB-DF [43], WILDDEEPFAKE [16], AND DFDC [44] DATASETS

Masking Ration (%)	Celeb-DF	WildDeepfake	DFDC
0%	74.95	70.27	75.78
15%	**76.45**	**71.51**	**76.03**
30%	75.39	71.01	75.84
50%	74.24	70.36	75.07
70%	73.47	70.15	73.21

D. Visualizations

1) Grad-CAM Visualizations: In this section, to gain deeper insights into the internal workings of our method and explore regions of interest for specific manipulation types, we employed visualizations using Grad-CAM [92] on the test set of the FF++ (LQ) dataset. As depicted in Figure 6, traces of various manipulation types are discernible in distinct regions. Notably, for the Face2Face and NeuralTextures manipulation methods, our approach effectively detects subtle alterations in areas such as the mouth, which may prove challenging for human observation. In comparison to ViT and the Baseline, the regions of interest identified by our CUTA exhibit greater accuracy and concentration, enabling precise localization of manipulated areas. As anticipated, the customized adapter tailored for the pretrained ViT adeptly enables interaction between general task-agnostic semantic information within ViT and texture details, alongside multi-scale perceptual features within the adapter. This culminates in the creation of a more comprehensive and adaptable representation for general deepfake detection.

Figure 7 shows the Grad-CAM visualization results on the test set of the WildDeepfake [16] dataset. We observe that our CUTA focuses on prominent features and overarching manipulated traces around facial landmarks, implying that our method can effectively capture the essential discrepancies between authentic and fake samples, even when encountering diverse unseen forgery types encountered in real life. The visualization, from another viewpoint, demonstrates the practicality and generalizability of our proposed method.

2) Distribution Visualizations: To validate the discriminative ability of the proposed CUTA, we utilize t-SNE [93] to visualize the semantic feature distributions of the ViT-Adapter [1] and our CUTA on FF++ (LQ) for two manipulation

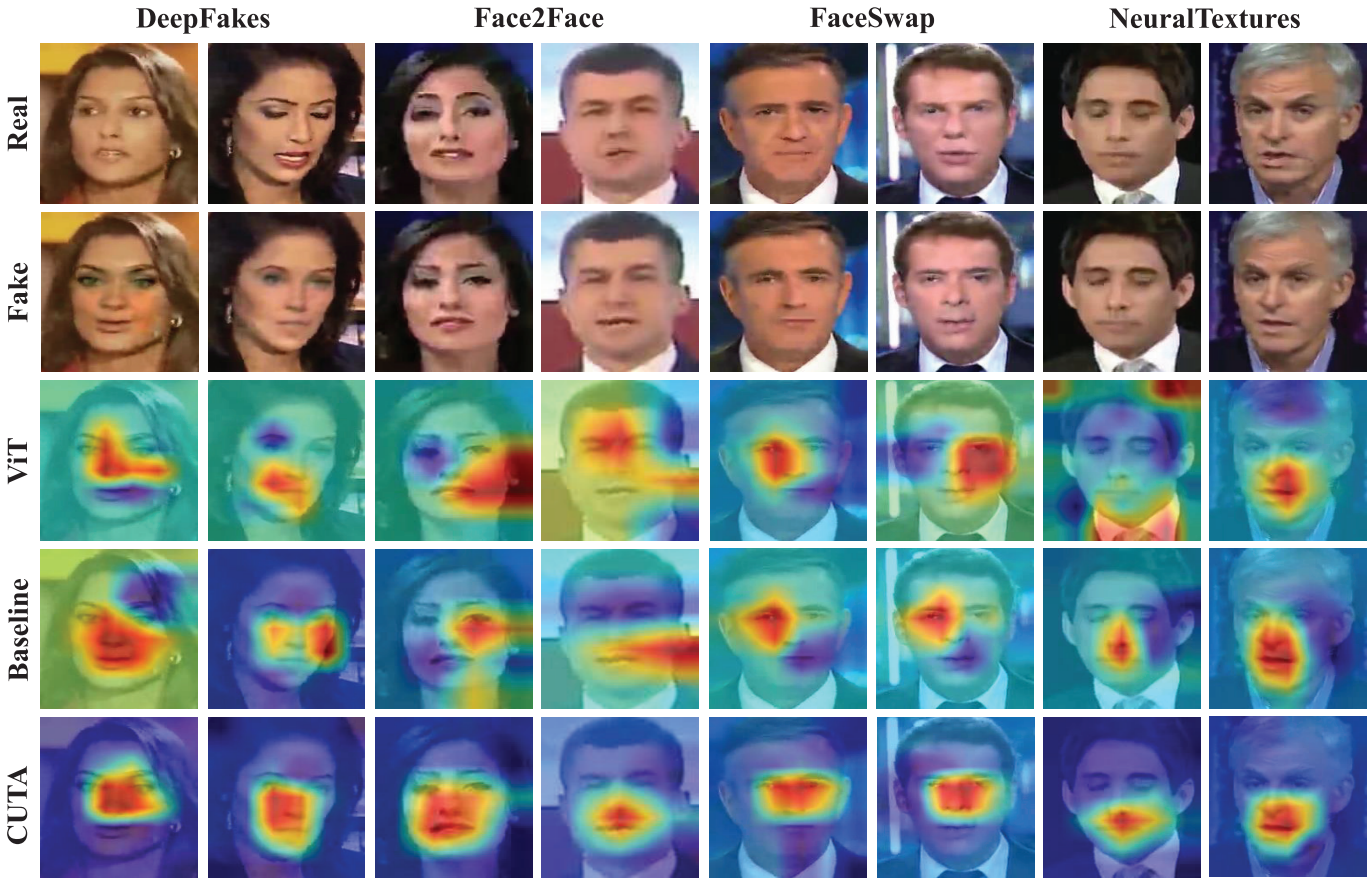


Fig. 6. Grad-CAM [92] visualizations were generated for ViT, Baseline, and our CUTA. These visualizations utilized the final layer features of the models and were applied to the test set of the FF++ [15] (LQ) dataset.

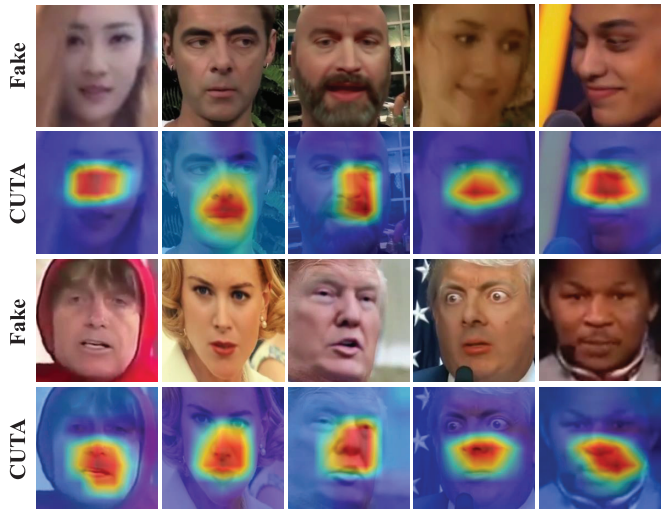


Fig. 7. Grad-CAM [92] visualizations of our CUTA on WildDeepfake [16] dataset. The first row represents the input images with specific manipulated patterns. The second row show the feature maps at final layer features.

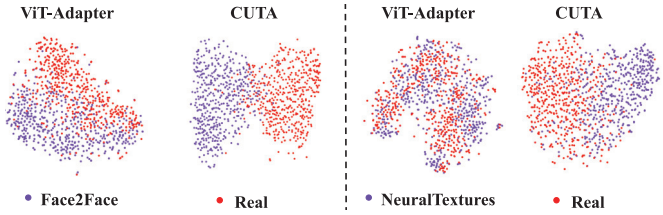


Fig. 8. t-SNE [93] visualization of the ViT-Adapter and our CUTA. The visualization was conducted using the final layer features of a model trained on FF++ [15] Face2Face and NeuralTextures.

methods: Face2Face and NeuralTextures. As depicted in Figure 8, compared to the ViT-Adapter, CUTA clusters samples from the same class into a more compact feature space. Therefore, our CUTA can effectively capture subtle differences between genuine and forged faces through the interaction of general semantics and multi-scale features, thereby exhibiting better detection performance and generalization capability.

E. Computational Complexity Analysis

In this section, we compare our method with RECCE [39], LTW [88], and DisGRL [41], using their publicly available source code to analyze computational complexity and inference time. Results are summarized in Table XI. Experiments were conducted on a cloud service platform using a single NVIDIA GeForce RTX 3090 GPU. The inference time is averaged over 500 randomly selected samples from the FF++ dataset. As shown in Table XI, our CUTA consumes less inference time compared to RECCE and DisGRL, despite higher parameters and FLOPs. *It is believed that inference*

TABLE XI
EXPERIMENTAL RESULTS OF COMPUTATIONAL COMPLEXITY."MS" REPRESENTS MILLISECONDS AND "M" REPRESENTS MILLION

Methods	Inference time	Parameters	FLOPs
RECCE [39]	29.6ms	30.24M	15.23G
LTW [88]	18.1ms	98.97M	63.45G
DisGRL [41]	20.7ms	95.87M	58.09G
CUTA(w/o TEM)	12.58ms	54.02M	52.07G
CUTA(w/o MSPM)	13.71ms	64.23M	57.82G
CUTA	16.06ms	99.78M	61.24G

time matters for deepfake detection and is more reasonable indicator than model parameters and FLOPs. This is because parameters and FLOPs alone cannot accurately measure the complex tricks used in many methods. The design of TEM and MSPM in CUTA indeed increases parameters and FLOPs. In future work, we plan to optimize computational complexity to better balance accuracy and complexity. Moreover, accuracy is also more important than computational complexity, *i.e.*, parameters and FLOPs. In summary, we believe the efficiency of CUTA is acceptable and practical for real-world applications.

V. CONCLUSION

In this paper, we proposed a novel customized transformer adapter (CUTA) coupled with frequency domain masking to exploit features from spatial and frequency domain to fine-tune the large pre-trained ViT for deepfake detection. The FDM was proposed with the input space to enhance training complexity and mine adaptive subtle forgery patterns from content-aware high-frequency information. A powerful task-specific adapter was formulated with TEM and MSPM, which effectively interacted texture information and multi-scale contextual features with general semantics of ViT. Comprehensive experiments and detailed visualizations on widely-used benchmarks demonstrated the effectiveness and generalization of our CUTA compared to state-of-the-art methods. In the future, we will focus on lightweight research for generalized deepfake detection using large visual models. We aim to develop a universal and lightweight deepfake detection model, with the hope of providing technical support for practical applications in real-world scenarios.

REFERENCES

[1] Z. Chen et al., "Vision transformer adapter for dense predictions," in *Proc. 11th Int. Conf. Learn. Represent.*, Jan. 2023, pp. 1–20.

[2] *Video Generation Models as World Simulators*, OpenAI, San Francisco, CA, USA, Feb. 2024.

[3] A. Luo, C. Kong, J. Huang, Y. Hu, X. Kang, and A. C. Kot, "Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 1168–1182, 2024.

[4] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu, and N. Yu, "F2Trans: High-frequency fine-grained transformer for face forgery detection," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1039–1051, 2023.

[5] C. Tan, Y. Zhao, S. Wei, G. Gu, and Y. Wei, "Learning on gradients: Generalized artifacts representation for GAN-generated images detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12105–12114.

[6] W. Guan, W. Wang, J. Dong, and B. Peng, "Improving generalization of deepfake detectors by imposing gradient regularization," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 5345–5356, 2024.

[7] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2387–2395.

[8] Y. Nirkin, Y. Keller, and T. Hassner, "FSGANv2: Improved subject agnostic face swapping and reenactment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 560–575, Jan. 2023.

[9] R. Chen, X. Chen, B. Ni, and Y. Ge, "SimSwap: An efficient framework for high fidelity face swapping," in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, Oct. 2020, pp. 2003–2011.

[10] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 484–492.

[11] C. Peng, Z. Miao, D. Liu, N. Wang, R. Hu, and X. Gao, "Where deepfakes gaze at? Spatial–temporal gaze inconsistency analysis for video face forgery detection," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 4507–4517, 2024.

[12] Y. Guo, C. Zhen, and P. Yan, "Controllable guide-space for generalizable face forgery detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 20761–20770.

[13] L. Verdoliva, "Media forensics and DeepFakes: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 910–932, Aug. 2020.

[14] J. Tian et al., "Learning to discover forgery cues for face forgery detection," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 3814–3828, 2024.

[15] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

[16] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A challenging real-world dataset for deepfake detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2382–2390.

[17] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 86–103.

[18] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2185–2194.

[19] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[20] D. Zhang, J. Tang, and K.-T. Cheng, "Graph reasoning transformer for image parsing," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 2380–2389.

[21] J. Wang et al., "M2TR: Multi-modal multi-scale transformers for deepfake detection," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2022, pp. 615–623.

[22] D. Zhang, F. Lin, Y. Hua, P. Wang, D. Zeng, and S. Ge, "Deepfake video detection with spatiotemporal dropout transformer," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 5833–5841.

[23] R. Shao, T. Wu, L. Nie, and Z. Liu, "DeepFake-adapter: Dual-level adapter for DeepFake detection," in *Proc. Int. J. Comput. Vis.*, Jan. 2023, pp. 1–16.

[24] C. Kong et al., "MoE-FFD: Mixture of experts for generalized and parameter-efficient face forgery detection," 2024, *arXiv:2404.08452*.

[25] A. Luo et al., "Forgery-aware adaptive learning with vision transformer for generalized face forgery detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 5, pp. 1–15, May 2025.

[26] C. Gao, Q. Xu, P. Qiao, K. Xu, X. Qian, and Y. Dou, "Adapter-based incremental learning for face forgery detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 4690–4694.

[27] L. Li et al., "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5000–5009.

[28] D. Zhang, P. Dong, L. Chen, and K.-T. Cheng, "Towards complementary knowledge distillation for efficient dense image prediction," 2024, *arXiv:2401.13174*.

[29] Z. Ba et al., "Exposing the deception: Uncovering more forgery clues for deepfake detection," in *Proc. 38th AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 2, pp. 719–728.

[30] L. Tan, Y. Wang, J. Wang, L. Yang, X. Chen, and Y. Guo, "Deepfake video detection via facial action dependencies estimation," in *Proc.37th AAAI Conf. Artif. Intell.*, vol. 37, Jun. 2023, pp. 5276–5284.

[31] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5039–5049.

[32] Q. Gu, S. Chen, T. Yao, Y. Chen, S. Ding, and R. Yi, "Exploiting fine-grained face forgery clues via progressive enhancement learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 735–743.

[33] H. Liu et al., "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 772–781.

[34] C. Shuai et al., "Locate and verify: A two-stream network for improved deepfake detection," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 7131–7142.

[35] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16317–16326.

[36] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent.*, Jan. 2021, pp. 1–21.

[37] J. Hu et al., "Mover: Mask and recovery based facial part consistency aware method for deepfake video detection," 2023, *arXiv:2303.01740*.

[38] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24480–24489.

[39] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4103–4112.

[40] Y. Wu, X. Song, J. Chen, and Y.-G. Jiang, "Generalizing face forgery detection via uncertainty learning," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 1759–1767.

[41] Z. Shi, H. Chen, L. Chen, and D. Zhang, "Discrepancy-guided reconstruction learning for image forgery detection," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, Aug. 2023, pp. 1387–1395.

[42] J. Cao, K.-Y. Zhang, T. Yao, S. Ding, X. Yang, and C. Ma, "Towards unified defense for face forgery and spoofing attacks via dual space reconstruction learning," *Int. J. Comput. Vis.*, vol. 132, no. 12, pp. 5862–5887, Dec. 2024.

[43] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3204–3213.

[44] B. Dolhansky et al., "The DeepFake detection challenge (DFDC) dataset," 2020, *arXiv:2006.07397*.

[45] Z. Chen et al., "DiffusionFace: Towards a comprehensive dataset for diffusion-based face forgery analysis," 2024, *arXiv:2403.18471*.

[46] Z. Shi, H. Chen, and D. Zhang, "Transformer-auxiliary neural networks for image manipulation localization by operator inductions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4907–4920, Sep. 2023.

[47] Y. Li, M. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.

[48] W. Bai, Y. Liu, Z. Zhang, B. Li, and W. Hu, "AUNet: Learning relations between action units for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24709–24719.

[49] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, and R. Ji, "Local relation learning for face forgery detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 2, pp. 1081–1088.

[50] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, and L. Ma, "Delving into the local: Dynamic inconsistency learning for deepfake video detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 744–752.

[51] Z. Liu, X. Qi, and P. H. S. Torr, "Global texture enhancement for fake face detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8057–8066.

[52] X. Zhu et al., "Face forgery detection by 3D decomposition and composition search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8342–8357, Jul. 2023.

[53] B. Zhang, S. Li, G. Feng, Z. Qian, and X. Zhang, "Patch diffusion: A general module for face manipulation detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 3, pp. 3243–3251.

[54] T. Wang and K. P. Chow, "Noise based deepfake detection via multi-head relative-interaction," in *Proc. 37th AAAI Conf. Artif. Intell.*, vol. 37, Jun. 2023, pp. 14548–14556.

[55] D. Liu et al., "FedForgery: Generalized face forgery detection with residual federated learning," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 4272–4284, 2023.

[56] Z. Sun et al., "Contrastive pseudo learning for open-world DeepFake attribution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 20825–20835.

[57] D. Zhang, H. Zhang, J. Tang, X.-S. Hua, and Q. Sun, "Causal intervention for weakly-supervised semantic segmentation," in *Proc. NIPS*, 2020, pp. 655–666.

[58] W. Zhuang et al., "UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 391–407.

[59] Y. Xu, G. Jia, H. Huang, J. Duan, and R. He, "Visual-semantic transformer for face forgery detection," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Aug. 2021, pp. 1–7.

[60] D. Zhang, C. Zuo, Q. Wu, L. Fu, and X. Xiang, "Unabridged adjacent modulation for clothing parsing," *Pattern Recognit.*, vol. 127, Jul. 2022, Art. no. 108594.

[61] Y. Wang, K. Yu, C. Chen, X. Hu, and S. Peng, "Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7278–7287.

[62] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15979–15988.

[63] C. Feichtenhofer, H. Fan, Y. Li, and K. He, "Masked autoencoders as spatiotemporal learners," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 35946–35958.

[64] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun, "Feature pyramid transformer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 323–339.

[65] P. Gao, T. Ma, H. Li, Z. Lin, J. Dai, and Y. Qiao, "MCMAE: Masked convolution meets masked autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 35632–35644.

[66] H. Wang, Y. Tang, Y. Wang, J. Guo, Z.-H. Deng, and K. Han, "Masked image modeling with local multi-scale reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2122–2131.

[67] J. Li, P. Chen, Z. He, S. Yu, S. Liu, and J. Jia, "Rethinking out-of-distribution (OOD) detection: Masked image modeling is all you need," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11578–11589.

[68] J. Xie, W. Li, X. Zhan, Z. Liu, Y.-S. Ong, and C. C. Loy, "Masked frequency modeling for self-supervised visual pre-training," in *Proc. 11th Int. Conf. Learn. Represent.*, Jan. 2022, pp. 1–22.

[69] J. Huang et al., "Masked generative adversarial networks are data-efficient generation learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 2154–2167.

[70] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2790–2799.

[71] A. Rosenfeld and J. K. Tsotsos, "Incremental learning through deep adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 651–663, Mar. 2020.

[72] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 280–296.

[73] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning," in *Proc. 38th AAAI Conf. Artif. Intell.*, vol. 38, M. J. Wooldridge, J. G. Dy, and S. Natarajan, Eds., Mar. 2024, pp. 5052–5060.

[74] C. T. Doloriel and N.-M. Cheung, "Frequency masking for universal deepfake detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 13466–13470.

[75] M. Guo, Z. Liu, T. Mu, and S. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5436–5447, May 2023.

[76] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[77] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[78] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[79] C. Wang and W. Deng, "Representative forgery mining for fake face detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14923–14932.

[80] K. Sun et al., "An information theoretic approach for attention-driven face forgery detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 111–127.
[81] J. Liu, J. Xie, Y. Wang, and Z.-J. Zha, "Adaptive texture and spectrum clue mining for generalizable face forgery detection," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 1922–1934, 2024.
[82] W. Zhao, Y. Rao, W. Shi, Z. Liu, J. Zhou, and J. Lu, "Diffswap: High-fidelity and controllable face swapping via 3D-aware masked diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 8568–8577.
[83] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-stage dense face localisation in the wild," 2019, *arXiv:1905.00641*.
[84] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, Dec. 2015, pp. 1–15.
[85] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, "Implicit identity leakage: The stumbling block to improving deepfake detection generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3994–4004.
[86] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "UCF: Uncovering common features for generalizable deepfake detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 22355–22366.
[87] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, and R. Ji, "Dual contrastive learning for general face forgery detection," in *Proc. 36th AAAI Conf. Artif. Intell.*, vol. 36, no. 2, Jun. 2022, pp. 2316–2324.
[88] K. Sun et al., "Domain general face forgery detection by learning to weight," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 2638–2646.
[89] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 3490–3497.
[90] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.
[91] M. Jia et al., "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 709–727.
[92] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[93] L. Van der Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, Jan. 2008.

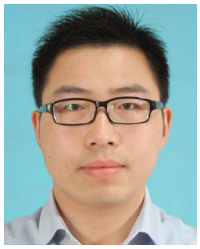
Zenan Shi received the M.E. degree in computing science and the Ph.D. degree in computer application technology with the College of Computer Science and Technology, Jilin University, China, in 2017 and 2021, respectively. Since 2021, she has been a Post-Doctoral Researcher at the Machine Learning and Visual Reasoning Laboratory, Jilin University, working with Prof. Haipeng Chen. Her research interests include multimedia forensics and computer vision, especially on deepfake detection.

Haipeng Chen received the Ph.D. degree in computer application technology with the College of Computer Science and Technology from Jilin University, China, in 2011. He was a Post-Doctoral Research Scientist at the College of Communication Engineering, Jilin University, where he is currently a Professor and the Ph.D. Advisor with the College of Computer Science and Technology. He has wide research interests including medical image segmentation, multimedia forensics, action recognition, human pose estimation, and video understand technology. In these areas, he has published more than 60 technical articles in leading journals and conferences, including IEEE TRANSACTIONS ON IMAGE PROCESSING, CVPR, AAAI, ACM MM, IJCAI, and ICCV.

Yixin Jia received the B.S. degree from Changchun University of Technology, Changchun, China, in 2023. He is currently pursuing the M.S. degree in software engineering with Jilin University, Changchun. His research interests include computer vision and multimedia forensics.

Dong Zhang (Member, IEEE) received the Ph.D. degree from NJUST in 2021. From January 2022 to November 2023, he was a Post-Doctoral Researcher at the Department of CSE. He is a Research Assistant Professor at the Department of Electronic and Computer Engineering, HKUST. He has published more than 30 technical articles in top journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MEDICAL IMAGING, IEEE TRANSACTIONS ON MULTIMEDIA, PR, IJCV, AAAI, ACM MM, IJCAI, CVPR, ECCV, ICCV, ICLR, and NeurIPS. His research focus on machine learning, computer vision, and medical image analysis, with a particular interest in image classification, semantic segmentation, object detection, and their potential applications.

Wei Lu (Member, IEEE) received the B.S. degree in automation from Northeast University, China, in 2002, and the M.S. and Ph.D. degrees in computer science from Shanghai Jiao Tong University, China, in 2005 and 2007, respectively. From 2006 to 2007, he was a Research Assistant with The Hong Kong Polytechnic University. He is currently a Professor with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. His research interests include multimedia forensics and security and AI security.



Xun Yang received the Ph.D. degree from Hefei University of Technology, Hefei, China, in 2017. From 2015 to 2017, he visited the University of Technology Sydney (UTS), Australia, as a joint Ph.D. Student. From 2018 to 2021, he was a Research Fellow with the NExT++ Research Center, National University of Singapore (NUS). He is currently a Professor with the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC). His current research interests include information retrieval, cross-media analysis and reasoning, and computer vision. He regularly serves as a PC Member and an Invited Reviewer for top-tier conferences and prestigious journals in multimedia and artificial intelligence, like ACM Multimedia, IJCAI, AAAI, CVPR, and ICCV. He serves as an Associate Editor for IEEE TRANSACTIONS ON BIG DATA journal and *Multimedia Systems* journal.