F^{2}Trans: High-Frequency Fine-Grained Transformer for Face Forgery Detection

Changtao Miao, *Member, IEEE*, Zichang Tan, *Member, IEEE*, Qi Chu, Huan Liu, Honggang Hu, and Nenghai Yu

***Abstract*— In recent years, face forgery detectors have aroused great interest and achieved impressive performance, but they are still struggling with generalization and robustness. In this work, we explore taking full advantage of the fine-grained forgery traces in both spatial and frequency domains to alleviate this issue. Specifically, we propose a novel High-Frequency Fine-Grained Transformer (F^{2}Trans) network which contains two important components, namely Central Difference Attention (CDA) and High-frequency Wavelet Sampler (HWS). The premier CDA module is capable of capturing invariant fine-grained manipulation patterns by aggregating both pixel-level intensity and gradient information of the query to generate key and value pairs. Subsequently, the proposed HWS discards the low-frequency components of wavelet transformation and hierarchically explores high-frequency forgery cues of feature maps, which prevents model confusion caused by low-frequency components and pays attention to local frequency information. In addition, HWS can be employed as a special pooling layer for the F^{2}Trans architecture to produce hierarchical feature representations in the spatial-frequency domain. Extensive experiments on multiple popular benchmarks demonstrate the generalization and robustness of the specially designed F^{2}Trans framework is well-tailored for face forgery detection when confronting the cross-dataset, cross-manipulation, and unseen perturbations.**

***Index Terms*— Face forgery detection, transformers, wavelet transform.**

I. Introduction

With the rapid development of deepfake techniques [4], [6], [7], various algorithms (*e.g.*, NeuralTextures [3] and Face2Face [5]) have been proposed to generate realistic face images and videos, which are indistinguishable to human eyes or face recognition system [8], [9]. These forged media may be abused for unethical and malicious uses, *e.g.*, spreading political propaganda and creating fake news, which poses huge threats to security. Under this background, face forgery detection came into being and has attracted more and more attention recently.

Most previous methods [10], [11], [12], [13] are constructed based on Convolutional Neural Networks (CNN) and achieve promising performance on intra-dataset evaluation. However, those methods usually suffer from inferior generalization performance on unseen datasets, due to image-specific inductive bias caused by the limited receptive field of CNN [14], [15]. Meanwhile, inspired by various successful applications of transformer architecture [1] in visual tasks, some researchers try to employ pure visual transformers [16], [17] (*e.g.*, ViT [18]) or combine the self-attention mechanism [1] of transformer architecture and CNN [19], [20], [21] for face forgery detection. But those transformer-based methods are not optimal for forgery detection. As clarified in previous work [22], [23], [24], the subtle and fine-grained artifacts are significant in face forgery detection. However, it is deficient for pure visual transformers in capturing these features since its self-attention mechanism [1] emphasizes coarse granularity while ignoring fine-grained feature details. To mitigate this issue, we introduce a Central Difference Attention (CDA) for capturing subtle and fine-grained forgery cues. The proposed CDA is greatly inspired by the central differential ideology [2], [25], [26] that has a strong ability to describe fine-grained and robust invariant manipulation artifacts (*e.g.*, blending boundary, checkboard, blur artifacts, etc.). To be specific, the proposed CDA first employs a vanilla convolution to generate the local texture features as the query of self-attention. Then the central differential operator models the local relationship among the neighbors of the query features to generate key and value pairs of the self-attention. In this way, the candidate keys and values possess detailed invariant forgery patterns, thus enhancing the fine-grained representation capability of the original self-attentive mechanism [1] to capture more informative features. With the proposed CDA, our method prevents the subtle forgery cues from disappearing in general visual transformers networks. Compared with vanilla self-attention [1] and central difference convolution [2], the attention maps of CDA could extract subtle fine-grained manipulated features from a global perspective, as shown in Fig. 1.

Considering that the spatial artifacts may be destroyed when the visual quality of a forged face is degraded, the frequency-domain information have been exploited in previous studies [23], [24], [27], [28], [29], [30], [31], [32], ranging from Fast Fourier transform (FFT) [27], [31] to

Manuscript received 27 July 2022; revised 21 November 2022; accepted 13 December 2022. Date of publication 2 January 2023; date of current version 10 January 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62002336 and Grant U20B2047, in part by the Fund Project of University of Science and Technology of China under Grant YD3480002001, and in part by the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Fernando Alonso-Fernandez. *(Corresponding author: Qi Chu.)*

Changtao Miao, Qi Chu, Honggang Hu, and Nenghai Yu are with the School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230026, China, and also with the Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, Hefei 230026, China (e-mail: miaoct@mail.ustc.edu.cn; qchu@ustc.edu.cn; hghu2005@ustc.edu.cn; ynh@ustc.edu.cn).

Zichang Tan is with the Department of Computer Vision Technology (VIS), Baidu Inc., Beijing 100094, China (e-mail: tanzichang@baidu.com).

Huan Liu is with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China (e-mail: 20112002@bjtu.edu.cn).

Digital Object Identifier 10.1109/TIFS.2022.3233774

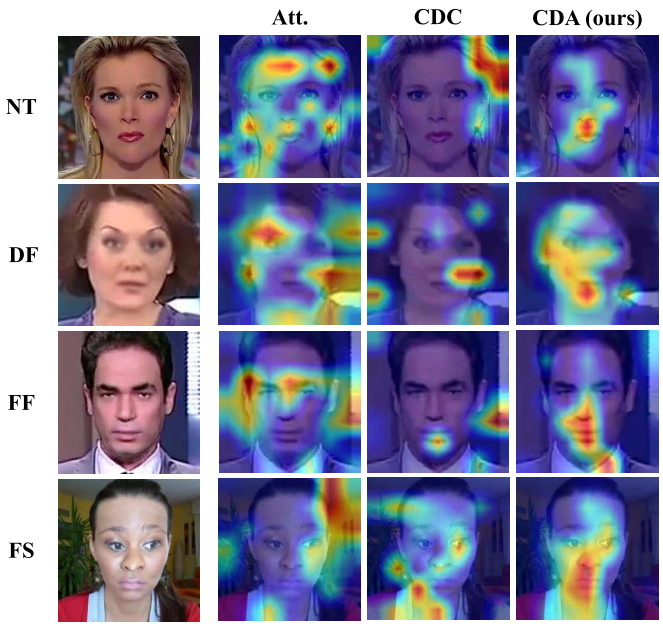


Fig. 1. Attention maps for different kinds of feature extracting operators, *i.e.*, vanilla self-attention (**Att.**) [1], central difference convolution (**CDC**) [2], and ours **CDA**, corresponding to each column. The testing images in each row are from the four forgery methods NT (NeuralTextures [3]), DF (Deepfakes [4]), FF (Face2Face [5]), and FS (FaceSwap [6]), respectively. The **Att.** column is the dispersive attention that learns coarse-grained global information. The **CDC** column is the deviated attention, which is due to the vulnerability of the convolutional network to the image-specific inductive bias. The **CDA** column is the focal attention that can capture fine-grained local forgery features. Figure best viewed in color.

Discrete Cosine Transform (DCT) [23], [28], [32]. Although these studies can deal with the fragile spatial artifacts mentioned above, the frequency-related forgery clues extracted by Fourier Transform (*i.e.*, FFT and DCT) may provide abundant frequency-domain information globally and miss the local spatial relations [33], [34], [35]. Fortunately, in contrast to the Fourier Transform, the Wavelet Transform [36], [37] can identify the local forgery artifacts more clearly and effectively in both spatial and frequency domains. As shown in Fig. 2, the 2D wavelet transformation decomposes an image into four frequency sub-bands, namely LL, LH, HL, and HH, where L and H indicate the low and high pass filters, respectively. For natural and manipulated face images, LL sub-bands present almost the same information while large differences are shown in LH, HL, and HH sub-bands. To achieve this, we design a novel High-frequency Wavelet Sampler (HWS), where the low-frequency components of the LL sub-bands are dropped and high-frequency LH, HL, and HH sub-bands are retained for capturing more general and robust forgery traces in the spatial-frequency domain. Specifically, we aggregate three high-frequency sub-bands of the input RGB features channel-wise together, and then feed it into a Layer Normalization [38] and a linear layer to learn multi-channel high-frequency patterns.

Moreover, the HWS also halves the spatial size of the input feature and doubles their feature dimensions, which can naturally play the role of the down-sampling layer. Thus, We design a High-Frequency Fine-Grained Transformer (**F^{2}Trans**) network with hierarchical representation by leveraging such property of the HWS and combining it with the CDA. Specifically, the proposed CDA and HWS modules are systematically bonded into a hierarchical F^{2}Trans network through four stages. In each stage, the CDA modules of the F^{2}Trans network capture fine-grained forged traces on feature maps of different resolutions, and the coordinated HWS modules extract high-frequency patterns while hierarchically reducing the feature space size. Two complementary modules are stratified to form the entire network, and high-frequency fine-grained forgery clues are sufficiently mined and amplified. Besides, to improve the feature representation capability of the F^{2}Trans under small face forgery datasets, we employ a Locality Skip Connection (LSC) strategic tactic to bolster the spatial-aware local information of the F^{2}Trans based on the ConvNeXt [39] pre-trained on ImageNet [40].

Our F^{2}Trans is a transformer-like network for face forgery detection, and extensive experiments have been conducted to show it outperforms other recent methods and achieves state-of-the-art performance on various challenging benchmarks. The main contributions of this paper can be summarized as follows:

- We propose a unified end-to-end framework named High-Frequency Fine-Grained Transformer (F^{2}Trans) network for face forgery detection.
- We propose a novel Central Difference Attention (CDA), which is suitable for face forgery detection due to its strong ability to capture intrinsic fine-grained features.
- We design a High-frequency Wavelet Sampler (HWS) by discarding the low-frequency while retaining high-frequency components, which helps to extract more general and robust forged traces in the spatial-frequency domain.
- Extensive experiments show the proposed method achieves state-of-the-art performance in both generalization and robustness.

II. Related Work

A. Face Forgery Detection in the Spatial Domain

Most face forgery detection approaches extract subtle artifacts in the spatial domain. Early methods [41], [42], [43], [44], [45], [46] utilize intrinsic statistics or hand-crafted features to model spatial manipulation patterns. Due to the rapid development of deep learning, some work [10], [13] utilize CNN-based models to learn discriminative forgery features from the forged inputs automatically. Besides, some work [22], [47], [48], [49] further introduce spatial attention to learn local forgery traces and achieve remarkable detection performance. Further, some approaches focus on the generalization of the detectors, especially the cross-dataset evaluation. Face X-ray [50] and PCL [51] provide an effective way for detecting the blending boundary with self-supervised datasets. FFD [52] and Local-Relation [32] learn manipulated region mask by a segmentation loss to get a general detector. However, most of them require additional training data which limits their application. Our approach is not only to learn spatial artifacts but also to seek high-frequency robust patterns to mitigate the vulnerability of spatial features.

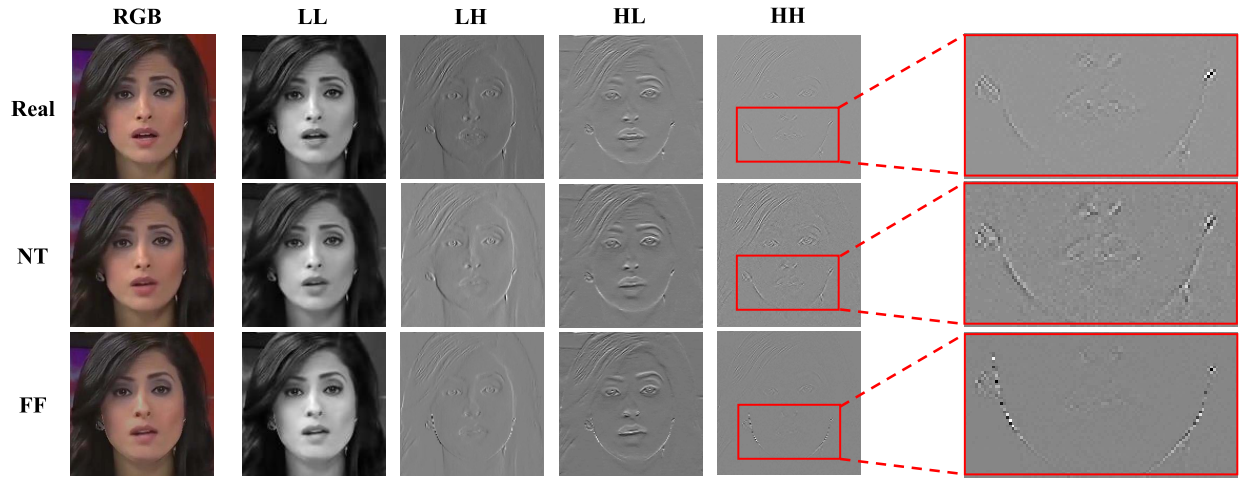


Fig. 2. Illustration of different frequency sub-bands of the wavelet transformation. The information of four frequency domains (*i.e.*, LL, LH, HL, and HH) decomposed from RGB images by implementing the Haar wavelet transformation, but it is employed to decompose the feature maps in our implementation. The real and fake (NT [3] and FF [5]) images are almost identical in low-frequency (*i.e.*, LL sub-band), but vary greatly in high-frequency, especially in HH sub-band. Note that we resize sub-band images in the same resolution as RGB images to compare the discrepancies more clearly. Figure best viewed in color.

B. Face Forgery Detection in the Frequency Domain

Recently, forgery clues in the frequency domain have received great attention [27], [28], [29]. Based on Fourier Transform (*i.e.*, FFT or DCT) some methods, such as Local-Relation [32] and GFF [24], combine the RGB and frequency information by two-stream networks. Different from Fourier Transform, the Wavelets [36], [37] decompose an image into various components and capture frequencies at different spatial resolutions. IAW [34] uses wavelet transform to process RGB images as input for forgery detection. However, most previous methods [24], [30], [34], [35], [53] adopt low-frequency and high-frequency components for analysis, which often causes model confusion. Meanwhile, these methods are coarse-grained for the exploitation of frequency information and cannot effectively capture the subtle forgery traces of different spatial and frequency resolutions comprehensively. Therefore, by discarding the low-frequency components in wavelet transform, the proposed method can better capture general fine-grained forgery artifacts in the frequency domain.

C. Transformers-Based Face Forgery Detection

The CNN architecture is good at learning local features by employing local receptive fields, shared weights, and spatial subsampling [54], [55]. But it experiences difficulty to capture global information due to the limited receptive field [14], [15]. Conversely, the self-attention mechanism of transformer architecture [1] models global relations and long-distance feature dependencies as visual representations. Specifically, the Vision Transformer (ViT) [18] is the first pure transformer architecture that obtains promising results for image classification by exploiting the transformer architecture [1].

For face forgery detection, the previous CNN-based methods [22], [23], [34], [47], [56] also fail to learn the global patterns due to the limited receptive field. Recently, [16] and [17] use pure ViT [18] as a backbone network to establish the global-range relation among the feature patches. M2TR [19] and LiSiam [21] further combine the transformer block and CNN backbone to detect the local forgery artifacts. FTCN [20] utilizes the transformer block to seek spatial-temporal inconsistency information. However, the self-attention mechanism [1] in visual transformers may not encode the local patterns well, leading to over-smoothed feature maps with a deeper network [14], [15]. Moreover, local and fine-grained information is required for the face forgery detection task [22], [23], [24], [31]. HFI-Net [57] and Trans-FCA [58] apply a dual-branch network consisting of ViT and convolution layers to overcome this, while this way is sub-optimal. Therefore, we specially design a single stream high-frequency fine-grained transformer by introducing the central difference operator to transformer architecture [1] to model both local details and global representations for face forgery detection.

III. High-Frequency Fine-Grained Transformer

In this section, we first illustrate the overall architecture of the proposed F^{2}Trans in Sec. III-A. Then, we introduce the Central Difference Attention (CDA) module and Central Difference Attention Transformer (CDAT) block in Sec. III-B. Finally, the High-Frequency Wavelet Sampler (HWS) module is presented in Sec. III-C.

A. Overall Architecture

An overview of the High-frequency Fine-grained Transformer (F^{2}Trans) architecture is presented in Fig. 3. We first embed an input $H \times W$ RGB image into the $\frac{H}{4} \times \frac{W}{4} \times C$ patch embeddings by the Stem layer, which is composed of a 4×4 non-overlapped convolution, a 7×7 depth-wise separable convolution block [59], [60] and a 1×1 convolution. Aiming to produce hierarchical representations, the backbone is divided into 4 stages along with a progressively increasing stride. Each stage contains several Central Difference Attention Transformer (CDAT) blocks to capture the fine-grained features. Between two consecutive stages, there is a High-frequency Wavelet Sampler (HWS) to downsample the feature map to halve the spatial size and double the feature dimensions. Furthermore, to supplement spatial-aware local information in

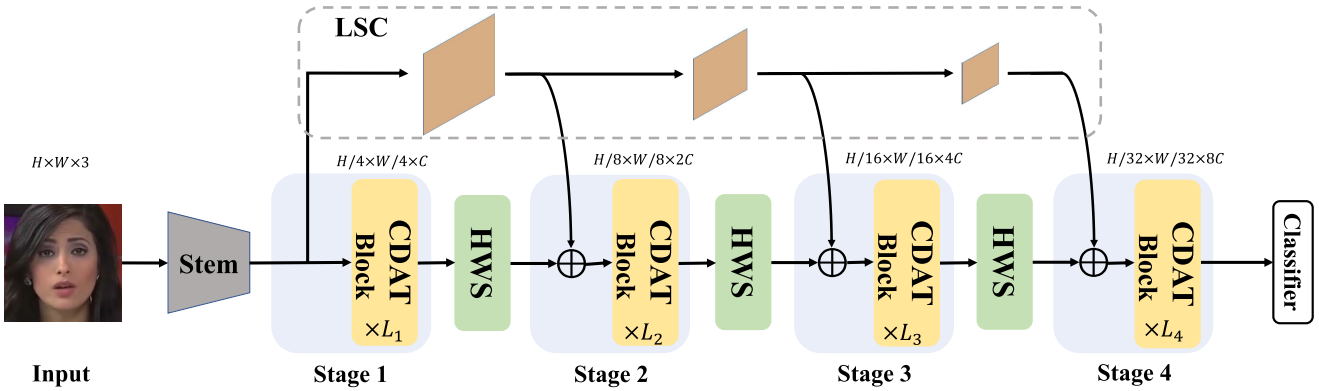


Fig. 3. Overview of the F^{2}Trans. The specially designed transformer network consists of three novel modules: Central Difference Attention Transformer (CDAT) block learns the high-frequency information and local fine-grained artifacts from a global perspective; High-frequency Wavelet Sampler (HWS) extracts multi-channel high-frequency forgery traces in the spatial-frequency domain; Locality Skip Connection (LSC) recruits the local spatial-aware patterns of the Stem to the hierarchical framework. The Stem and LSC consist of spatial depth-wise convolution and linear layers. The $\oplus$ denotes element-wise addition.

the multi-stage network, we leapingly connect the output of the Stem to the hierarchical architecture via LSC. Following ConvNeXt [39], we employ spatial depth-wise convolution and linear layer to construct the LSC complementary structure. Lastly, we employ a linear classifier following the output feature maps of the last stage to produce the predicted results. In the training stage, the cross-entropy loss is employed as the loss function.

B. Central Difference Attention

Local forgery patterns and global-relation representations are critical to the discriminative and generalization capabilities of face forgery detectors. Although several transformer-based studies [19], [20], [57], [58] have achieved remarkable progress by integrating ViT [18] network or part transformer layers with CNN, they did not optimize the basic self-attention mechanism of transformer architecture [1]. Recent improved visual transformer backbone networks [55], [61], [62], [63] try to introduce convolutions into transformer architecture [1], while they focus on generic vision tasks (*e.g.*, classification, detection, and segmentation) and are not specifically designed for face forgery detection task. Inspired by the powerful ability of the central difference operator in fine-grained representation [2], [25], [26], [64], [65], we specifically design a Central Difference Attention (CDA) to extract local fine-grained forgery cues from a global perspective for face forgery detection task. A detailed illustration of the proposed CDA is shown in Fig. 4.

1) Preliminaries of the Vision Transformer (ViT): The ViT [18] is the first pure transformer architecture for image classification tasks. Specifically, ViT [18] splits each input image into a sequence of non-overlapping patches, and then applies multiple standard transformer blocks, each of which consists of a Multi-Head Self-Attention (MHSA) mechanism [1], a Multilayer Perceptron (MLP), and a Layer Normalization (LN) [38] to model these patches.

We first revisit the MHSA mechanism in ViT [18]. Taking a flattened feature map $\boldsymbol{x} \in \mathbb{R}^{N\times D}$ as the input (N indicates the number of tokens and D denotes the dimension of each token), an MHSA block with M heads is formulated as

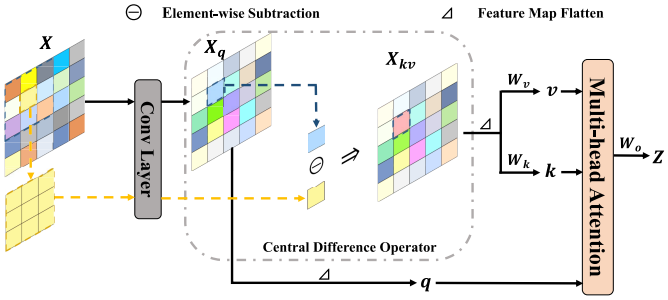


Fig. 4. Illustration of our CDA which learns the invariant fine-grained manipulation patterns from a global perspective. Feature Map Flatten indicates that converts a 2D feature map to a 1D token sequence. The solid line represents the feature flow and the dashed line represents the central difference operator flow.

$$\boldsymbol{q} = \boldsymbol{x}\boldsymbol{W}_q, \quad \boldsymbol{k} = \boldsymbol{x}\boldsymbol{W}_k, \quad \boldsymbol{v} = \boldsymbol{x}\boldsymbol{W}_v, \tag{1}$$

$$z_m = \sigma(\boldsymbol{q}_m \boldsymbol{k}_m^{\top}/\sqrt{d})\boldsymbol{v}_m, \quad m = 1, \ldots, M, \tag{2}$$

$$z = \text{cat}(z_1, \ldots, z_M)\,\boldsymbol{W}_o, \tag{3}$$

where $\sigma(\cdot)$ denotes the softmax function, and $d = D/M$ is the dimension of each head. z_m denotes the embedding output from the m-th attention head, $\boldsymbol{q}_m, \boldsymbol{k}_m, \boldsymbol{v}_m \in \mathbb{R}^{N\times d}$ denote query, key, and value embeddings, respectively. $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, $\boldsymbol{W}_v$, $\boldsymbol{W}_o \in \mathbb{R}^{D\times D}$ are the projection matrices. Then the output of the MHSA [1] is normalized by LN [38] and fed into the MLP (contains two linear layers with a GELU non-linearity) to generate the input to the next transformer block. Meanwhile, the residual connections are employed in both the MHSA and MLP.

2) Central Difference Attention Module: The proposed Central Difference Attention (CDA) effectively models the fine-grained association among tokens through invariant information in the feature maps. Firstly, we exploit the vanilla convolution layer to extract the pixel intensity information to describe local texture information as the queries. These important permanent forgery patterns are determined by the central difference descriptor which combines the pixel intensity information and the pixel gradient information from the local query features. Then these features are fed to the key

and value projections to get the shared fine-grained keys and values. Finally, we apply Eq. (2) and Eq. (3) of the standard MHSA (see Sec. III-B.1) to attend queries to the learned keys and aggregate features from the fine-grained values.

Specifically, as shown in Fig. 4, given the input feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, we first adopt a 2D spatial vanilla convolution to generate the query features $\mathbf{X}_q$, which is formulated as

$$\mathbf{X}_q(p_0) = \sum_{p_n\in\mathcal{R}} \boldsymbol{W}(p_n)\cdot \mathbf{X}(p_0+p_n), \tag{4}$$

where $\boldsymbol{W}(p_n)$ stands for the weight and p_0 denotes the current location on both input and output feature maps, while p_n enumerates the locations in receptive field region $\mathcal{R}$. Then we flatten $\mathbf{X}_q$ to a series of tokens $\boldsymbol{q} \in \mathbb{R}^{(H\times W)\times C}$ with regarding each pixel as a token. Meanwhile, we employ a local central difference operator on $\mathbf{X}_q$ to generate fine-grained invariant features $\mathbf{X}_{kv} \in \mathbb{R}^{H\times W\times C}$ with shared weight $\boldsymbol{W}(p_n)$ of the above 2D spatial vanilla convolution, which can be represented as

$$\mathbf{X}_{kv}(p_0) = \mathbf{X}_q(p_0) - \sum_{p_n\in\mathcal{R}} \boldsymbol{W}(p_n)\cdot \mathbf{X}(p_0). \tag{5}$$

Then the $\mathbf{X}_{kv}$ is flattened and further projected as key tokens $\boldsymbol{k} \in \mathbb{R}^{(H\times W)\times C}$ and values tokens $\boldsymbol{v} \in \mathbb{R}^{(H\times W)\times C}$ by weight shared projection matrices $\boldsymbol{W}_k$ and $\boldsymbol{W}_v$, respectively. This process of producing $\boldsymbol{q}$, $\boldsymbol{k}$, and $\boldsymbol{v}$ above replaces the original position-wise linear projection, *i.e.*, Eq. (1). The Eq. (2) and Eq. (3) of the MHSA are conducted to obtain fine-grained attentive features $\mathbf{Z}$. In this way, our CDA obtains local fine-grained spatial context while maintaining the advantages of the transformer (*i.e.*, dynamic attention, and global relation).

3) Central Difference Attention Transformer (CDAT) Block: To fully exploit the fine-grained forgery features in transformers, we further introduce a CDAT block based on the proposed CDA. More specifically, each CDAT block consists of a CDA module, an MLP layer, and an LN [38] layer. We similarly employ LN [38] layer to normalize the output features of the CDA module, then it is fed into the MLP layer. The residual connections are still adopted in both CDA and MLP. As shown in Fig. 3, four-level CDAT blocks are employed in our framework with each one corresponding to a specific stage. In addition, the input to the CDAT block is complemented with spatially-aware local information through the LSC strategy. The CDAT blocks and LSC strategy sequentially extract fine-grained features to provide general and robust modeling capacity for face forgery detection.

C. High-Frequency Wavelet Sampler

The previous face forgery detection methods [23], [27], [28], [30], [31], [32] extract the frequency-related forgery cues based on the Fourier Transform (*i.e.*, FFT or DCT). Traditional Fourier Transform provides abundant global frequency information, but it lacks the extraction of local fine-grained frequency details. Therefore, we decide to employ Wavelet Transform [36], [37] for face forgery detection, which performs component analysis hierarchically in both spatial and frequency domains. In this way, the spatial decomposition properties of Wavelet Transform [36], [37] can also be performed as a down-sampler for the proposed hierarchical

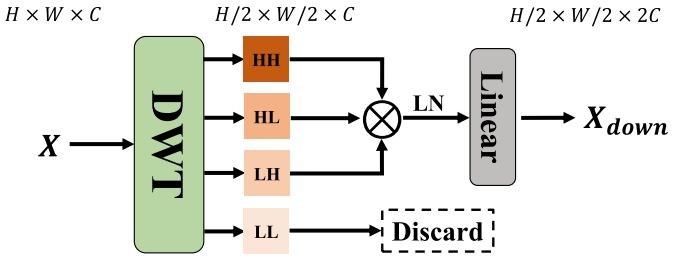


Fig. 5. Proposed HWS extracts multi-channel high-frequency forgery artifacts and halves the size of the input feature maps. DWT indicates a one-level Discrete Wavelet Transform which decomposes the input signals into LL, LH, HL, and HH sub-bands. Linear indicates a single linear layer which reduces the number of channels for high-frequency information. LN denotes Layer Normalization [38] and $\otimes$ denotes channel-wise concatenation.

F^{2}Trans framework. As shown in Fig. 5, we proposed a multi-channel High-Frequency Wavelet Sampler (HWS) layer which is capable of being naturally fit to the above proposed CDA module.

1) Discrete Wavelet Transform (DWT): Wavelets [36], [37] are powerful time-frequency analysis tools, which have wide applications in various visual tasks [33], [66], [67], [68], [69], [70]. The 2D DWT decomposes an image into various components in the frequency domain, essentially capturing different frequencies at different resolutions. The classic 2D DWT contains two kinds of filters, namely **L** and **H**, which indicate the low and high pass filters, respectively. More specifically, the low-pass filter (**L**) concentrates on the smooth surface which is mostly related to low-frequency signals while the high-pass filter (**H**) captures most high-frequency signals like vertical, horizontal, and diagonal edges. These two types of filters can be combined arbitrarily, which forms to four kernels, *i.e.*, **LL**, **LH**, **HL**, and **HH**.

2) Multi-Channel High-Frequency Sample Layer: In our framework, we adopt the 2D Haar Wavelet transformation to obtain **LL**, **LH**, **HL**, and **HH** sub-bands after the one-level of the wavelet decomposition. To be specific, the low (**L**) and high (**H**) pass filters of 2D DWT are

$$\mathbf{L} = \begin{pmatrix} \cdots & \cdots & \cdots & & & \\ \cdots & l_{-1} & l_0 & l_1 & \cdots & \\ & & \cdots & l_{-1} & l_0 & l_1 & \cdots \\ & & & & \cdots & \cdots \end{pmatrix}, \tag{6}$$

$$\mathbf{H} = \begin{pmatrix} \cdots & \cdots & \cdots & & & \\ \cdots & h_{-1} & h_0 & h_1 & \cdots & \\ & & \cdots & h_{-1} & h_0 & h_1 & \cdots \\ & & & & \cdots & \cdots \end{pmatrix}, \tag{7}$$

where the symbols $l_{-1}, l_0, l_1, \ldots$ and $h_{-1}, h_0, h_1, \ldots$ are wavelet basis. Given a feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ where C, H, and W represent the channel number, height, and width of the feature map, respectively, we do DWT for it channel by channel. Specifically, for the features $\mathbf{X}^i$ of the i^{th} channel, the sub-band features are generated by a one-level decomposition as follows:

$$\mathbf{X}^i_{ll} = \mathbf{L}\mathbf{X}^i\mathbf{L}^T, \quad \mathbf{X}^i_{lh} = \mathbf{H}\mathbf{X}^i\mathbf{L}^T, \tag{8}$$

$$\mathbf{X}^i_{hl} = \mathbf{L}\mathbf{X}^i\mathbf{H}^T, \quad \mathbf{X}^i_{hh} = \mathbf{H}\mathbf{X}^i\mathbf{H}^T, \tag{9}$$

where $i \in \{0, 1, \cdots, C-1\}$. Then, these features are stacked among the channel dimension, and we denote them as $\mathbf{X}_{ll}, \mathbf{X}_{lh}, \mathbf{X}_{hl}$, and $\mathbf{X}_{hh}$.

As shown in Fig. 2, by analyzing the wavelet sub-bands of a real and its corresponding manipulated face image, we find that **LL** mainly consists of low-frequency information, depicting the overall appearance of real and fake face images, while **LH**, **HL**, and **HH** contain information representing subtle artifacts and forgery traces (*e.g.*, blending boundary, checkboard, blur artifacts, etc.) of fake face images. As the low-frequency of manipulated face images are, in essence, approximations of the original face, and many kinds of research [24], [30], [53] indicate that the high-frequency (**LH**, **HL**, and **HH**) are helpful for face forgery detection. Therefore, we do not consider the **LL** sub-band for face forgery detection. For the features of **LH**, **HL**, and **HH**, we aggregate them together by a channel-wise concatenation (cat), which can be represented as:

$$\widetilde{\mathbf{X}} = cat(\mathbf{X}_{lh}, \mathbf{X}_{hl}, \mathbf{X}_{hh}), \tag{10}$$

where $\widetilde{X} \in \mathbb{R}^{H/2 \times W/2 \times 3C}$. It not only aggregates high-frequency channel features but also reduces the resolution of input feature maps.. Following [39] and [61], we employ a Layer Normalization (LN) [38] and a Linear layer to align the channel dimension of $\widetilde{X}$ and get the final down-sampling features $\mathbf{X}_{down} \in \mathbb{R}^{H/2 \times W/2 \times 2C}$. After that, the output features of the HWS module are represented as:

$$\mathbf{X}_{down} = Linear(LN(\widetilde{\mathbf{X}})). \tag{11}$$

So far, the HWS module could produce hierarchical feature representations in the spatial domain, while decomposing the feature data in the frequency domain and modeling the local frequency space relativity.

3) Comparison to Vanilla Sampler: There are various down-sampling operations in deep networks, such as Max Pooling, Average Pooling, strided-convolution, etc. Max Pooling and Average Pooling are effective and primitive, but they may ignore beneficial details of the images [33]. Although Mixed Pooling [71], Stochastic Pooling [72], and MaxBlur Pooling [73] are introduced to address these issues, they do not consider the inconsistency between the real face and manipulated face in the frequency domain. Our HWS is significantly different from these vanilla samplers. The proposed HWS employs DWT to not only conduct feature down-sampling but also decompose the image into low-frequency and high-frequency components. Moreover, our HWS further eliminates the low-frequency (**LL**) components but captures forgery traces from high-frequency (**LH**, **HL**, and **HH**) components. Vanilla samplers do not consider the frequency domain information, which may easily result in the aliasing between low-frequency and high-frequency features. It should be noted that our method is underpinned by rigorous wavelet theory [33], [36], [37]. The ablation study is in Table VII to show the superior performance of HWS for face forgery detection.

IV. Experiments

A. Datasets

In this paper, we divide the experiments into four parts: intra-dataset, unseen datasets (cross-dataset), unseen manipulations (cross-manipulation), and real-world perturbations (unseen perturbations). For data preprocessing, we use the MTCNN [74] to detect the face and adopt a conservative crop to enlarge the face region by a factor of 1.3 around the center of the detected face, following [56], [75].

1) Intra-Dataset Evaluation: FaceForensics++ (FF++) [75] is adopted with two challenging quality levels, *i.e.*, High Quality (C23) and Low Quality (C40), following [24], [28], [30], [31], [32]. FF++ consists of 1,000 original videos and 4,000 corresponding fake videos that are generated through four typical manipulation methods, including Deepfakes (DF) [4], Face2Face (FF) [5], FaceSwap (FS) [6], and NeuralTextures (NT) [3]. We follow [75] that randomly selects 720 training videos, 140 validation videos, and 140 testing videos for every 1000 videos. The number of frames in each video is between 300 and 700, so for each video, we take 300 frames for testing evaluation. But we only select 20 frames for training.

2) Unseen Datasets Evaluation: We train the proposed method on FF++ and then test it on three unseen datasets, including Celeb-DF (CDF), DeepFake Detection Challenge (DFDC) [76], and DeepFake-TIMIT-HQ (DFT-H) [77]. Specifically, CDF contains two versions, namely CDF-1 [78] and CDF-2 [79], all of which are taken for generalization evaluation. DFDC adopts the augmentations of the approximate degradation in real-life video distribution. DFT-H includes 320 high-quality fake videos and 320 real videos. We adopt the standard testing set of CDF and DFDC and all fake and real videos of DFT-H as unseen test sets.

3) Unseen Manipulations Evaluation: These experiments are conducted on the FF++ datasets with four forgery methods including DF, FF, FS, and NT. We follow [80], [81] to adopt two protocols, *i.e.*, GID-DF and GID-FF, for evaluation. To be specific, GID-DF and GID-FF represent training on the other three forgery methods but testing on DF and FF, respectively. This is because the FF++ datasets provide four different manipulation methods based on the same source video, this prevents possible bias from different source videos.

4) Real-World Perturbations Evaluation: We apply extensive real-world perturbations to the DeeperForensics-1.0 (DFo) [82], *i.e.*, the standard set without distortions (std) and the standard set with single-level distortions (std/sing), random-level distortions (std/rand), the mixture of three random-level (std/mix3), or the mixture of four random-level (std/mix4). Following work [65], [82], the models are trained on the std set and tested on the perturbation sets.

B. Experimental Setups

1) Implementation Details: During experiments, HWS modules and CDAT blocks are randomly initialized, while the Stem convolution layers and LSC strategic tactic are pre-trained on the ImageNet-1K [40]. The Adam optimizer with a learning rate of $2e-5$ and a weight decay of $1e-5$ is

TABLE I

COMPARISON WITH THE SOTAS ON THE FF++ DATASET. IL INDICATES THE IMAGE-LEVEL AUC AND ACC. IVL INDICATES THE IMAGE-BASED VIDEO-LEVEL AUC AND ACC QUANTITATIVE RESULTS. EXTRA SUP. SHOWS THE EXTRA MASK SUPERVISION. EXTRA LOSS SHOWS THE EXTRA LOSS FUNCTIONS AS SECOND-ORDER SUPERVISION. THE SYMBOL $\star$ INDICATES THE RESULT OF REPRODUCTION

Types	Methods	Ref.	Input Size	Extra Sup.	Extra Loss	C40		C23	
						AUC	ACC	AUC	ACC
IL	Steg. Features [44]	TIFS 12	512	No	No	55.98	–	70.97	–
	Cozzolino et al. [11]	IH&MMSec 17	128	No	No	58.69	–	78.45	–
	Bayar & Stamm [83]	IH&MMSec 16	256	No	No	66.84	–	82.97	–
	Rahmouni et al. [84]	WIFS 17	100	No	No	70.47	–	83.10	–
	MseoNet [10]	WIFS 18	256	No	No	–	70.47	–	83.10
	Xception [75]	ICCV 19	299	No	No	81.76	80.32	94.86	92.39
	SPSL [31]	CVPR 21	299	No	No	82.82	81.57	95.32	91.50
	F^3-Net* [28]	ECCV 20	299	No	No	87.22	84.53	97.80	93.12
	M2TR* [19]	ICMR 22	320	No	No	87.97	85.09	97.84	93.22
	LiSiam [21]	TIFS 22	299	No	No	89.02	86.50	–	–
	F^{2}Trans-S	–	224	No	No	89.60	86.98	99.18	96.09
	F^{2}Trans-B	–	224	No	No	**89.91**	**87.20**	**99.24**	**96.60**
IVL	IAW [34]	TBIOM 21	224	Yes	Yes	92.97	88.96	99.60	96.75
	FRLM [47]	TBIOM 21	299	Yes	Yes	93.69	90.71	99.50	97.63
	LiSiam [21]	TIFS 22	299	Yes	Yes	94.65	91.29	99.52	97.57
	Local-Relation [32]	AAAI 21	299	Yes	Yes	95.21	91.47	99.46	97.59
	M2TR [19]	ICMR 22	320	Yes	Yes	94.22	92.35	99.48	98.23
	MADD [22]	CVPR 21	299	No	Yes	87.26	86.95	98.97	96.37
	Two-branch [29]	ECCV 20	256	No	Yes	91.10	86.34	99.10	96.43
	FDFL [30]	CVPR 21	299	No	Yes	92.40	89.00	99.30	96.69
	Xception [75]	ICCV 19	299	No	No	89.30	86.86	94.80	93.86
	Add-Net [85]	MM 20	224	No	No	91.01	87.50	97.74	96.78
	F^3-Net [28]	ECCV 20	299	No	No	93.30	90.43	98.10	97.52
	M2TR* [19]	ICMR 22	320	No	No	92.46	88.71	99.19	96.14
	DCL [81]	AAAI 22	299	No	No	–	–	99.30	–
	MTD-Net [65]	TIFS 21	224	No	No	–	–	99.38	–
	LiSiam [21]	TIFS 22	299	No	No	93.99	88.86	–	–
	F^{2}Trans-S	–	224	No	No	**94.11**	90.14	99.73	98.14
	F^{2}Trans-B	–	224	No	No	94.04	**90.57**	**99.74**	**98.71**

adopted. The input size is 224 × 224 and the batch size is set to 12. Two F^{2}Trans variants are designed, including F^{2}Trans-S with $C = 128$ and layer numbers $= \{1, 1, 1, 1\}$, and F^{2}Trans-B with $C = 128$ and layer numbers $= \{1, 1, 3, 1\}$, where C is the channel number of the hidden layers in the Stem and Stage 1, and the layer numbers are the number of repetitions (L_1, L_2, L_3, and L_4 in Fig. 3) of the corresponding Stage. We implement our framework and experiments with open-source PyTorch and a single NVIDIA Tesla V100 GPU.

2) Evaluation Metrics: The Accuracy (ACC) and Area Under the Receiver Operating Characteristic Curve (AUC) are reported for comparison metrics. Since our F^{2}Trans is essentially an image-based model, and thus we use the image-level (IL) way to generate the predicted score via a single frame. But some approaches [28], [56] presented the video-level performance. For a fair comparison, we apply the image-based video-level (IVL) way to compute the video-level prediction by averaging the score of a sequence of frames.

C. Comparison With Recent Work

In this sub-section, we construct generalization and robustness experiments, including intra-dataset evaluation, cross-dataset evaluation, cross-manipulation evaluation, and real-world perturbations evaluation. For a fair comparison with state-of-the-art (SOTA) methods, we select the image-level (IL) and image-based video-level (IVL) evaluation metrics in the following experiments. Note that the results of comparisons are obtained from their papers, and we specify our implementation by $\star$ otherwise.

1) Intra-Dataset Evaluation: The image-level (IL) experimental results are shown in Table I row IL. The image-level evaluation is more challenging, but our method outperforms all compared methods on the FF++ (C23 and C40) datasets. For instance, our F^{2}Trans-B is about **7.09%** and **2.69%** AUC higher than the classical frequency-related SPSL [31] and F^3-Net [28] on the highly compressed C40 dataset, respectively. M2TR [19] and LiSiam [21] with larger input sizes, our method still can achieve better performance than theirs. These excellent behaviors are derived from the proposed F^{2}Trans that excludes compression interference of the C40 dataset and learns invariant fine-grained forgery features.

Meanwhile, many kinds of research reported video-level performance metrics, although their models are image-based. So we adopt the image-based video-level (IVL) way to compute the video-level results in the testing phase, as shown in Table I row IVL. Some studies (*e.g.*, IAW [34], FRLM [47], M2TR [19], LiSiam [21], and Local-Relation [32]) achieve remarkable accuracy through applying the specific

TABLE II
COMPARISON WITH THE SOTAS ON THE UNSEEN DATASETS. THE SYMBOL ⋆ INDICATES THE RESULT OF REPRODUCTION

Image-level (IL) AUC						Image-based Video-level (IVL) AUC					
Methods	Training Set	Unseen Testing Set				Methods	Training Set	Unseen Testing Set			
		CDF-2	CDF-1	DFDC	DFT-H			CDF-2	CDF-1	DFDC	DFT-H
SMIL [56]	FF++ (C40)	56.30	–	–	–	Two-branch [29]	FF++ (C40)	76.70	–	–	–
Add-Net [85]		57.83	–	51.60	–	M2TR⋆ [19]		72.05	74.49	66.02	76.77
Xception [75]		65.50	–	69.70	70.50	F^3-Net⋆ [28]		77.92	78.74	67.35	81.05
M2TR⋆ [19]		66.15	67.42	63.74	72.29	**F^2Trans-S**		80.98	**91.68**	**76.26**	70.95
PEL [23]		69.18	–	63.31	–	**F^2Trans-B**		**83.96**	91.09	74.20	**88.06**
FRLM [47]		70.58	76.52	69.81	65.48	PatchForensics [86]	FF++ (C23)	69.60	–	65.60	–
F^3-Net⋆ [28]		71.29	72.29	65.39	76.59	CNN-GRU [87]		69.80	–	68.90	–
Two-branch [29]		73.41	–	–	–	F^3-Net⋆ [28]		75.68	66.04	70.88	97.54
SPSL [31]		76.88	–	–	–	M2TR⋆ [19]		75.76	70.88	73.12	96.73
F^2Trans-S		75.61	86.29	**71.78**	67.69	Xception [75]		73.70	–	70.90	–
F^2Trans-B		**77.61**	**86.46**	70.39	**83.78**	GFF [24]		75.31	–	71.58	–
Capsule [88]	FF++ (C23)	57.50	–	–	74.40	STIL [89]		75.58	–	–	–
Xception [75]		65.30	–	72.20	94.40	MADD [22]		76.65	–	67.34	–
MADD [22]		67.44	–	–	–	LTW [80]		77.14	–	74.58	–
F^3-Net⋆ [28]		68.69	63.57	67.45	93.50	Local-Relation [32]		78.26	–	76.53	–
M2TR⋆ [19]		69.94	68.57	69.94	93.75	Face X-ray [50]		79.50	–	65.50	–
MTD-Net [65]		70.12	–	–	–	DCL [81]		82.30	–	76.71	–
IAW [34]		–	72.30	–	–	LipForensics [90]		82.40	–	73.50	–
LiSiam [21]		78.21	81.14	–	–	FTCN [20]		86.90	–	74.00	–
F^2Trans-S		80.72	81.77	**73.69**	96.44	**F^2Trans-S**		87.46	87.95	**77.69**	97.37
F^2Trans-B		**83.05**	**83.92**	71.81	**97.04**	**F^2Trans-B**		**89.87**	**88.03**	76.15	**98.38**

manipulated mask supervision to guide the learning of capturing discriminative features. But the mask ground truth they rely on is difficult to access in real-world situations. MADD [22] and FDFL [30], and Two-branch [29] employ additional loss functions as second-order supervision, all of which obtain good results. The difference is that our approach focuses on designing face forgery detection networks suitable for wild scenarios and only employs the Cross-Entropy Loss function without additional supervision signals. We fairly compare against others methods and achieve better results than Xception [75], Add-Net [85], F^3-Net [28], DCL [81], MTD-Net [65], and LiSiam [21]. In addition, we also reproduce the video-level results of M2TR [19] without additional supervision signals, and our F^2Trans outperforms it.

2) Generalization to Unseen Datasets: In the real world, many manipulated face data are completely unknown, *i.e.*, they are created by anonymous forgery methodology based on unknown source data. To evaluate the generalization capacity of the face forgery detector in this scenario, we constructed a cross-dataset testing protocol. Concretely, the model is trained on all the four types of fake data in FF++ (C40 or C23) train sets while tested on four unseen datasets, including CDF-1 [78], CDF-2 [79], DFDC [76], and DFT-H [77]. As shown in Table II, compared with the intra-dataset evaluation, most previous methods like Local-Relation [32], MTD-Net [65], DCL [81], LiSiam [21], and M2TR [19] show a drastic performance drop on unseen datasets. From Table II column IL, our F^2Trans significantly outperforms all the competitors on all unseen testing datasets with image-level evaluation metric. Moreover, FTCN [20] and LipForensics [90] employ temporal coherence information to detect fake face videos, so we adopt the image-based video-level evaluation metric for fair comparison in Table II column IVL. The results show that the proposed F^2Trans-B also exceeds the recent SOTA FTCN [20] by **2.97%** on the CDF-2 dataset. On the more challenging DFDC dataset, our F^2Trans-S outperforms DCL [81] by **0.98%**. Overall, our method achieves the best performance on every unseen dataset, which mainly benefits from the fine-grained high-frequency forgery traces explored by the proposed CDA and HWS modules.

3) Generalization to Unseen Manipulations: The generalization performance of forgery detection models is equally affected by the emergence of new manipulation methods. To further demonstrate the generalization ability of detection models when facing unseen manipulation types, we conduct cross-manipulation experiments on FF++ (C23 and C40) following [80], [81]. In Table III, we can observe that the proposed F^2Trans outperforms other forgery detectors by a large margin. Although GID-DF and GID-FF have different compression rates and different manipulation methods, our method can as well learn fine-grained high-frequency features and generalize to detect unseen forgery data. For example, the proposed F^2Trans-B improves by **2.66%** in terms of video-level AUC compared with the current SOTA DCL, on highly compressed GID-FF (C40).

4) Robustness to Real-World Perturbations: During the transmission of manipulated face data in real-world situations, many noises are introduced into the data, such as blurring, compression, etc. Accordingly, it is essential for face forgery detectors to be robust to unseen kinds of perturbations. To further verify the robustness of F^2Trans network, a series of experiments are designed on distorted data in DFo [82]. In this setting, all models are trained on the standard set without distortions and tested on different level perturbation sets (std/sing, std/rand, std/mix3, and std/mix4), as shown in Table IV. Our F^2Trans-B achieves SOTA performance and is about **5.65%** higher than MTD-Net [65] on the std/mix3

TABLE III
IMAGE-BASED VIDEO-LEVEL AUC AND ACC PERFORMANCE ON CROSS-MANIPULATION DATASETS. THE SYMBOL ⋆ INDICATES THE RESULT OF REPRODUCTION

Methods	GID-DF (C23)		GID-DF (C40)		GID-FF (C23)		GID-FF (C40)	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
EfficientNet [91]	91.11	82.40	75.30	67.60	80.1	63.32	67.40	61.41
ForensicTransfer [92]	–	72.01	–	68.20	–	64.50	–	55.00
Multi-task [12]	–	70.30	–	66.76	–	58.74	–	56.50
MLDG [93]	91.82	84.21	73.12	67.15	77.10	63.46	61.70	58.12
LTW [80]	92.70	85.60	75.60	69.15	80.20	65.60	72.40	65.70
DCL [81]	94.90	87.70	83.82	75.90	82.93	68.40	75.07	67.85
M2TR⋆ [19]	94.91	81.07	84.85	74.29	76.99	55.71	71.70	66.43
F^3-Net⋆ [28]	94.95	83.57	85.77	77.50	81.20	61.07	73.70	64.64
F^{2}Trans-S	97.47	89.64	86.92	77.86	90.55	81.43	76.52	66.79
F^{2}Trans-B	**98.92**	**92.86**	**88.77**	**82.14**	**94.08**	**86.07**	**77.73**	**70.36**

TABLE IV
IMAGE-BASED VIDEO-LEVEL ACC RESULTS ON REAL-WORLD PERTURBATIONS DATASETS. THE SYMBOL ⋆ INDICATES THE RESULT OF REPRODUCTION

Methods	std (training set)			
	std/rand	std/sing	std/mix3	std/mix4
C3D [95]	92.38	87.63	–	–
TSN [96]	95.00	91.50	–	–
I3D [97]	96.88	90.75	–	–
ResNet+LSTM [82]	97.13	90.63	–	–
F^3-Net⋆ [28]	92.79	86.82	79.10	66.42
Xception [75]	94.75	88.38	82.32	–
M2TR⋆ [19]	94.78	91.79	86.32	78.36
MTD-Net [65]	95.22	–	86.89	–
F^{2}Trans-S	97.26	95.77	93.28	90.55
F^{2}Trans-B	**98.76**	**97.76**	**95.27**	**92.54**

TABLE V
ABLATION EXPERIMENT ON THE DIFFERENT MODULES OF THE PROPOSED F^{2}TRANS. THE 'W/O' DENOTES 'WITHOUT'

Methods	FF++(C40)		CDF-1	std/mix4
	AUC	ACC	AUC	ACC
w/o HWS	88.85	86.77	77.99	83.49
w/o CDA	88.86	86.15	77.58	82.89
w/o LSC	88.55	86.01	80.42	84.06
F^{2}Trans-S	**89.60**	**86.98**	**86.29**	**87.75**

testing set, which reflects that our model is more robust to unseen perturbations. It is worth mentioning that our novel F^{2}Trans framework is adept at robustness, which benefits from CDA and HWS modules learning the high-frequency detailed forgery cues.

5) Attention Visualizations: The visualizations for DF [4], FF [5], FS [6], NT [3], and CDF [78], [79] are shown in Fig. 6. Detecting the faces manipulated by FF and NT is especially challenging due to these methods are local tampering technologies (*i.e.*, facial-reenactment), which only manipulate the expression or the lips movements of a face. However, our method focuses well on microcosmic regions, especially in subtle manipulation parts (*e.g.*, nose and mouth) of the forged face. Besides, our model also learns the general artifacts for the unseen CDF datasets as shown in the CDF column of Fig. 6.

D. Ablation Study

In this sub-section, we perform comprehensive experiments to analyze the influence of different components of our F^{2}Trans on the FF++ (C40) [75], CDF-1 [78], and the std/mix4 of the DFo [82]. Since generalization and robustness are the primary challenges of the face forgery detector, our ablation experiments concentrate on analyzing the performance of different models on CDF-1 and std/mix4 datasets. Note that we train F^{2}Trans-S on the train set of FF++ (C40) and evaluate it on intra-dataset (test set of FF++ (C40)) and cross-dataset (unseen CDF-1). For the std/mix4 of the DFo, our models train on the std set of the DFo and then test on std/mix4 of the DFo. All ablation experiments are evaluated with image-level (IL) AUC and ACC evaluation metrics.

1) Analysis on Proposed Modules: As shown in Table V, we conduct experiments to analyze the different modules of the proposed F^{2}Trans. 'w/o HWS' indicates that the vanilla Max Pooling is employed rather than the HWS module, 'w/o CDA' indicates the F^{2}Trans network without the CDA transformer block, and 'w/o LSC' means the locality skip connection is dropped. When any one of the above three modules is removed, the performance degrades, especially on generalization and robustness evaluation. For example, when HWS is dropped, the AUC reduces from 86.29% to 77.99% on the CDF-1 dataset, which shows that the wavelet-based sampler is suitable for face forgery detection because it extracts high-frequency artifacts in the spatial-frequency domain. Overall, our specially designed F^{2}Trans architecture is an organic whole, and each module plays important in learning high-frequency fine-grained features.

2) Analysis on Different Frequency Components: In order to seek suitable frequency sub-bands of 2D DWT for face forgery detection task, we employ various combinations of frequency components for comparison, as shown in Table VI. We can observe that the high-frequency (LH+HL+HH) outperforms other combinations of frequency components in both

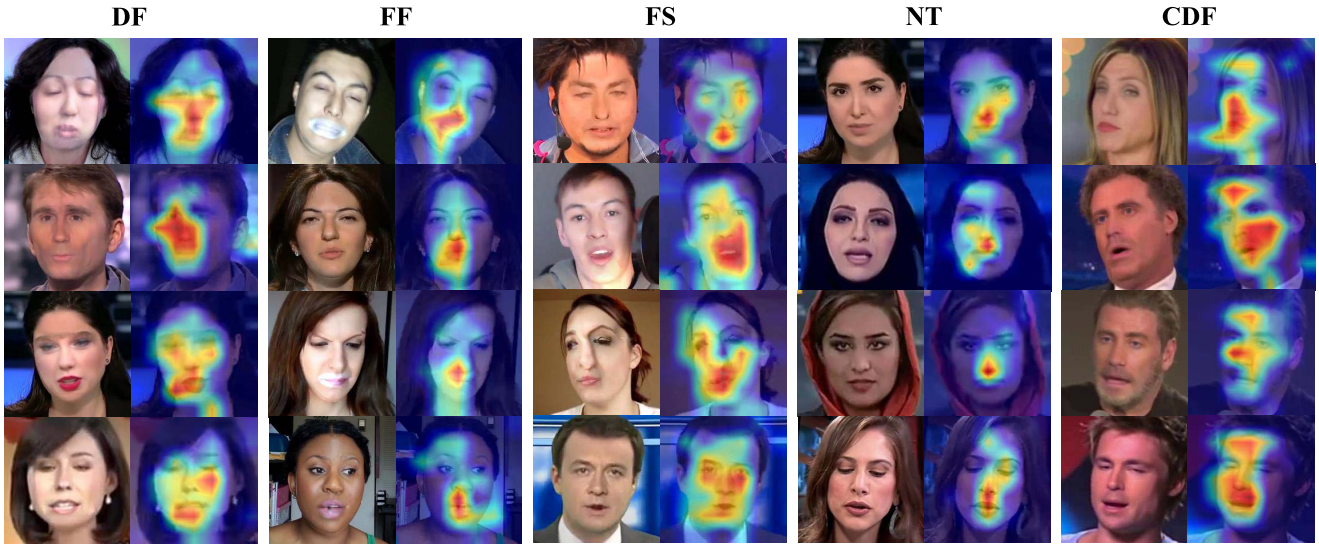


Fig. 6. Attention visualizations of our method via Grad-CAM [94]. The images are randomly selected from DF [4], FF [5], FS [6], NT [3], and CDF [78], [79], with each corresponding to a column. And each column includes RGB images and corresponding heat maps. Note that the model is trained on FF++ (C40) and highlighted regions indicate the area the model pays attention to. Figure best viewed in color.

TABLE VI

ABLATION STUDY OF THE ONE-LEVEL 2D DWT WITH DIFFERENT FREQUENCY COMPONENTS

Frequency Components				FF++(C40)		CDF-1	std/mix4
LL	LH	HL	HH	AUC	ACC	AUC	ACC
✓				89.06	86.37	78.61	82.63
	✓			89.26	86.33	80.79	83.88
		✓		89.39	86.63	84.23	84.84
			✓	89.55	86.89	84.36	85.78
✓	✓			89.15	85.84	79.33	82.80
		✓	✓	89.48	86.68	85.81	87.04
✓	✓	✓		88.90	86.35	80.51	84.88
✓	✓	✓	✓	89.59	86.62	84.33	86.03
	✓	✓	✓	**89.60**	**86.98**	**86.29**	**87.75**

TABLE VII

ABLATION STUDY OF THE DIFFERENT DOWN-SAMPLING METHODS

Methods	FF++(C40)		CDF-1	std/mix4
	AUC	ACC	AUC	ACC
Average Pool	89.17	86.02	82.50	85.86
Mixed Pool	88.25	86.08	83.19	84.93
Stochastic Pool	88.78	86.27	83.03	83.45
MaxBlur Pool	89.33	86.76	84.62	86.58
Strided Conv.	89.38	86.90	83.74	84.07
HWS	**89.60**	**86.98**	**86.29**	**87.75**

TABLE VIII

ABLATION STUDY OF THE DIFFERENT WAYS TO EXPLOITING FINE-GRAINED FORGERY INFORMATION

Methods	FF++(C40)		CDF-1	std/mix4
	AUC	ACC	AUC	ACC
Conv	89.20	85.99	76.97	84.36
CDC	89.21	86.83	79.45	83.09
Att.	89.26	86.37	80.40	82.03
CDA	**89.60**	**86.98**	**86.29**	**87.75**

TABLE IX

ABLATION STUDY OF THE DEPTH AND COMPUTATIONAL COMPLEXITY OF NETWORK

Methods	Flops	Param	CDF-1	std/mix4
			AUC	ACC
ViT-B [18]	16.86 G	85.8 M	74.16	86.57
ViT-L [18]	59.67 G	303.3 M	74.10	84.58
Swin-B [61]	15.14 G	86.75 M	73.84	74.63
Swin-L [61]	34.04 G	195.0 M	73.86	69.15
$\{1, 1, 1, 1\}$	19.72 G	117.52 M	86.29	87.75
$\{1, 1, 3, 1\}$	21.78 G	128.01 M	**86.46**	**89.60**
$\{2, 2, 2, 2\}$	23.84 G	145.39 M	83.75	84.76
$\{2, 2, 6, 2\}$	27.95 G	166.39 M	82.44	83.12

generalization and robustness evaluation across the CDF-1 and std/mix4 datasets. For different settings, the single-frequency sub-band does not perform as well as the others. In addition, the LL sub-band may lead to an inferior generalization and robustness performance of the model on the unseen CDF-1 and std/mix4 datasets. On the contrary, the HH sub-band will enhance these performances. Since the low-frequency component contains a natural image background and real parts, these forgery-independent features have a negative impact on the discriminative artifact features [47], [98]. Therefore, our strategy of discarding the LL sub-band can help the model learn more generalized and robust manipulation traces in the spatial-frequency domain.

3) Analysis on Different Sampler: To verify the proposed HWS module, we take several vanilla samplers (see Sec. III-C) including Average Pool, Mixed Pool [71], Stochastic Pool [72], MaxBlur Pool [73], and Strided Conv. for comparisons. Those vanilla samplers are entirely different from our HWS module, in which they cannot down-sample in both the spatial and frequency domains. Our HWS not only halves the size of the feature maps in the spatial domain but also removes the low-frequency components of the feature maps in the frequency domain. As shown in Table VII, our HWS module boosts the generalization and robustness capacity by extracting the more discriminative forgery patterns in the spatial-frequency domain. The other methods experience a dramatic performance drop on CDF-1 and std/mix4 datasets,

because they may result in the aliasing among low-frequency and high-frequency components. The low-frequency information of the manipulated and corresponding pristine face images is semblable, so our HWS module inhibits their shared parts in the spatial-frequency domain by dropping the LL sub-band of the 2D DWT.

4) Analysis on Proposed CDA: We replace our CDA transformer block with central difference-related methods (*i.e.*, vanilla Conv, central difference convolution (CDC) [2], or traditional self-attention mechanism (Att.) [1]) to verify why the CDA module is specially designed for the face forgery detection task, as shown in Table VIII. The vanilla Conv focuses on the deeper semantic features which suffer from poor generalization and robustness. Meanwhile, the CDC can enhance the performance a little but it is weak in capturing the global-range relation. The Att. models more general forgery features from a global perspective while fine-grained information is still lacking. The proposed CDA module explores detailed fine-grained information between original and manipulated faces via invariant fine-grained keys and values, so it achieves remarkable generalization and robustness ability on the unseen CDF-1 and std/mix4 datasets.

5) Analysis on Depth and Computational Complexity of the Network: The parameters and FLOPs of the proposed F^{2}Trans with different depths are presented in Table IX. It can be observed that as the depth of the model increases, the critical generalization and robustness performance do not get better, instead, $\{1, 1, 1, 1\}$ (F^{2}Trans-S) and $\{1, 1, 3, 1\}$ (F^{2}Trans-B) are the best structures. The larger the model is, the more difficult it is to train it, thus preventing satisfactory representation of forgery features, especially on smaller face forgery detection datasets. Moreover, we further report the computational complexity of classic transformer backbone ViT [18] and Swin [61], which are pre-trained on ImageNet [40]. But they hardly achieve satisfactory performance in terms of generalization and robustness power. This may be because they are targeted for generic vision tasks (*e.g.*, classification, detection, and segmentation), in contrast to our F^{2}Trans which is specifically designed for face forgery detection tasks.

V. Conclusion

In this paper, we specifically design a High-Frequency Fine-Grained Transformer (F^{2}Trans) to extract the invariant high-frequency fine-grained forgery cues in both spatial and frequency domains. To be specific, the CDA transformer block captures the invariant fine-grained forged patterns from a global perspective. The HWS module decomposes the aliasing features into high-frequency components and produces a hierarchical spatial-frequency representation. Meanwhile, the LSC tactic further compensates the spatial-aware local information to multi-level networks. These three modules organically form the F^{2}Trans, which is suitable for face forgery detection and achieves state-of-the-art performance on unseen datasets, unseen manipulations, and real-world perturbations.

References

[1] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[2] Z. Yu et al., "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5295–5305.

[3] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, 2019.

[4] (2019). *DeepFakes*. [Online]. Available: https://github.com/deepfakes/

[5] J. Thies, M. Zollhofer, 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.

[6] (2019). *FaceSwap*. [Online]. Available: https://github.com/MarekKowalski/FaceSwap

[7] J. Zhao et al., "Dual-agent GANs for photorealistic and identity preserving profile face synthesis," in *Proc. NeurIPS*, vol. 30, 2017, pp. 1–11.

[8] Q. Wang, P. Zhang, H. Xiong, and J. Zhao, "*Face.evoLVe*: A high-performance face recognition library," 2021, *arXiv:2107.08621*.

[9] J. Zhao, "Deep learning for human-centric image analysis," Ph. D. dissertation, Dept. Elect. Comput. Eng., Nat. Univ. Singapore, Singapore, 2018.

[10] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, Dec. 2018, pp. 1–7.

[11] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection," in *Proc. 5th ACM Workshop Inf. Hiding Multimedia Secur.*, 2017, pp. 159–164.

[12] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," 2019, *arXiv:1906.06876*.

[13] R. Wang et al., "FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1–8.

[14] C. Gong, D. Wang, M. Li, V. Chandra, and Q. Liu, "Improve vision transformers training by suppressing over-smoothing," 2021, *arXiv:2104.12753*.

[15] D. Zhou et al., "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.

[16] C. Miao, Q. Chu, W. Li, T. Gong, W. Zhuang, and N. Yu, "Towards generalizable and robust face manipulation detection via bag-of-feature," in *Proc. Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2021, pp. 1–5.

[17] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," 2021, *arXiv:2102.11126*.

[18] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

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

[20] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15044–15054.

[21] J. Wang, Y. Sun, and J. Tang, "LiSiam: Localization invariance Siamese network for deepfake detection," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2425–2436, 2022.

[22] 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.

[23] Q. Gu, S. Chen, T. Yao, Y. Chen, S. Ding, and R. Yi, "Exploiting fine-grained face forgery clues via progressive enhancement learning," 2021, *arXiv:2112.13977*.

[24] 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.

[25] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[26] F. Juefei-Xu, V. N. Boddeti, and M. Savvides, "Local binary convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 19–28.

[27] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, "Unmasking deepfakes with simple features," 2019, *arXiv:1911.00686*.

[28] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. ECCV*. Cham, Switzerland: Springer, 2020, pp. 86–103.

[29] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Proc. ECCV*. Cham, Switzerland: Springer, 2020, pp. 667–684.

[30] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6458–6467.

[31] 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.

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

[33] Q. Li, L. Shen, S. Guo, and Z. Lai, "Wavelet integrated CNNs for noise-robust image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7245–7254.

[34] G. Jia et al., "Inconsistency-aware wavelet dual-branch network for face forgery detection," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 3, pp. 308–319, Jul. 2021.

[35] M. Wolter, F. Blanke, R. Heese, and J. Garcke, "Wavelet-packets for deepfake image analysis and detection," 2021, *arXiv:2106.09369*.

[36] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA, USA: SIAM, 1992.

[37] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[38] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[39] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11976–11986.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[41] T. Carvalho, F. A. Faria, H. Pedrini, R. D. S. Torres, and A. Rocha, "Illuminant-based transformed spaces for image forensics," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 4, pp. 720–733, Apr. 2016.

[42] D. Cozzolino, D. Gragnaniello, and L. Verdoliva, "Image forgery localization through the fusion of camera-based, feature-based and pixel-based techniques," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 5302–5306.

[43] P. Ferrara, T. Bianchi, A. D. Rosa, and A. Piva, "Image forgery localization via fine-grained analysis of CFA artifacts," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1566–1577, Oct. 2012.

[44] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 868–882, Jun. 2012.

[45] X. Pan, X. Zhang, and S. Lyu, "Exposing image splicing with inconsistent local noise variances," in *Proc. IEEE Int. Conf. Comput. Photography (ICCP)*, Apr. 2012, pp. 1–10.

[46] B. Peng, W. Wang, J. Dong, and T. Tan, "Optimized 3D lighting environment estimation for image forgery detection," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 2, pp. 479–494, Feb. 2017.

[47] C. Miao et al., "Learning forgery region-aware and ID-independent features for face manipulation detection," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 4, no. 1, pp. 71–84, Jan. 2022.

[48] 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.

[49] W. Zhuang et al., "UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection," in *Proc. ECCV*. Cham, Switzerland: Springer, 2022, pp. 391–407.

[50] 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. 5001–5010.

[51] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15023–15033.

[52] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5781–5790.

[53] T. Dzanic, K. Shah, and F. Witherden, "Fourier spectrum discrepancies in deep network generated images," in *Proc. NeurIPS*, vol. 33, 2020, pp. 3022–3032.

[54] Y. LeCun et al., "Object recognition with gradient-based learning," in *Shape, Contour and Grouping in Computer Vision*. Berlin, Germany: Springer, 1999, pp. 319–345.

[55] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.

[56] X. Li et al., "Sharp multiple instance learning for deepfake video detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1864–1872.

[57] C. Miao, Z. Tan, Q. Chu, N. Yu, and G. Guo, "Hierarchical frequency-assisted interactive networks for face manipulation detection," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 3008–3021, 2022.

[58] Z. Tan, Z. Yang, C. Miao, and G. Guo, "Transformer-based feature compensation and aggregation for deepfake detection," *IEEE Signal Process. Lett.*, vol. 29, pp. 2183–2187, 2022.

[59] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[60] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[61] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[62] Z. Peng et al., "Conformer: Local features coupling global representations for visual recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 367–376.

[63] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. ICML*, 2021, pp. 10347–10357.

[64] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao, "NAS-FAS: Static-dynamic central difference network search for face anti-spoofing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 9, pp. 3005–3023, Sep. 2021.

[65] J. Yang, A. Li, S. Xiao, W. Lu, and X. Gao, "MTD-Net: Learning to detect deepfakes images by multi-scale texture difference," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4234–4245, 2021.

[66] Y. Duan, F. Liu, L. Jiao, P. Zhao, and L. Zhang, "SAR image segmentation based on convolutional-wavelet neural network and Markov random field," *Pattern Recognit.*, vol. 64, pp. 255–267, Apr. 2017.

[67] Y. Gao et al., "High-fidelity and arbitrary face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16115–16124.

[68] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-CNN for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 773–782.

[69] T. Williams and R. Li, "Wavelet pooling for convolutional neural networks," in *Proc. ICLR*, 2018, pp. 1–12.

[70] J. Yoo, Y. Uh, S. Chun, B. Kang, and J.-W. Ha, "Photorealistic style transfer via wavelet transforms," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9036–9045.

[71] D. Yu et al., "Mixed pooling for convolutional neural networks," in *Proc. Int. Conf. Rough Sets Knowl. Technol.* Cham, Switzerland: Springer, 2014.

[72] M. D. Zeiler and R. Fergus, "Stochastic pooling for regularization of deep convolutional neural networks," 2013, *arXiv:1301.3557*.

[73] R. Zhang, "Making convolutional networks shift-invariant again," in *Proc. ICML*, 2019, pp. 7324–7334.

[74] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[75] A. Rossler, 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.

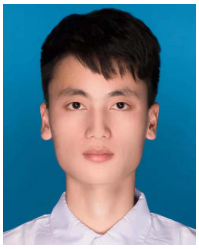
[76] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.

[77] P. Korshunov and S. Marcel, "DeepFakes: A new threat to face recognition? Assessment and detection," 2018, *arXiv:1812.08685*.

[78] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," 2019, *arXiv:1909.12962*.

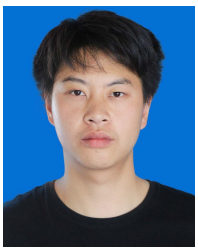
[79] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3207–3216.

[80] K. Sun et al., "Domain general face forgery detection by learning to weight," in *Proc. AAAI*, vol. 35, no. 3, 2021, pp. 2638–2646.
[81] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, and R. Ji, "Dual contrastive learning for general face forgery detection," in *Proc. AAAI*, vol. 36, no. 2, 2022, pp. 2316–2324.
[82] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2889–2898.
[83] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proc. 4th ACM Workshop Inf. Hiding Multimedia Secur.*, 2016, pp. 5–10.
[84] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in *Proc. IEEE Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2017, pp. 1–6.
[85] 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.
[86] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? Understanding properties that generalize," in *Proc. ECCV*. Cham, Switzerland: Springer, 2020, pp. 103–120.
[87] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot…for now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 7, Jun. 2020, pp. 1–10.
[88] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2307–2311.
[89] Z. Gu et al., "Spatiotemporal inconsistency learning for deepfake video detection," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3473–3481.
[90] 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.
[91] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019, pp. 6105–6114.
[92] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "ForensicTransfer: Weakly-supervised domain adaptation for forgery detection," 2018, *arXiv:1812.02510*.
[93] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proc. AAAI*, 2018, pp. 1–8.
[94] 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.
[95] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.
[96] L. Wang et al., "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. ECCV*. Cham, Switzerland: Springer, 2016, pp. 20–36.
[97] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.
[98] K. Xu, M. Qin, F. Sun, Y. Wang, Y.-K. Chen, and F. Ren, "Learning in the frequency domain," in *Proc. CVPR*, 2020, pp. 1740–1749.



Changtao Miao (Member, IEEE) received the B.S. degree from Anhui University in 2019. He is currently pursuing the M.S. degree in cyber science and technology with the University of Science and Technology of China. His research interests include face forgery forensics and face manipulation.

Zichang Tan (Member, IEEE) received the B.E. degree from the Department of Automation, Huazhong University of Science and Technology (HUST), in 2016, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CAS), in 2021. Since July 2021, he has been with Baidu Inc., as a Researcher. His main research interests include deep learning, computer vision, and biometrics in particular. He was named as an Outstanding Graduate of the College when he graduated, and a Winner of the 2020 CAS President Award.



Qi Chu received the B.S. degree in electronic engineering and the Ph.D. degree in information and communication engineering from the University of Science and Technology of China in 2014 and 2019, respectively. He is currently an Associate Research Fellow with the School of Cyber Science and Technology, University of Science and Technology of China. His research interests include deep learning, computer vision, and artificial intelligence security.

Huan Liu received the B.S. degree from Beijing Jiaotong University, Beijing, China, in 2020, where he is currently pursuing the Ph.D. degree with the Institute of Information Science. His research interests are forgery forensics and deep learning.

Honggang Hu received the B.S. degree in mathematics and the B.E. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in June 2000 and June 2001, respectively, and the Ph.D. degree in electrical engineering from the Graduate School of Chinese Academy of Sciences, Beijing, China, in July 2005. From July 2005 to April 2007, he was a Post-Doctoral Fellow at the Institute of Software, Chinese Academy of Sciences, Beijing. From May 2007 to July 2007, he was a Research Assistant Professor at the Institute of Software, Chinese Academy of Sciences. From August 2007 to July 2009, he was a Post-Doctoral Fellow at the University of Waterloo, Canada. From August 2009 to August 2011, he was a Research Associate at the University of Waterloo. Since September 2011, he has been a Full Professor with the University of Science and Technology of China. His research interests include cryptography and coding theory.

Nenghai Yu received the Ph.D. degree from the University of Science and Technology of China in 2004. He was a Visiting Scholar at the Faculty of Engineering, Institute of Production Technology, The University of Tokyo, in 1999, and did cooperative research as a Senior Visiting Scholar at the Department of Electrical Engineering, Columbia University, from April to October 2008. He is currently a Full Professor with the University of Science and Technology of China. His research interests include image processing and video analysis, multimedia communication, media content security, internet information retrieval, data mining and content filtering, network communication, and security.