Hindawi
IET Biometrics
Volume 2024, Article ID 6523854, 14 pages
https://doi.org/10.1049/2024/6523854

*Research Article*

# Face Forgery Detection with Long-Range Noise Features and Multilevel Frequency-Aware Clues

**Yi Zhao**,[1] **Xin Jin**,[2] **Song Gao**,[2] **Liwen Wu**,[2] **Shaowen Yao**,[2] **and Qian Jiang**[2]

[1]*School of Information Science and Engineering, Yunnan University, Kunming, Yunnan 650091, China*
[2]*The Engineering Research Center of Cyberspace and the School of Software, Yunnan University, Kunming, Yunnan 650091, China*

Correspondence should be addressed to Shaowen Yao; yaosw@ynu.edu.cn and Qian Jiang; jiangqian@ynu.edu.cn

Received 23 November 2023; Revised 22 December 2023; Accepted 18 January 2024; Published 5 February 2024

Academic Editor: Vincenzo Conti



The widespread dissemination of high-fidelity fake faces created by face forgery techniques has caused serious trust concerns and ethical issues in modern society. Consequently, face forgery detection has emerged as a prominent topic of research to prevent technology abuse. Although, most existing face forgery detectors demonstrate success when evaluating high-quality faces under intra-dataset scenarios, they often overfit manipulation-specific artifacts and lack robustness to postprocessing operations. In this work, we design an innovative dual-branch collaboration framework that leverages the strengths of the transformer and CNN to thoroughly dig into the multimodal forgery artifacts from both a global and local perspective. Specifically, a novel adaptive noise trace enhancement module (ANTEM) is proposed to remove high-level face content while amplifying more generalized forgery artifacts in the noise domain. Then, the transformer-based branch can track long-range noise features. Meanwhile, considering that subtle forgery artifacts could be described in the frequency domain even in a compression scenario, a multilevel frequency-aware module (MFAM) is developed and further applied to the CNN-based branch to extract complementary frequency-aware clues. Besides, we incorporate a collaboration strategy involving cross-entropy loss and single center loss to enhance the learning of more generalized representations by optimizing the fusion features of the dual branch. Extensive experiments on various benchmark datasets substantiate the superior generalization and robustness of our framework when compared to the competing approaches.

## 1. Introduction

Face forgery refers to a series of computer graphics-based or deep learning-based techniques that can reenact the expression or swap the identity of the source face in an image to the target face [1, 2]. The advent of face forgery techniques has brought about revolutionary transformations in the entertainment industry and visual arts. Nonetheless, the potential for unscrupulous abuse of these techniques, e.g., creating fake news and spreading false political propaganda, which poses grave threats to personal privacy and information security [3]. Against this background, the development of effective methods for detecting facial forgery holds paramount significance, particularly within real-world scenarios, as they play a critical role in enhancing the trustworthiness of digital facial media.

In the current period, many face manipulation detection works [4–8] rely on convolutional neural networks (CNN) to extract forgery traces and distinguish manipulated faces. Unfortunately, these works suffer a severe performance drop when evaluating cross-dataset or cross-quality scenarios. Although, other CNN-based approaches alleviate these issues by introducing different prior knowledge into the backbone network [9, 10] and improving the representation learning paradigm [11–13], the extracted features still overfit to manipulation-specific artifacts. This is because CNN-based architectures have a limited receptive field, making it challenging to capture the global representation and easier to be disturbed by manipulation-specific inductive bias [14, 15]. Meanwhile, inspired by the advantage of the visual transformer in capturing long-range interregion relationships in the visual task, some methods combine the self-attention mechanism with CNN-based architectures or insert a few transformer layers into CNNs for face forgery detection [16, 17]. However, extracting global content information in RGB space through vision transformers is not optimal for

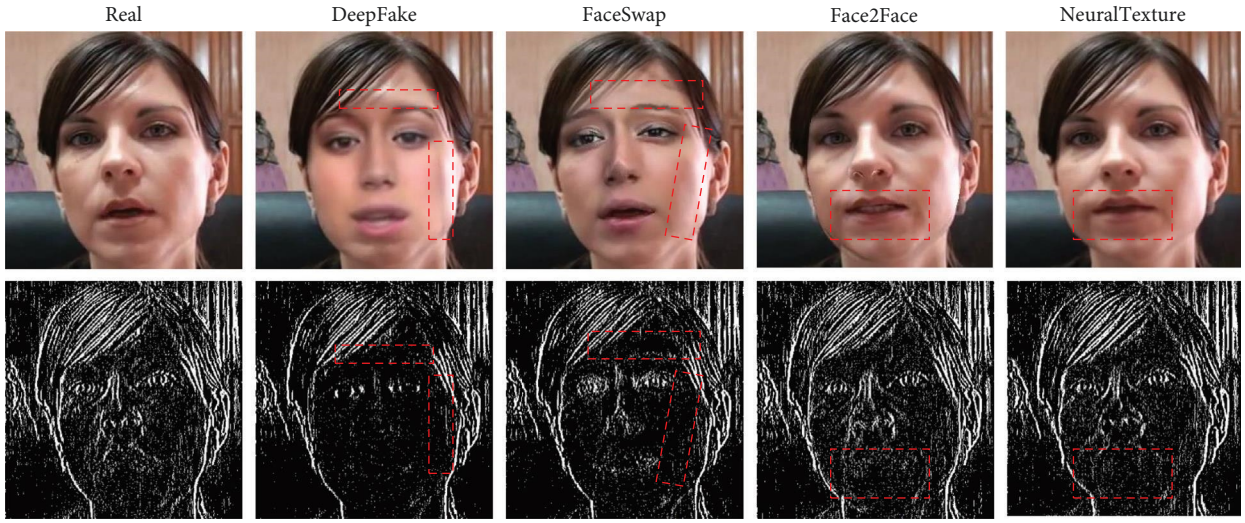


FIGURE 1: Visualization examples of SRM noise maps for real face and four manipulation scenes in the FF++ [20] low-compression subdatasets. Red boxes mark the disparities in noise domain between certain artifact regions and the remaining regions in the face.

forgery detection, as regional differential features in high-fidelity forged faces are difficult to expose in RGB space. Furthermore, the transformer-based architecture presents limitations in capturing local forgery clues since the self-attention mechanism emphasizes global context information while ignoring fine-grained features [18]. To this end, extracting forgery clues from both local and global perspectives is complementary and critically essential for face forgery detection.

Our work is motivated by two primary considerations. On the one hand, capturing interregion relationships in the noise domain from a global perspective contributes to improving the generalization ability of the detector. As clarified by recent approaches [9, 19], image noises can remove high-level semantic content while amplifying forgery artifacts. We show some examples in Figure 1 where the forged faces would expose noise differences between certain artifact regions and the remaining regions in the face. In contrast, the noise distribution of an authentic image remains continuous across the whole face. On the other hand, in the scenario where the visual quality of the manipulated face degrades, subtle forgery artifacts can be described within the frequency domain. As illustrated in Figure 2, the low-frequency components primarily capture the face content information, which easily causes confusion in discriminating between authentic and forged images. Conversely, the middle and high-frequency components can describe fine-grained forged details in multiple scales. Therefore, our study would emphasize both long-range noise trace and multilevel frequency-aware clues, as they play a crucial role in generalization and robustness.

Under these two motivations, we devise an innovative dual-branch collaboration framework that leverages the strengths of the transformers and CNNs to fully explore multimodal forgery artifacts from both global and local perspectives. Specifically, we design a novel adaptive noise trace enhancement module (ANTEM) to extract tampering artifacts from the noise domain for the transformer-based branch. The ANTEM casts off incomprehensive prior knowledge by introducing the restrictedly learnable steganalysis rich model (RSRM) filters and further reinforces the forged traces in the noise domain with a feature reuse block (FRB). Subsequently, the enhanced noise features are split into patches to extract long-range noise features by transformer-based branch. In the CNN-based branch, we devise a multilevel frequency-aware module (MFAM) consisting of data preprocessing and a multilevel feature refinement block (MFRB). The data preprocessing aims to decompose faces in the frequency domain and extract forgery clues from the middle and high-frequency components. Considering that middle and high-frequency components contain forged clues of different scales, the MFRB set convolutions with different dilation ratios to refine the multiscale frequency-aware clues. Then, the CNN-based branch thoroughly extracted the frequency-aware forged clues from the local perspective. Finally, the multimodal features of the dual branches are further fuzed and projected into a compact embedding space. Through the supervision of a collaborative strategy involving cross-entropy loss and single-center loss [21], our framework acquires a more generalized and robust representation of forged clues.

In order to validate the effectiveness of our framework, we perform thorough evaluations on diverse benchmark datasets, encompassing FaceForensics++ [20], Celeb-DF [22], and DFDC [23]. The experimental outcomes clearly indicate that our framework surpasses the performance of competing approaches. The main contributions of our work can be succinctly outlined as follows:

(1) There are two key features for detecting forged faces: long-range noise features and multilevel frequency-aware clues. A novel dual-branch collaboration framework is proposed that takes full advantage of both the transformer and CNN to mine the multimodal forgery artifacts from a global and local perspective.
(2) To cast off introducing incomprehensive prior knowledge, we design an ANTEM, which extracts and

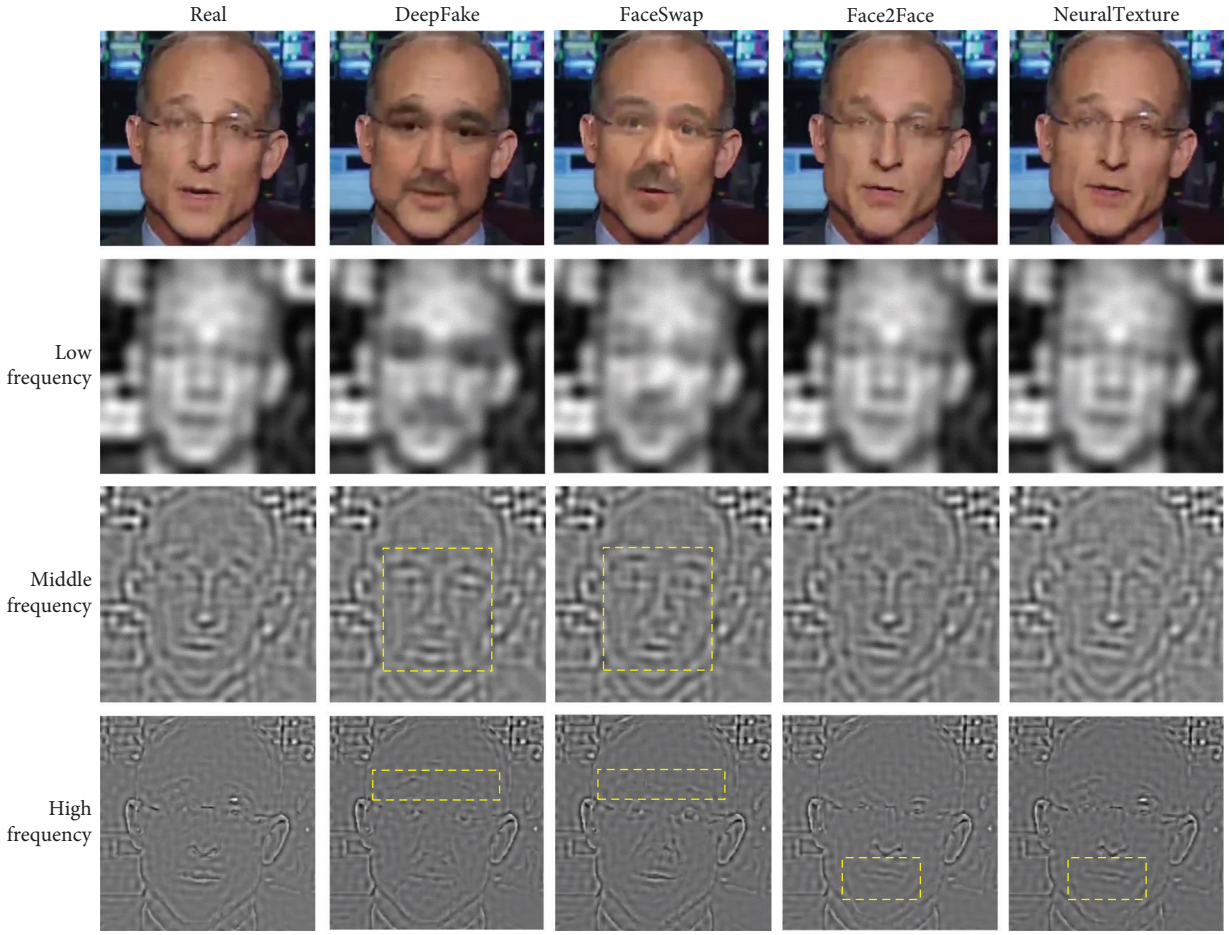


FIGURE 2: Visualization examples of frequency decomposed components for the real and four manipulation scenes in the FF++ [20] high-compression subdatasets. Capturing the differences between the real and forged images is challenging within the low-frequency component. Two face-swapping methods expose subtle differences in facial textures within the middle-frequency component. Meanwhile, four tampering methods expose subtle manipulation traces in different facial regions within the high-frequency component.

reinforces the generalized noise features fitting the forged clues in a data-driven fashion.

(3) We propose an MFAM to decompose faces and refine forgery clues from the middle and high-frequency components, fully leveraging robust frequency-aware features to escape the vulnerability of spatial artifacts.

(4) Extensive experiments on a range of benchmark datasets confirm the outstanding performance of our framework compared to that of competitors.

## 2. Related Work

*2.1. CNN-Based Face Forgery Detection.* Given the rapid evolution of deep learning, the majority of face forgery detection methods rely on CNNs to detect manipulated faces. Early methods attempted to capture forgery discriminant features directly in the spatial domain. Afchar et al. [4] proposed Meso-4 and Meso-Inception-4, which are shallow neural networks designed to extract mesoscopic features. Nguyen et al. [5] utilized the power representation of the capsule network and combined it with the pretrained VGG19 for face forensics. Rössler et al. [20] are the first to introduce the Xception network into deepfake detection, achieving superior performance at that time. Furthermore, some methods have incorporated spatial attention mechanisms to capture local forgery artifacts. For instance, Dang et al. [6] proposed three supervised intensity strategies to train attention maps that direct the network's attention toward tampered regions [7]. have integrated texture features with multiple attention maps to steer the network's attention toward fine-grained features in distinct local regions. However, these approaches tend to overfit to specific forgery methods, and the distinctive features extracted from the spatial domain may deteriorate as the quality of the manipulated faces decreases.

Subsequent studies have noticed these problems and attempted to address them by incorporating different prior knowledge and improving the representation learning paradigm within the backbone networks to extract generalized and robust forgery detection features. For the former, several approaches [9, 19, 24] integrated steganalysis rich model (SRM) filters into the analysis framework, enabling the exploration of generalized forgery features in the noise

domain. Zhao et al. [25] proposed a cross-modality feature pyramid block interacting shallow subtle noise features with deep semantic features to obtain generalized forgery detection representation. Yang et al. [26] identified that the multiscale texture diversity exists between real and manipulated faces, and they introduced the central difference convolution to model the texture difference descriptor. Besides, certain studies have incorporated frequency information in an effort to enhance the robustness of low-quality data. In [10], two complementary branches, namely the frequency-aware decomposition (FAD) and local frequency statistics (LFS), were introduced to detect forged patterns in the frequency domain at multiple scales. SPSL [27] also integrated RGB images with phase spectra to extract the upsampling anomalies of forged faces, thereby enhancing the transferability of the detection model. As for the latter, inspired by the application of transfer learning in computer vision applications, Kim et al. [28] have proposed an innovative method called feature representation transfer adaptation learning (FReTAL) that combines knowledge distillation with a long-life learning strategy to extract generalized features. Taking inspiration from the contrastive learning, Wang et al. [29] introduced a novel approach called the localization invariance Siamese network (LiSiam). This network was designed to enforce localization invariance against a variety of image degradations for the purpose of deepfake detection. Their framework utilized pairwise images of varying qualities and a localization consistency loss was proposed to ensure consistent localization between the two segmentation maps. However, owing to the restricted receptive field inherent in CNN-based architectures, the above-mentioned methods have limitations in capturing global representations and are easily disturbed by manipulation-specific inductive bias.

*2.2. Transformer-Based Face Forgery Detection.* The transformer architecture initially showcased its outstanding performance in various natural language processing tasks [30, 31]. Its remarkable capability to capture extensive long-range and global contextual information has stimulated the adaptation of transformer to computer vision (CV) tasks [15, 32] and derived a series of architectures called vision transformers (ViTs). For face forgery detection, some works have [17, 33] directly reshaped the features extracted by CNN into a series of low-dimensional patches and passed them to VITs encoder, achieving certain levels of generalization performance. M2TR [34] operated patches of different sizes to model a multiscale transformer. They further combined the frequency information with RGB features using a cross-modality fusion module to detect local inconsistency. However, the transformer-based architecture has limitations in capturing local forgery clues due to the inherent nature of its self-attention mechanism, which emphasizes global context information while potentially ignoring fine-grained features. Therefore, Trans-FCA [35] proposes a local adjustment block containing a global-–local cross-attention that focuses on fuzing local convolution and global features at each stage of the transformer backbone. F2Trans [36] designed an innovative high-frequency fine-grained transformer. They enhance fine-grained representation ability by replacing the basic self-attention mechanism with the central differential attention mechanism, which aggregates pixel intensity and gradient information. However, extracting global content information in RGB space through vision transformers is not optimal for forgery detection, as regional differential features in high-fidelity forged faces are difficult to expose in RGB space.

## 3. Methods

In this section, we first introduce the ANTEM in Section 3.1. In Section 3.2, the MFAM is devised, which comprises data preprocessing and an MFRB. Section 3.3 introduces the dual-branch collaborative learning framework to comprehensively dig into the multimodal forgery artifacts from both a global and local perspective.

*3.1. Adaptive Noise Trace Enhancement Module.* As the visual quality of forged faces continues to improve, extracting visual forgery clues within the RGB domain poses a greater challenge. Nevertheless, the hidden traces caused by tampering can still be captured in the noise domain [9, 19, 37]. Given such a situation, an ANTEM is proposed to remove face semantic content and amplify more generalized forgery clues.

The structure of ANTEM is shown in the bottom left of Figure 3. ANTEM introduces the RSRM filters within its first layer. SRM is commonly acknowledged as a form of residual extractor utilized in steganalysis tasks, with the goal of suppressing the semantic components of images and constructing a more robust and compact statistical descriptor. The procedure for calculating the residual $R_{ij}$ is outlined as follows:

$$\begin{cases} R_{ij} = \widehat{X_{ij}}\left(N_{ij}\right) - cX_{ij} \\ R_{ij} \longleftarrow \text{trunc}\left(\text{round}\left(\dfrac{R_{ij}}{q}\right)\right) \end{cases}, \tag{1}$$

where $N_{ij}$ denotes the neighboring pixels of pixel $X_{ij}$ and $\widehat{X}_{ij}$ is defined as a predictor of $cX_{ij}$. This predictor acts as the weights of the SRM filter in collaboration with the residual order $c$. The parameter $q$ is introduced to enhance the sensitivity of residuals to spatial inconsistencies. The application of truncation and round functions aims to compute co-occurrence matrices in subsequent steganalysis steps. Diverse weights of SRM filters can be meticulously designed to capture diverse relationships between the central pixel and its neighboring pixels. For example, a second-order residual can be defined as follows: $R_{ij} = X_{i,j-1} + X_{i,j+1} - 2 \times X_{ij} + 0 \times \widehat{N_{ij}}$.

Although some studies have applied SRM filters with manually defined weights to capture manipulated artifacts [9, 19], these approaches introduce incomprehensive prior knowledge and have limitations in adapting to different manipulated methods. In response to these issues, we generalize the residual extractors to learning-based convolutional filters, allowing for trainable weights within the SRM filter. Initially, within the entire residual calculation process, the truncation function is employed for co-occurrence matrix calculations, a feature not utilized within our framework. Moreover, with the introduction of learnability, the round

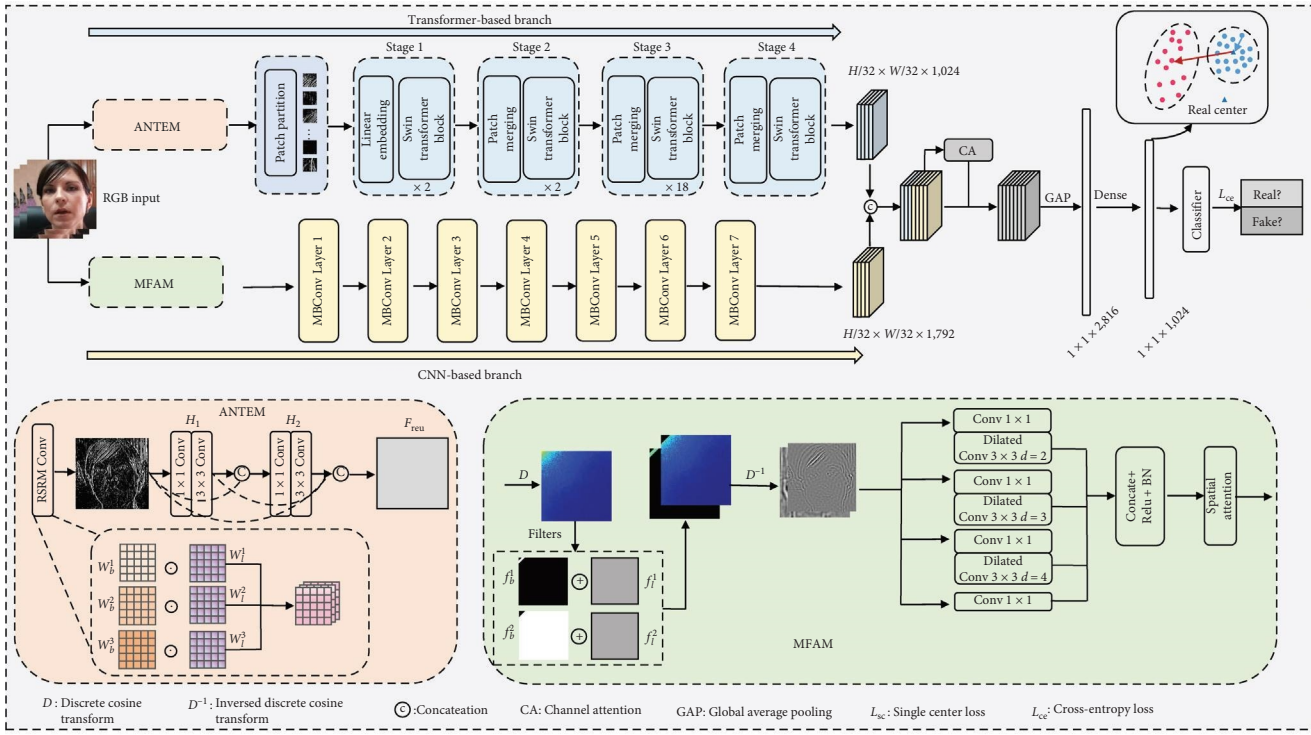


FIGURE 3: The pipeline of the proposed framework. We design a dual-branch collaborative learning framework to comprehensively dig into the long-range noise features and multilevel frequency-aware clues. ANTEM represents the adaptive noise trace enhancement module. MFAM represents the multilevel frequency-aware module. The collaboration strategy of single center loss $L_{sc}$ and cross-entropy loss $L_{ce}$ supervise the framework to learn more generalized and robust features.

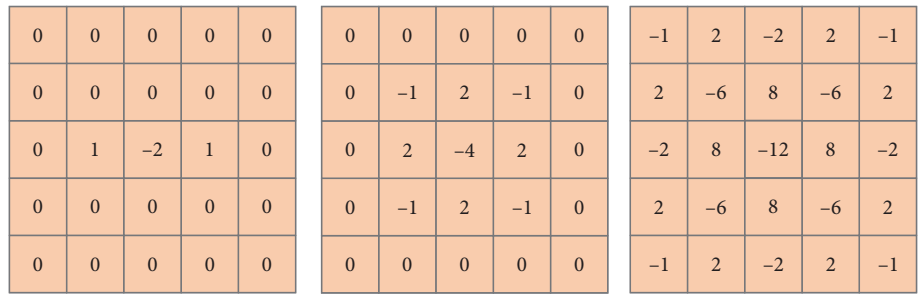


FIGURE 4: The three hand-crafted SRM filters serve as the base kernels.

function and parameter $q$ become redundant. Consequently, we have abandoned the truncation and round functions as well as the parameter $q$ in Equation (1) while ensuring that all operations during the learning process remain differentiable.

After that, in order to prevent the disruption of the original characteristics of SRM filter kernels used for noise computation during the learning process, we have introduced a constrained learning strategy. Specifically, following Zhou et al. [19], we select three SRM kernels that achieved decent performance in the image manipulation detection task and then extend them to $5 \times 5 \times 3$ as the base kernels $W_b \in \mathbb{R}^{5\times5\times3}$. The base kernels are shown in Figure 4. Subsequently, we employ three learnable matrices with dimensions matching the base kernel as the learnable kernel $W_l \in \mathbb{R}^{5\times5\times3}$. The parameters of all three kernels within $W_l$ are individually initialized to 1/2, 1/4, and 1/12, respectively. Consequently, we modify Equation (1), and the definition of RSRM residual is as follows:

$$R = I * (W_b \odot W_l), \tag{2}$$

where $\odot$ denotes element-wise product, $R \in \mathbb{R}^{H\times W\times C}$ represents the residual maps produced by the convolution operation $*$ applied to input image $I \in \mathbb{R}^{H\times W\times C}$ with both base kernels $W_b \in \mathbb{R}^{5\times5\times3}$, and learnable kernels $W_l \in \mathbb{R}^{5\times5\times3}$. Under this strategy, the zero-values within the base kernels remain constant throughout the training process to preserve the originally designed characteristics of these three SRM filter kernels for detecting image manipulation.

Given the fragility of features extracted by the RSRM layer, utilizing them directly might lead to training instability. Drawing inspiration from Huang et al. [38], we devise a FRB to further enhance the noise features. As depicted in the bottom left portion of Figure 3, the FRB comprises two convolutional layers, each consisting of a $1 \times 1$ convolution (kernels = 3, stride = 1, padding = 0) and a $3 \times 3$ convolution (kernels = 3, stride = 1, padding = 1). We define $H_i(\bullet)$ as the composite function of the $i$ convolution layer and $[\alpha_1, \alpha_2, \ldots, \alpha_n]$ denote the concatenation of the n feature maps. The process can be formulated as follows:

$$F_{\text{reu}} = [R, H_1(R), H_2([H_1(R), R])], \tag{3}$$

where $F_{\text{reu}} \in \mathbb{R}^{H \times W \times 3C}$ is the enhanced noise feature maps obtained from the FRB. It is worth noting that the FRB does not incorporate a nonlinear layer, as nonlinear operations could potentially distort the manipulated clues within the noise feature maps. AMTEN, in an adaptive manner, acquires and amplifies manipulation traces within the noise domain, a more appropriate approach for effective face forgery detection.

*3.2. Multilevel Frequency-Aware Module.* In real-world scenarios, forged faces may undergo compression before being shared on social media. Therefore, maintaining robustness to compression is crucial for face forgery detectors. As stated in prior studies [10, 34], subtle forgery artifacts could be captured in the frequency domain even in compressed scenarios. To this end, we have devised an MFAM comprising data preprocessing and an MFRB to extract subtle forgery artifacts from the frequency domain.

As illustrated in the bottom right of Figure 3, the data preprocessing step initially utilizes discrete cosine transform (DCT) to convert the input image channel-wise from RGB to the frequency domain $D(X) \in \mathbb{R}^{H \times W \times 3}$. According to the spectrum characteristics of DCT, the low frequency to high-frequency components progressively distribute from the upper-left corner to the bottom-right corner of the spectrum. Following the approach in [10], we partition the spectrum into low, middle, and high-frequency bands using three hand-crafted binary base filters $\{f_b^i | 1 \le i \le 3\}$ based on a roughly equal energy principle. To be specific, the entire spectrum is empirically divided into three parts: (1) The low-frequency band encompasses the initial 1/16 of the entire spectrum. (2) The middle-frequency band spans from 1/16 to 1/8 of the entire spectrum. (3) The high-frequency band encompasses the remaining 7/8 of the entire spectrum. Given that forgery artifacts typically appear in the middle to high-frequency portions, we utilize two binary base filters $\{f_b^1, f_b^2\}$ to capture the middle and high-frequency information. Subsequently, two learnable filters $\{f_l^1, f_l^2\}$ are incorporated alongside these two binary base filters. These learnable filters serve the purpose of adaptively selecting frequencies of interest beyond the fixed base filters. The resulting decomposed frequency components are defined as follows:

$$C_i = D(X) \odot \left(f_b^i + \theta\left(f_l^i\right)\right), i = \{1, 2\}, \tag{4}$$

where $D$ is DCT, the $\theta(f) = \frac{1-\exp(-f)}{1+\exp(-f)}$ is used for normalizing the value of $f$ between $-1$ and $+1$. To maintain the local consistency and shift-invariance of natural images, we reverse the decomposed components $C_i \in \mathbb{R}^{H \times W \times 3}$ back into the RGB domain using IDCT and reassemble them along the channel axis to obtain the desired representation $C^{-1} \in \mathbb{R}^{H \times W \times 6}$. The process can be calculated as per the equation:

$$C^{-1} = [D^{-1}(C_1), D^{-1}(C_2)]. \tag{5}$$

After data preprocessing, the CNN-compatible frequency representations contain forged clues within the middle to high-frequency range. The MFRB processes $C^{-1}$ using three parallel dilated convolution layers with different dilation rates (2, 3, and 4) to extract multiscale frequency-aware clues. Each dilated convolution layer comprises of a $1 \times 1$ convolution (kernel = 3, stride = 1, padding = 0) and a $3 \times 3$ dilated convolution (kernel = 3, stride = 1, padding = i, dilation = $i$, $i = 2$, 3, 4). Furthermore, an extra $1 \times 1$ convolution (kernel = 3, stride = 1, padding = 0) acts as a skip connection between consecutive blocks. Subsequently, we concatenate these feature maps and apply BatchNormalization and ReLU nonlinearity to integrate multiscale frequency-aware feature maps $C^{-1'} \in \mathbb{R}^{H \times W \times 12}$. Finally, a spatial attention layer is adopted to refine and highlight the manipulation traces in the $C^{-1'}$, as follows:

$$F_{\text{ref}} = \sigma\left(f^{7\times7}\left(\left[\text{AvgPool}\left(C^{-1'}\right); \text{MaxPool}\left(C^{-1'}\right)\right]\right)\right), \tag{6}$$

here, $f^{7\times7}$ denotes convolution with a $7 \times 7$ filter. AvgPool() and MaxPool() refer to average and maximum pooling, respectively.

*3.3. Dual-Branch Collaborative Learning Framework.* The two kinds of forgery clues, namely long-range noise features and multilevel frequency-aware clues, are pivotal in improving generalization and robustness. Considering the distinct characteristics of these two forgery features, we implement a dual-branch collaborative learning framework as shown in Figure 3. This framework is designed to leverage the strengths of both the transformer and CNN, enabling comprehensive exploration of these multimodal forgery clues from both global and local perspectives.

To be specific, for the transformer-based branch, we have employed Swin-B [32] as the backbone of the encoder. The enhanced noise feature maps $F_{\text{reu}} \in \mathbb{R}^{H \times W \times 3C}$ obtained from the ANTEM are subdivided into nonoverlapping regions of size $4 \times 4 \times 3$ to transform the $F_{\text{reu}}$ into sequence embeddings. Then, a linear embedding layer is utilized to convert the embedded dimension into 128. Following this, the features are passed through four consecutive Swin transformer layers, forming hierarchical feature maps that begin with smaller patches and progressively merge them as the network gets deeper in layers. Figure 5 illustrates the two sequential Swin blocks of the Swin transformer layer. Each Swin block is of a layer normalization (LN), a multihead self-attention

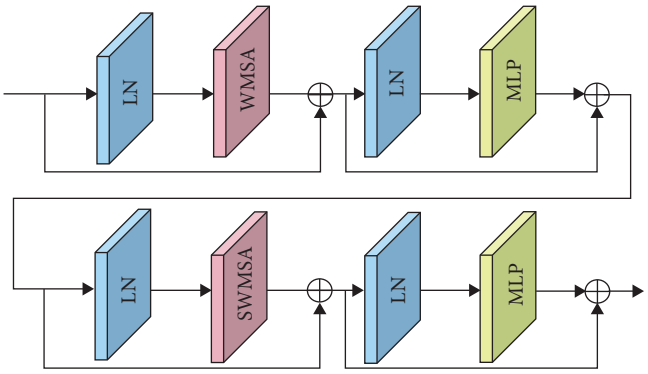


FIGURE 5: Architecture of consecutive Swin transformer blocks.

(MSA) unit, residual connections, and a 2-layer MLP employing GELU nonlinearity. The window-based MHSA (W-MSA) unit and the shifted window-based MHSA (SW-MSA) unit are alternately applied to two consecutive Swin-transformer blocks. Both units perform self-attention within nonoverlapping windows, leading linear computation complexity. Meanwhile, the SW-MHSA unit promotes cross–window interaction without incurring additional computational costs. Compared with the VIT architecture, the Swin-transformer can effectively integrate hierarchical local features while capturing long-range dependencies of noise features.

For the CNN-based branch, we adopt EfficientNetB4 [39] as the encoder's backbone due to its remarkable classification performance with fewer parameters and low-FLOPs costs. As depicted in Figure 2, the refined frequency feature maps $F_{\text{ref}}$ obtained from the MFAM are fed into seven consecutive layers, comprising 2 MBConv1 blocks and 30 MBConv6 blocks. The architecture of MBConv1 block and MBConv6 block are illustrated in Figure 6. The MBConv1 block consists of depthwise convolution, batch normalization (BN), squeeze-and-excitation module (SE) [40], and pointwise convolution with a BN layer. In contrast, the MBConv6 block incorporates an additional pointwise convolution and BN layer when compared to the MBConv1 block. Unlike the vision transformer-based architecture, CNNs still maintain their advantage in extracting spatial local features, which is crucial for capturing subtle forged artifacts.

As described above, the input image $I \in \mathbb{R}^{H\times W\times C}$ is processed through a dual-branch network, getting long-range noise features $F_{\text{noi}} \in \mathbb{R}^{H/32\times W/32\times 1,024}$ and multilevel frequency-aware features $F_{\text{fre}} \in \mathbb{R}^{H/32\times W/32\times 1,792}$, respectively. We concatenate them and apply the channel attention on the connected features $F_c \in \mathbb{R}^{H/32\times W/32\times 2,816}$ to strengthen the pertinent discriminant features based on data characteristics. The process is defined as follows:

$$F_m = \begin{aligned} &\delta(W_1(W_0(\text{AvgPool}(F_c))) \\ &+ W_1(W_0(\text{MaxPool}(F_c))) \end{aligned}, \tag{7}$$

where $W_0 \in \mathbb{R}^{C\times C/r}$ and $W_1 \in \mathbb{R}^{C/r\times C}$ are the MLP weights. Subsequently, the multimodal fusion feature, generated by applying a global average and a dense layer, is utilized for binary class prediction. Considering that the cross-entropy

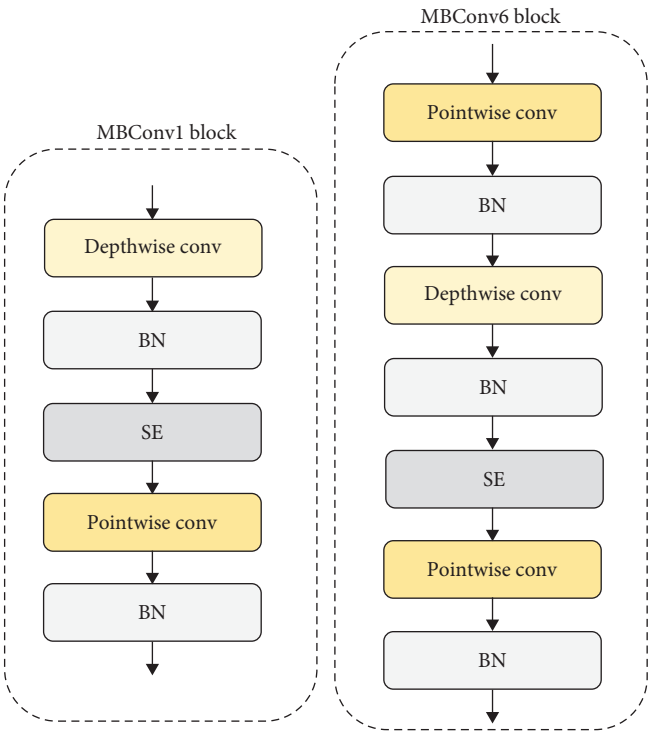


FIGURE 6: The architectural design of the MBConv1 block and MBConv6 block employed within the MBConv layers.

loss primarily guides the network in distinguishing between two specified data distributions in binary classification tasks, this potentially leading to overfitting to data specific to certain forgery methods in the face forgery detection task. Inspired by the insight that the single-center loss [21] can learn decision boundaries suitable for both forged and authentic faces in the face forgery detection task, we have employed a collaborative strategy involving cross-entropy loss and single-center loss to acquire more generalized forgery representations. In detail, given the diverse feature distributions of manipulated faces created by different manipulation methods, aggregating all manipulated faces usually leads to erroneous optimization directions for the network. The single-center loss compresses intra-class variations of real faces within the embedding space, while also promoting interclass differences between fake and real faces. The single-center loss is formulated as follows:

$$L_{\text{sc}} = M_r + \max\left(M_r - M_f + m\sqrt{L}, 0\right), \tag{8}$$

where $M_r$ and $M_f$ represent the mean Euclidean distance from features of real faces and fake faces, respectively, to the center point C within a batch. The margin is designed as $m\sqrt{L}$, where $m$ controls the margin and $L$ is the dimension of features. The mean Euclidean distance function between the two is formulated as follows:

$$\begin{cases} M_r = \dfrac{1}{|\Omega_r|}\sum\limits_{i\in\Omega_i} \|f_i - C\|_2 \\ M_f = \dfrac{1}{|\Omega_f|}\sum\limits_{i\in\Omega_f} \|f_i - C\|_2 \end{cases}, \tag{9}$$

where $\Omega_r$ and $\Omega_f$ represent the set of real faces and fake faces within a batch, respectively. Therefore, the total loss function equation can be expressed as follows:

$$L_{\text{total}} = L_{\text{ce}} + \lambda L_{\text{sc}}, \tag{10}$$

where $\lambda$ serves to control the balance between the collaborative supervised loss functions.

## 4. Experiments

In this section, we initially outline the experimental setups and offer implementation specifics. Subsequently, we will present thorough experimental results to validate the superior performance of our proposed approach.

### 4.1. Experimental Settings

*4.1.1. Datasets.* For a comprehensive evaluation of the performance of our framework in detecting manipulated faces, as well as its generalization across datasets and robustness in different perturbation scenarios, we conduct comprehensive experiments on three widely used face forgery benchmark databases.

(1) FaceForensic++ (FF++) [20] is a widely used facial manipulation database and currently serves as a benchmark for face forgery detection tasks. FF++ comprises 1,000 original videos sourced from YouTube, along with 4,000 videos subjected to manipulation using four distinct methods. These manipulation techniques include two graphics-based approaches, namely FaceSwap and Face2Face, as well as two deep learning-based methods, DeepFakes and NeuralTextures. The database categorizes videos into three quality levels based on their compression levels: c0 (RAW), lightly compressed c23 (HQ), and heavily compressed c40 (LQ). Following the settings outlined in previous work [20], out of the 1,000 videos for each category, 720 videos are used as the training set, and the remaining 280 videos are equally divided for the validation and test sets. During the training process, we address category imbalance between original and manipulated data by augmenting original videos four times. We sample 270 frames from each training video and 100 frames from each video in the validation and testing sets. Given that forgeries encounter in real-world scenarios often exhibit restricted quality, we present the performance metrics for both the lightly compressed and heavily compressed versions of the database.

(2) Celeb-DF [22]: Celeb-DF is a challenging new dataset centered around deepfake-based videos, meticulously crafted using advanced deepfake algorithms to produce high-quality fake videos. This dataset encompasses 590 authentic videos sourced from YouTube and an additional 5,639 fake videos. We utilize the officially disclosed Celeb-DF test set to assess the cross-database generalization performance of our method. From each video within this test set, we sample a total of 32 frames for our evaluation.

(3) DFDC [23]: DFDC is a large-scale dataset initially introduced for the deepfake detection challenge. The DFDC comprises 19,154 authentic videos and 100,000 synthetic videos. The authentic videos exhibit a wide array of subjects and backgrounds, simulating real-world scenarios and encompassing diversity in factors such as skin tone, gender, and lighting conditions. On the other hand, the forged videos are generated using various deepfake techniques, further amplifying the complexity of the detection task. In our experiments, we follow the partitioning scheme in prior research [41]. For evaluating cross-database generalization, we employ the folders designated from 40 to 49 as our test set and sample 32 frames for each video.

*4.1.2. Implement Details.* For video processing and the execution of frame-level experiments, we employ the open-source face detector DLIB [42] to detect and extract faces from each frame. The detected face region is then expanded by a factor of 1.3 around the center of the detected face. Subsequently, the detected faces are resized to $224 \times 224$. In our experiments, we utilize Swin-B, pretrained on ImageNet, as the backbone network for the transformer-based branch and EfficientNetB4, pretrained on ImageNet, as the backbone network for the CNN-based branch. The parameter $m$ in Equation (8) is set to 0.3. The parameter $\lambda$ in the Equation (10) is set to 0.5 to control the tradeoff between $L_{\text{sc}}$ and $L_{\text{ce}}$. During the training process, the framework is optimized using the Adam [43] optimizer with the specified hyperparameter settings ($\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$). The framework was trained for 30, 000 iterations, with a batch size of 32. The initial learning rate is halved if the validation loss does not decrease for three consecutive validation iterations (validation is conducted every 500 iterations). Training is terminated when the learning rate reach $1 \times 10^{-12}$. Our experiments are implemented using the open-source PyTorch platform on a workstation equipped with an Intel (R) i7-10700k CPU and a single NVIDIA RTX 3090 GPU.

*4.1.3. Evaluation Metrics.* To evaluate our proposed method and compare it with other state-of-the-art facial forgery detection approaches, we employ the evaluation metrics of accuracy (Acc) and the area under receiver operating characteristic curve (AUC), consistent with recent research [7, 9, 34]. (1) Acc: We utilize the frame-level Acc the most direct metric to assess the detection performance within the intra-class scene of FaceForensic++. (2) AUC: Frame-level AUC is employed as an additional metric to assess the effectiveness within the intra-dataset scenario and evaluate the generalization performance across different benchmarks, including Celeb-DF and DFDC.

### 4.2. Comparison with Previous Methods.

In this subsection, we perform a series of comparative experiments with prior face forgery detection methods on the three aforementioned datasets to validate the effectiveness, generalization, and robustness of our method.

Table 1: Quantitative comparison results for different quality settings in FF++ dataset.

| Methods | Level | HQ (c23) | | LQ (c40) | |
|---|---|---|---|---|---|
| | Metric | Acc | AUC | Acc | AUC |
| Steg. Features [44] | | 70.97 | — | 55.98 | — |
| LD-CNN [45] | | 78.45 | — | 58.69 | — |
| Constrained Conv [46] | | 82.97 | — | 66.84 | — |
| MesoNet [4] | | 83.1 | — | 70.47 | — |
| DSP-FWA [47] | | — | 57.49 | — | 62.34 |
| Face X-ray [48] | | — | 87.4 | — | 61.6 |
| Xception [20] | | 94.93 | 97.32 | 83.52 | 86.02 |
| EfficientNetB4 [39] | | 95.84 | 98.31 | 85.14 | 87.12 |
| Vit [15] | | 84.32 | 87.73 | 76.53 | 79.81 |
| Swin-B [32] | | 90.64 | 92.32 | 81.68 | 83.74 |
| SPSL [27] | | 91.5 | 95.32 | 81.57 | 82.82 |
| GFFD [9] | | 96.18 | 98.56 | 86.16 | 87.94 |
| MADD [7] | | 97.12 | 99.05 | 85.78 | 87.31 |
| Our | | **97.37** | **99.34** | **88.21** | **89.84** |

The best results are marked in bold fonts.

Table 2: Quantitative comparison results in terms of Acc and AUC on FF++ dataset with four manipulation methods.

| Methods | Manipulations (LQ) | Deepfake | | Face2Face | | FaceSwap | | NeuralTexture | |
|---|---|---|---|---|---|---|---|---|---|
| | Metrics | Acc | AUC | Acc | AUC | Acc | AUC | Acc | AUC |
| Xception [20] | | 92.81 | 94.32 | 85.21 | 87.04 | 91.84 | 93.83 | 75.21 | 77.67 |
| EfficientNet [39] | | 93.12 | 95.64 | 85.32 | 87.21 | 92.38 | 94.23 | 76.41 | 79.13 |
| Vit [15] | | 79.86 | 82.78 | 67.91 | 69.34 | 76.65 | 79.18 | 65.43 | 68.78 |
| Swin-B [32] | | 85.25 | 88.32 | 76.68 | 78.12 | 83.43 | 85.16 | 72.31 | 75.14 |
| GFFD [9] | | 94.02 | 96.01 | 86.02 | 88.23 | 92.52 | 94.22 | 79.21 | 82.97 |
| MADD [7] | | 94.47 | 96.43 | 86.87 | 89.42 | 93.66 | 95.45 | 81.21 | 83.61 |
| Ours | | **95.58** | **97.01** | **88.23** | **90.81** | **94.74** | **96.16** | **83.46** | **86.02** |

The best results are marked in bold fonts.

*4.2.1. Evaluation and Comparison on FF++.*

*(1) Different Video Quality Settings*. We conduct a comparative analysis of our framework with previous detection methods on both the high quality (c23) and low quality (c40) versions of FF++. The results of the comparison are listed in Table 1, with the best results highlighted in bold font. Our approach obviously surpasses the previously referenced methods, including Steg.Features [44], LD-CNN [45], constrained Conv [46], and MesoNet [4], in terms of both Acc and AUC metrics. Xception and EfficientNetB4 are two potent CNN-based backbone networks commonly used in current state-of-the-art face forgery detectors. Under two quality versions, the proposed method achieves a 2.02% and 3.82% improvement in AUC compared to Xception, and a 1.03% and 2.72% improvement compared to EfficientNetB4, respectively. Similarly, compared to transformer-based backbone networks, our approach outperforms ViT by 11.61% and 10.03% and Swin-B by 7.02% and 6.1% in terms of AUC under the two quality versions. Our method also outperforms top-performing methods, including Face X-ray [48] and SPSL [27]. Face X-ray relies on distinguishing differences around mixed boundaries, which can lead to the disruption of mixed traces in high-compression scenarios, limiting its detection performance. On the other hand, SPSL suppresses face semantic content by discarding several convolutional blocks and focusing on local textures. While this direct approach improves generalization, it simultaneously constrains the network's feature extraction capability. Furthermore, compared with the open-source methods GFFD and MADD, we reproduce them in our experimental conditions for a fair comparison. GFFD and MADD can achieve promising performance in the c23 version due to the introduction of different prior knowledge. Nonetheless, the features learned by these two approaches are sensitive to high-compression rates. In contrast, our framework achieves a leading position in both c23 and c40 scenarios. This can be attributed to the introduction of multimodal features from both global and local perspectives, which enhance the effectiveness of intra-dataset scene detection. Additionally, the incorporation of multilevel frequency-aware clues enhances the robustness of our framework in the high-compression scenarios.

*(2) Different Manipulation Methods*. Furthermore, we also assess our framework against four different manipulation methods in FF++. For each manipulation method, the proposed framework is trained and tested on the LQ version to validate the robustness of our framework to compression. The test results are listed in Table 2. It is evident that our approach

Table 3: Quantitative cross-dataset comparison results for AUC metric on Celeb-DF and DFDC, with training on FF++(c23).

| Methods | Training Set | Celeb-DF | DFDC |
|---|---|---|---|
| Xception [20] | FF++(c23) | 65.23 | 68.21 |
| EfficientNetB4 [39] | FF++(c23) | 66.31 | 69.45 |
| Vit [15] | FF++(c23) | 69.14 | 70.31 |
| Swin-B [32] | FF++(c23) | 68.13 | 70.29 |
| M2TR [34] | FF++(c23) | 65.70 | — |
| SPSL [27] | FF++(c23) | **76.88** | 66.16 |
| MD-CSDNet [49] | FF++(c23) | 68.77 | — |
| F3-Net [10] | FF++(c23) | 65.17 | — |
| MTD-Net [26] | FF++(c23) | 70.12 | — |
| MADD [7] | FF++(c23) | 68.21 | 71.02 |
| GFDD [9] | FF++(c23) | 70.13 | <u>72.17</u> |
| Ours | FF++(c23) | <u>73.16</u> | **74.89** |

The best results are denoted in bold, and the second-best results are underlined.

surpasses the CNN-based backbone networks, including Xception and EfficientNet, as well as the transformer-based backbones, such as ViT and Swin-B, for each manipulation type. GFFD enhances generalization by extracting multiscale noise features and suppressing method-specific textures. However, when detecting local manipulation types, such as Face2Face and NeuralTexture, its performance improvement is limited. Noting that NeuralTextures pose the greatest challenge, as they only modify the pixels in the lip region corresponding to facial expressions, leaving behind synthesized realistic faces without noticeable forgery artifacts. While MADD has improved its ability to detect local artifacts by using multiple attention maps and enhancing shallow texture features, it still struggles to capture texture differences effectively in high compression scenarios. Therefore, MADD also achieved limited improvements in the detection of local manipulation types. As shown in Table 2, the test results demonstrate the robustness of our proposed framework. Compared to MADD, our method achieves a 2.26% improvement in AUC for detecting NeuralTexture manipulation. This performance improvement can be primarily attributed to MFAM, which captures multiscale local forgery artifacts in the frequency domain.

*4.2.2. Evaluation and Comparison on Cross-Dataset.* In real-world scenarios, many manipulated facial data remain entirely unknown, as they are generated through Unspecified forgery methods based on undisclosed source faces. Therefore, the generalization performance is crucial for deepfake detection tasks. To assess the generalization capability of our framework in real-world scenes, we conduct cross-dataset evaluation experiments. In particular, the models are trained using four manipulations from FF++ (c23) and tested on the high-quality dataset Celeb-DF, as well as the large-scale dataset DFDC. As shown in Table 3, we highlighted the best results in bold font, and also underlined the second-best results among all the approaches listed. All methods exhibit a noticeable drop in performance when evaluated on unseen datasets compared to the intra-dataset evaluation. The transformer-based backbone networks achieve better generalization performance compared to CNN-based backbone networks, especially ViT. This indicates that the long-range dependent features improve generalization in deepfake detection tasks. Our framework achieves an AUC performance improvement of 6.85% and 5.44% compared to EfficientNetB4 on the Celeb-DF and DFDC datasets, respectively. Similarly, on the Celeb-DF and DFDC datasets, our AUC performance is 5.02% and 4.6% higher than that of Swin-B. Furthermore, in most instances, our method achieves leading-edge performance when contrasted with recent face forgery detectors. This can be attributed to the utilization of long-range noise features and cooperative supervision strategies, which have been discussed in the ablation study. It is worth noting that while SPSL exhibits better generalization performance on Celeb-DF compared to our method, its shallow network architecture comes at the cost of significantly lower AUC scores within the intra-dataset scene.

*4.3. Ablation Study.* In this subsection, we conduct thorough experiments to analyze the effectiveness of various components of our framework in both intra-dataset (FF++ (c40)) and cross-dataset (Celeb-DF) scenarios. Note that our framework and its variants are trained on the FF++ (c40) train set and evaluated on both the intra-dataset (FF++ (c40) test set) and cross-dataset (Celeb-DF test set) scenarios. All ablation experiments are evaluated using AUC and Acc metrics.

*4.3.1. Analysis on Different Components.* We conduct experiments to analyze the proposed modules, including ANTEM, MFAM, the dual-branch network, and the collaboration strategy of single center loss and cross-entropy loss. The experimental results are shown in Table 4. First, we observed in Variant 3 that combining the CNN-based and transformer-based branches improves performance in the cross-dataset scenario, suggesting that integrating both local and global features can enhance the generalization capability of the deepfake detector. Simultaneously, the nearly unchanged performance in FF++ (c40) indicates that the features extracted from the RGB spatial domain lack robustness under high compression. Then, the Variant 4 and Variant 5 further validate that the long-range noise features extracted by the transformer-based branch and the multilevel frequency-aware clues extracted by the CNN-based

Table 4: Ablation studies on different branches and proposed components.

| Methods | Variant | | | | | FF++(c40) | | Celeb-DF |
|---|---|---|---|---|---|---|---|---|
| | EfficientNetB4 | Swin-B | ANTEM | MFAM | SC + CE | Acc | AUC | AUC |
| 1 | ✓ | | | | | 85.14 | 87.12 | 64.12 |
| 2 | | ✓ | | | | 81.68 | 83.74 | 66.41 |
| 3 | ✓ | ✓ | | | | 86.21 | 87.34 | 67.11 |
| 4 | ✓ | ✓ | ✓ | | | 86.23 | 87.32 | 68.78 |
| 5 | ✓ | ✓ | | ✓ | | 87.27 | 88.46 | 67.43 |
| 6 | ✓ | ✓ | ✓ | ✓ | | 87.45 | 88.61 | 69.72 |
| 7 | ✓ | ✓ | ✓ | ✓ | ✓ | **88.21** | **89.84** | **71.02** |

We progressively add each component and compare the detection performance in the intra-dataset (FF++(c40)) and cross-dataset (Celeb-DF) scenarios. The best performances are marked with bold fonts.

Table 5: Ablation studies of ANTEM were carried out in the transformer-based branch within intra-dataset (FF++(40)) and cross-dataset (Celeb-DF) scenarios.

| ID | Variant | FF++(c40) | | Celeb-DF |
|---|---|---|---|---|
| | | Acc | AUC | AUC |
| 1 | Swin-B | 81.68 | 83.74 | 66.41 |
| 2 | Swin-B + SRM-Fix | 80.21 | 82.17 | 67.02 |
| 3 | Swin-B + RSRM | 82.51 | 84.66 | 67.42 |
| 4 | Swin-B + RSRM + FRB | **83.42** | **85.02** | **68.13** |

The best performances are marked with bold fonts.

branch individually enhance the model's generalization capability in deepfake forgery detection and its robustness in compressed scenarios. Furthermore, the simultaneous utilization of ANTEM and MFAM leads to a notable improvement in the framework's capacity for facial forgery detection, as shown in Variant 6. Specifically, when compared to EfficientNetB4, there is an improvement of approximately 1.49% in AUC in the intra-dataset scenario and a 5.6% increase in generalization performance on the AUC score in the cross-dataset scenario. Similarly, compared to Swin-B, there is an improvement of approximately 4.87% in AUC in the intra-dataset scenario and a 3.31% increase in generalization performance on the AUC score in the cross-dataset scenario. These improvements validate the complementary nature of long-range noise features and multilevel frequency-aware clues. Furthermore, our framework achieves state-of-the-art performance when the single center (SC) loss is introduced. This indicates that the SC loss compels the framework to learn classification boundaries within the fuzed feature embedding that are more suitable for deepfake detection tasks, thus avoiding overfitting specific data distributions.

*4.3.2. Analysis on ANTEM.* To validate the effectiveness of the ANTEM, we analyzed the effect of each component in the ANTEM module, including the RSRM filters and FRB. Experiments are conducted on the transformer-based branch, supervised by cross-entropy loss alone. The experimental results are presented in Table 5. From Variant 2 to Variant 3, it can be observed that long-range noise features, as opposed to long-range dependencies extracted from the RGB domain, contribute to enhancing the model's generalization capability in the deepfake detection task. Compared to hand-crafted SRM filters [9] (denoted as SRM-Fix), RSRM learns to fit forged clues more effectively without compromising the essence of noise feature extraction. The results demonstrate that RSRM, by avoiding the introduction of insufficient prior knowledge, exhibits enhanced robustness not only in the FF++ (c40) scenario but also superior generalization in the cross-dataset scenario (Celeb-DF). Furthermore, with the introduction of the FRB in Variant 4, there is an improvement of approximately 0.36% in AUC in the intra-dataset scenario and a 0.71% increase in generalization capability on the AUC score in the cross-dataset scenario compare to Variant 3. This suggests that the FRB further strengthens the noise features extracted by RSRM.

*4.3.3. Analysis on MFAM.* MFAM involves a data preprocessing step to decompose different frequency components and the MFRB for enhancing multilevel frequency-aware clues. Ablation experiments are conducted to assess the impact of different decomposed frequency components and MFRB. We referred to the low, middle, and high-frequency components as LF, MF, and HF. These experiments were performed on the CNN-based branch, with supervision solely by the cross-entropy loss. The experimental results are presented in Table 6. The results from Variant 2 to Variant 3 indicate that extracting forgery clues from mid to high-frequency components contributes to improving the model's detection performance under the high-compression scenario. However, when considering low, mid, and high-frequency information simultaneously in Variant 4, the model's performance even decreased by 0.15% in terms of AUC in the FF++ (c40) scenario and by 1.59% in Celeb-DF compared to Variant 3. This implies that low-frequency components introduce content information unrelated to forgery, thereby confusing the model's ability to distinguish between authentic and

Table 6: Ablation studies of MFAM were carried out in the CNN-based branch within intra-dataset (FF++(40)) and cross-dataset (Celeb-DF) scenarios.

| ID | Variant | | | FF++(c40) | | Celeb-DF |
|---|---|---|---|---|---|---|
| | EfficientNetB4 | Data preprocessing | MFRB | Acc | AUC | AUC |
| 1 | ✓ | | | 85.14 | 87.12 | 64.12 |
| 2 | ✓ | HF | | 86.52 | 88.13 | 65.23 |
| 3 | ✓ | HF + MF | | 86.65 | 88.31 | 65.81 |
| 4 | ✓ | HF + MF + LF | | 85.71 | 87.16 | 64.22 |
| 5 | ✓ | HF + MF | ✓ | **87.22** | **88.41** | **66.32** |

The best performances are marked with bold fonts.

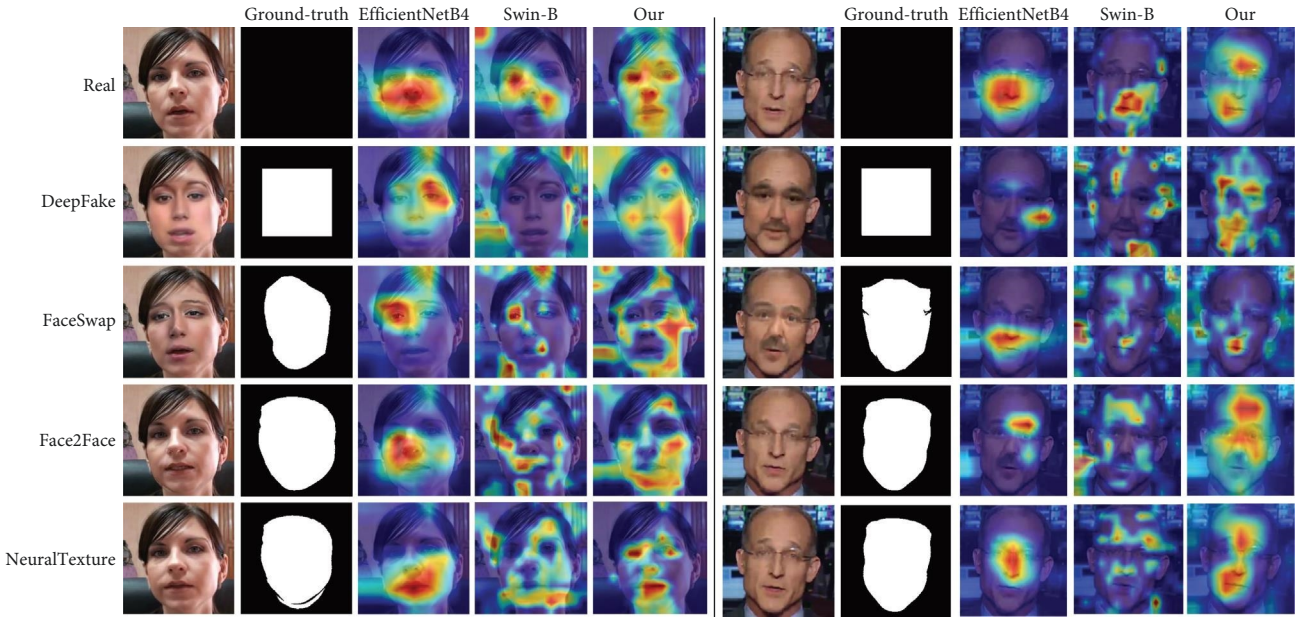


Figure 7: The visualization includes ground-truth, Grad-CAM visualizations of feature maps learned by the baseline models (EfficientNetB4 and Swin-B), and our framework, corresponding to different columns. We separately display the Grad-CAM visualizations of real and four different forgery methods in each row for scenarios c23 and c40. Note that the models are trained on different methods under two compression rates in FF++.

manipulated faces. Furthermore, with the introduction of MFRB in Variant 5, there is a noticeable enhancement in robustness under FF++ (c40) and generalization performance in Celeb-DF compared to Variant 3. This suggests that MFRB further refines and aggregates multilevel frequency-aware clues, allowing the backbone network to extract forgery clues more effectively from the local regions.

*4.4. Visualization Experiments.* To gain deeper insights into the proposed framework, we conducted a comparative analysis of the feature maps learned by our framework and the baseline models (EfficientNetB4 and Swin-B) under different quality settings, as depicted in Figure 7. We utilized gradient-weighted class activation mapping (Grad-CAM) [50], an advanced visualization technique that generates heatmaps to enhance visual explanations of network behavior. Moreover, we separately added ground truth masks in the second column for scenes c23 and c40. In the ground truth masks, white portions represent tampered areas, while black portions indicate unaltered background regions. We observe that, compared to the ground truth, EfficientNetB4 localizes distinct specific areas for various forgery methods. This indicates that CNN-based approaches are prone to introducing method-specific inductive biases, leading to overfitting on particular forgery methods. In contrast to CNNs, it can be observed that Swin-B have a notable capacity to capture long-range relationships through a self-attention mechanism among image patch tokens. However, Swin-B ignores crucial artifacts in the facial region while learning coarse-grained global information. Furthermore, as mentioned in Section 4.3.1, forged artifacts in the RGB spatial domain exhibit vulnerability in high compression scenarios. Therefore, compared to the c23 scenario, both EfficientNet and Swin-B extract fewer informative discriminative features in the c40 scenario. Compared with EfficientNetB4 and Swin-B, our framework not only extracts global features but also prevents the disappearance of subtle facial forgery discriminative features. For example, in the c23 scenario, our framework can capture the blending traces of face-swapping manipulation methods compared to the ground truth, and even in the c40

scenario, our method can still capture local manipulation clues in the lip area of facial reenactment methods. This is attributed to our framework's comprehensive utilization of multimodal forgery features from both global and local perspectives.

## 5. Conclusion

In this paper, we have proposed an innovative dual-branch collaboration framework that leverages the strengths of both transformer and CNN to thoroughly explore multimodal forgery artifacts from both global and local perspectives. Specifically, a novel ANTEM is proposed in the transformer-based branch to remove high-level face content while amplifying more generalized forgery artifacts in the noise domain. An MFAM is developed and further applied to the CNN-based branch to extract complementary frequency-aware clues in middle and high components. Additionally, a collaboration strategy involving cross-entropy loss and single center loss is introduced to enhance the learning of more generalized and robust representations by optimizing the fusion features of the dual branch. Extensive ablation experiments confirm the effectiveness of each component and comprehensive comparative experiments demonstrate the generalization and robustness of our framework.

In our upcoming research, we will explore how to integrate multimodal clues with mask information. We aim to locate the manipulated regions, extract forgery-relevant features with higher precision, and filter out forgery-irrelevant features.

## Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This study is supported by the National Natural Science Foundation of China (No. 62101481), the Basic Research Project of Yunnan Province (Nos. 202301AW070007, 202301AU070210), and the Major Scientific and Technological Project of Yunnan Province (No. 202202AD080002).

## References

[1] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A survey on deepfake video detection," *IET Biometrics*, vol. 10, no. 6, pp. 607–624, 2021.

[2] F. Z. Mehrjardi, A. M. Latif, M. S. Zarchi, and R. Sheikhpour, "A survey on deep learning-based image forgery detection," *Pattern Recognition*, vol. 144, Article ID 109778, 2023.

[3] L. Verdoliva, "Media forensics and DeepFakes: an overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7, IEEE, Hong Kong, China, December 2018.

[5] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: using capsule networks to detect forged images and videos," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2307–2311, IEEE, Brighton, UK, May 2019.

[6] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5781–5790, IEEE Computer Society, Los Alamitos, CA, USA, 2020.

[7] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2185–2194, IEEE Computer Society, Los Alamitos, CA, USA, June 2021.

[8] H. Qi, Q. Guo, F. Juefei-Xu et al., "Deeprhythm: exposing deepfakes with attentional visual heartbeat rhythms," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 4318–4327, ACE Multimedia, 2020.

[9] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16312–16321, IEEE, Nashville, TN, USA, June 2021.

[10] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: face forgery detection by mining frequency-aware clues," in *European Conference on Computer Vision*, pp. 86–103, Springer, Cham, 2020.

[11] G. Lan, S. Xiao, J. Wen, D. Chen, and Y. Zhu, "Data-driven deepfake forensics model based on large-scale frequency and noise features," *IEEE Intelligent Systems*, pp. 1–8, 2022.

[12] M. Le Binh and S. Woo Simon, "Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 122–130, AAAI Press, 2022.

[13] M. Zhang, H. Wang, P. He, A. Malik, and H. Liu, "Exposing unseen GAN-generated image using unsupervised domain adaptation," *Knowledge-Based Systems*, vol. 257, Article ID 109905, 2022.

[14] Z. Peng, Z. Guo, W. Huang et al., "Conformer: local features coupling global representations for recognition and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9454–9468, 2023.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16x16 words: transformers for image recognition at scale," 2020.

[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 *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pp. 1–5, IEEE, Munich, Germany, December 2021.

[17] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining efficientnet and vision transformers for video deepfake detection," in *International Conference on Image Analysis and Processing*, pp. 219–229, Springer, Cham, 2022.

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

[19] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1053–1061, IEEE, Salt Lake City, UT, USA, June 2018.

[20] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "Faceforensics++: learning to detect manipulated facial images," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–11, IEEE, Seoul, Korea (South), 2019.

[21] J. Li, H. Xie, L. Yu, X. Gao, and Y. Zhang, "Discriminative feature mining based on frequency information and metric learning for face forgery detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12167–12180, 2021.

[22] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: a large-scale challenging dataset for deepfake forensics," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3204–3213, IEEE, Seattle, WA, USA, June 2020.

[23] "Deepfake detection challange," 2022, Accessed https://www.kaggle.com/c/deepfake-detection-challenge/.

[24] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1831–1839, IEEE, Honolulu, HI, USA, July 2017.

[25] Y. Zhao, X. Jin, S. Gao, L. Wu, S. Yao, and Q. Jiang, "TAN-GFD: generalizing face forgery detection based on texture information and adaptive noise mining," *Applied Intelligence*, vol. 53, pp. 19007–19027, 2023.

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

[27] H. Liu, X. Li, W. Zhou et al., "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 772–781, IEEE, Nashville, TN, USA, June 2021.

[28] M. Kim, S. Tariq, and S. S. Woo, "FReTAL: generalizing deepfake detection using knowledge distillation and representation learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1001–1012, IEEE Computer Society, Los Alamitos, CA, USA, June 2021.

[29] J. Wang, Y. Sun, and J. Tang, "LiSiam: localization invariance siamese network for deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2425–2436, 2022.

[30] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[31] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, vol. 1, pp. 4171–4186, Association for Computational Linguistics, Minneapolis, MN, USA, 2019.

[32] Z. Liu, Y. Lin, Y. Cao et al., "Swin transformer: hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10022, IEEE Computer Society, Los Alamitos, CA, USA, 2021.

[33] Y.-J. Heo, Y.-J. Choi, Y.-W. Lee, and B.-G. Kim, "Deepfake detection scheme based on vision transformer and distillation," 2021.

[34] J. Wang, Z. Wu, W. Ouyang et al., "M2TR: multi-modal multi-scale transformers for deepfake detection," in *ICMR '22: Proceedings of the 2022 International Conference on Multimedia Retrieval*, pp. 615–623, Association for Computing Machinery, New York, NY, USA, June 2022.

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

[36] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu, and N. Yu, "$F^2$ Trans: high-frequency fine-grained transformer for face forgery detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1039–1051, 2023.

[37] B. Han, X. Han, H. Zhang, J. Li, and X. Cao, "Fighting fake news: two stream network for deepfake detection via learnable SRM," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 320–331, 2021.

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, IEEE, Honolulu, HI, USA, July 2017.

[39] M. Tan and Q. V. Le, "Efficientnet: rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pp. 6105–6114, Long Beach, June 2019.

[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, IEEE, Salt Lake City, UT, USA, 2018.

[41] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 5012–5019, IEEE, Milan, Italy, January 2021.

[42] D. E. King, "Dlib-ml: a machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[43] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2014.

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

[45] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection," in *IH&MMSec '17: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, pp. 159–164, Association for Computing Machinery, New York, NY, USA, June 2017.

[46] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *IH&MMSec '16: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pp. 5–10, Association for Computing Machinery, New York, NY, USA, June 2016.

[47] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," 2018.

[48] L. Li, J. Bao, T. Zhang et al., "Face X-ray for more general face forgery detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5000–5009, IEEE, Seattle, WA, USA, June 2020.

[49] A. Agarwal, A. Agarwal, S. Sinha, M. Vatsa, and R. Singh, "MD-CSDNetwork: multi-domain cross stitched network for deepfake detection," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pp. 1–8, IEEE, Jodhpur, India, December 2021.

[50] 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 *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, IEEE, Venice, Italy, October 2017.