Proceedings of the 42nd Chinese Control Conference
July 24-26, 2023, Tianjin, China

FAClue: Exploring Frequency Clues by Adaptive Frequency-Attention for Deepfake Detection

Weiyun Liang, Yanfeng Wu, Jiesheng Wu, Jing Xu
College of Artificial Intelligence, Nankai University, Tianjin 300350, P. R. China
E-mail: xujing@nankai.edu.cn

Abstract: Detecting fake faces produced by face forgery technologies attracts intensive attention in recent years. Deep learning approaches have shown their effectiveness in deepfake detection task. Some previous deep learning-based methods exploit forgery artifacts in spatial domain but easily overfit the specific forgery patterns. Therefore, some works utilize additional frequency domain information to obtain generalized features. We consider to improve the frequency-based methods in two aspects: 1) extracting discriminative frequency features comprehensively; 2) mining complementary features in different domains sufficiently. In this paper, we propose a dual-stream network named FAClue for deepfake detection, which extracts comprehensive frequency information to complement spatial domain features. Specifically, the FAClue consists of three main components. A Frequency-Attention Extractor (FAE) is proposed to adaptively highlight prominent frequency bands from both global and local perspectives. A RGB-Frequency Complementary Enhancement (RFCE) module is developed to mine complementary information between RGB and frequency domains in an explicit manner. A Frequency Guided Attention (FGA) module is designed to fuse different domain features and generate discriminative features for detection. Extensive experiments on three benchmark datasets demonstrate the FAClue achieves competitive performance compared with state-of-the-art methods.

Key Words: Deepfake Detection, Frequency Domain, Attention Mechanism, Feature Fusion

1 Introduction

In recent years, fake face images are widely disseminated online, posing a threat to portrait rights and personal property. Therefore, it is imperative to develop effective methods for detecting fake faces.

Deepfake detection is often defined as a binary classification task that judges whether a face image is real or fake [1]. Recently, deep learning-based methods have achieved high detection performance and shown robustness to multiple face forgery technologies. Prior works [1–3] exploit forgery artifacts in spatial domain but easily overfit the specific forgery patterns in the training data, degrading the performance on cross-testing [4, 5]. Fortunately, frequency information contains invisible forgery artifacts, which might be useful in complex scenes (e.g., low-quality videos and unseen forgery types) [6]. To alleviate overfitting and obtain generalized features, some works [6–10] exploit frequency domain information extracted by discrete cosine transform (DCT) [11] to assist the spatial domain features. These frequency-based methods extract convolutional neural network (CNN)-compatible frequency representation, and utilize the feature representation capacity of CNNs for frequency-aware deepfake detection [6]. Existing works [6–10] mainly focus on obtaining rich frequency features that are complementary to RGB features. Despite the outstanding performance of these methods, there are two aspects that are worthy of further investigation to enhance the performance of deepfake detection.

Specifically, one is extracting discriminative frequency features comprehensively. Some works [6–8] apply handcrafted or learnable filters on DCT coefficients. Empirically, this filter-based frequency extraction strategy might be coarse, limiting the capacity of extracting discriminative frequency features [9]. Some works [9, 10] perform DCT [11] on local image patches to form manual frequency maps and design a module to further interacts information in different frequencies. We notice that some image patches may only contain real regions. The local frequency features extracted from these patches may lack discriminative information between real and fake regions, while the global frequency features can assist these local features to enhance the feature discrimination. Motivated by these observations and [6, 12], we design a Frequency-Attention (FA) module to adaptively highlight prominent frequency components. We further apply the designed FA modules to global and local DCT coefficients, and integrate the global and local frequency features to generate discriminative frequency features comprehensively.

The other is mining complementary features in different domains sufficiently. Existing works [7, 10] often combine RGB and frequency features by concatenation and convolutions sequentially, and obtain their attention maps by splitting the combined features along channel dimension in multiple stages. The straight convolutions may reduce the complementary features in RGB and frequency domains. Therefore, we mine complementary information along spatial and channel dimensions in a more explicit manner to enhance the complementarity.

Specifically, we propose a dual-stream network for deepfake detection, called FAClue. The FAClue consists of three key parts, i.e., Frequency-Attention Extractor (FAE), RGB-Frequency Complementary Enhancement (RFCE) module, and Frequency Guided Attention (FGA) module. Our main contributions are summarized as follows:

- The FAE is proposed to adopt FA modules to adaptively highlight prominent frequency bands and generate discriminative frequency features from both global and local perspectives.
- The RFCE module is proposed to mine complementary features along spatial and channel dimenssions, making

This work was supported in part by National Natural Science Foundation of China under Grant 62002177, and Natural Science Foundation of Tianjin City under Grant 19JCQNJC00300 and 21JCYBJC00110.

different domain features focus on prominent information in their own domain. Moreover, an FGA module is designed to fuse different domain features and generate discriminative forgery features.

- Extensive experiments demonstrate FAClue achieves competitive performance against state-of-the-art deepfake detection methods on three benchmark datasets.

2 Related Work

2.1 Deepfake Detection

With the rapid development of face forgery technologies, deepfake detection methods based on CNNs achieve excellent performance in deepfake detection task. Prior works [1–3] exploit forgery artifacts in spatial domain but easily overfit the specific forgery patterns in the training data. In order to obtain generalized features, current methods usually use three approaches, i.e., auxiliary task supervision [13–15], multi-modality mining [16, 17], and multi-domain feature learning [6–10, 18–21]. In terms of multi-domain feature learning, different domain features are applied to assist spatial domain features, e.g., steganalysis features [18, 19], temporal clues [20, 21], and frequency features [6–10].

Frequency features have shown effectiveness in computer vision tasks, e.g., image manipulation detection [22] and camouflaged object detection [23]. For deepfake detection, Qian et al. [6] utilize fixed and learnable filters to obtain frequency features from global and local DCT coefficients. Later developments mainly focus on two directions. On the one hand, some works [7, 8] focus on global DCT coefficients, and adopt a well-designed framework with auxiliary task supervision for accurate deepfake detection. Jia et al. [5] propose a frequency adversarial attack method that uses learnable filters for robust deepfake detection. On the other hand, some works [9, 10] focus on local DCT coefficients and propose an elaborate module to extract fine-grained frequency features. Inspired by these pioneering works, we consider to design an attention module for extracting frequency features from DCT coefficients. We further fuse frequency features from both global and local perspectives to leverage complementary information in global and local DCT coefficients.

2.2 Attention Mechanism

Attention mechanism is commonly used in computer vision tasks to adaptively highlight important features, e.g., SE block [24] and CBAM [12]. In deepfake detection task, attention mechanism is widely applied to enhance the feature representation. Chen et al. [7] propose a RGB-frequency attention module (RFAM) to fuse information in both RGB and frequency domains for more comprehensive feature representation. In order to enhance the feature complementarity between different domains, Gu et al. [10] design a mutual-enhancement module (MEM) to fuse RGB and frequency features. Different from them, we consider to mine complementary information along spatial and channel dimensions in a more explicit manner to enhance the complementarity.

3 Methodology

3.1 Overview

The architecture of the proposed FAClue is shown in Fig. 1. Specifically, given an input RGB image, the FAClue first adopts an FAE to generate discriminative frequency features. Moreover, we design a Frequency-Attention (FA) module to adaptively highlight prominent frequency bands in FAE. Multiple FA modules are applied to global and local DCT coefficients to extract frequency features from global and local perspectives, respectively. Next, the RGB and frequency features are separately fed into two backbones to extract hierarchical forgery features in spatial and frequency domains. To mine complementary information in different domains, the RFCE modules are applied between the two branches. Finally, an FGA module is utilized to fuse different domain features for detection.

3.2 Frequency-Attention Extractor (FAE)

3.2.1 Frequency-Attention (FA) module

Frequency information contains unseen forgery clues in frequency domain, which reveals the abnormal frequency distribution between real and fake faces [6]. Motivated by this and the CBAM [12], we design an FA module, which is directly applied to DCT coefficients, to model the inter-frequency relationships between different frequency bands. The FA module assigns different weights to different positions of DCT coefficients corresponding to different frequency bands. By this way, the FA module adaptively emphasizes the prominent frequency bands.

The inputs of the FA module are DCT coefficients $\mathbf{D} \in \mathbb{R}^{H\times W\times 3}$, where H and W are the height and width, respectively. First, $\mathbf{D}$ is separately fed into a global average pooling (GAP) layer and a global max pooling (GMP) layer. Next, the outputs of the two pooling layers are concatenated, and a 7×7 convolutional layer is adopted to model the inter-frequency relationships between different frequency bands. Then a *sigmoid* function is utilized to generate the weights for different frequency components. Finally, the learned DCT coefficients $\mathbf{D}_{FA} \in \mathbb{R}^{H\times W\times 3}$ are obtained by multiplying the weights and the original DCT coefficients. This procedure can be formulated as

$$\mathbf{D}_{FA} = \mathbf{D} \otimes \sigma(f_{7\times 7}(\mathrm{Concat}[\mathrm{GAP}(\mathbf{D}), \mathrm{GMP}(\mathbf{D})])), \quad (1)$$

where $f_{7\times 7}(\cdot)$ denotes a 7×7 convolutional layer followed by a batch normalization layer. $\otimes$ denotes element-wise multiplication, and $\sigma(\cdot)$ represents the *sigmoid* function. The choice of the 7×7 convolution is discussed in Sec. 4.3.1.

3.2.2 Global and Local Perspectives

Previous works [7–10] utilize single local or global DCT coefficients in frequency domain. Local DCT coefficients are obtained from image patches and some patches may only contain real regions, resulting in lack of forgery information. The global DCT coefficients are extracted from the entire image and can assist these local features. Therefore, we integrate the global and local frequency features to generate discriminative frequency features comprehensively.

The input RGB image is first transformed to YCbCr space. We adopt an FA module on DCT coefficients of the input image, and obtain the global frequency features $\mathbf{D}_{FA}^{global} \in \mathbb{R}^{H\times W\times 3}$. For the local frequency feature extraction, the input image is first divided into $P\times P$ patches. The DCT and an FA module are sequentially performed on each patch.

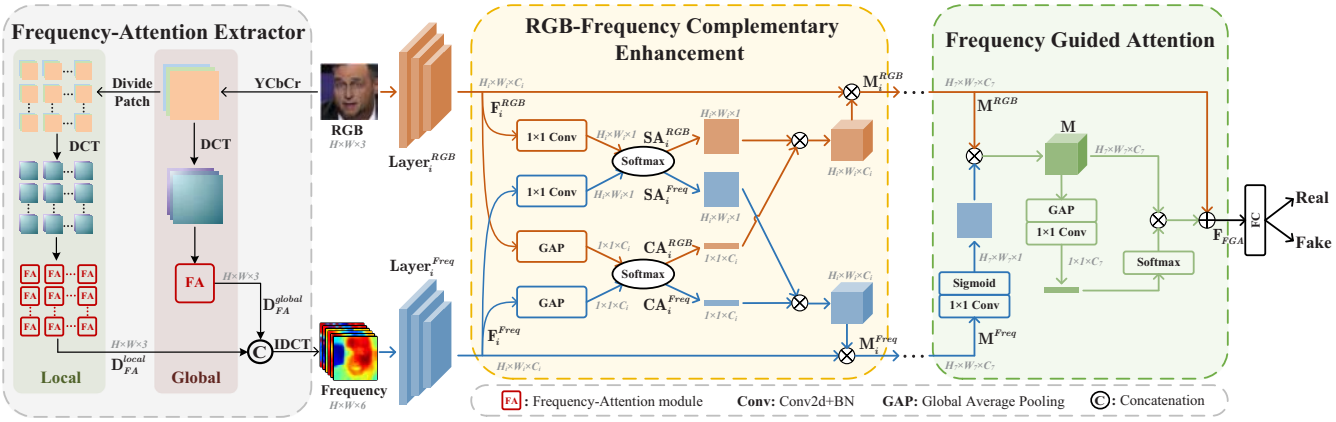

Fig. 1: The overall framework of the proposed FAClue. The FAE learns discriminative frequency features by FA modules from both global and local perspectives comprehensively. The RFCE modules make RGB and frequency features complementary. The FGA module fuses two domain features and generates discriminative forgery features for detection.

Then the learned DCT coefficients of the $P \times P$ patches are stitched together to generate the local DCT coefficients $\mathbf{D}_{FA}^{local} \in \mathbb{R}^{H \times W \times 3}$. The FAE concatenates $\mathbf{D}_{FA}^{global}$ and $\mathbf{D}_{FA}^{local}$, and then performs inverse discrete cosine transform (IDCT) [11] to obtain the final frequency features $\mathbf{F}^{Freq} \in \mathbb{R}^{H \times W \times 6}$. This procedure is defined as

$$\mathbf{F}^{Freq} = \mathrm{IDCT}(\mathrm{Concat}[\mathbf{D}_{FA}^{global}, \mathbf{D}_{FA}^{local}]). \quad (2)$$

3.3 RGB-Frequency Complementary Enhancement (RFCE)

The RGB and frequency features in two streams contain characteristic artifacts that can complement each other [10]. Inspired by [7, 19], we propose an RFCE module to mine complementary information in spatial and frequency domains along spatial and channel dimensions.

We adopt EfficientNet-B0 [25] as the backbone network, which consists of 7 main layers. The inputs of an RFCE module are the outputs of a backbone layer. The input RGB and frequency features of the RFCE module are denoted as $\mathbf{F}_i^{RGB} \in \mathbb{R}^{H_i \times W_i \times C_i}$ and $\mathbf{F}_i^{Freq} \in \mathbb{R}^{H_i \times W_i \times C_i}$, respectively, where i denotes the i-th layer of the backbone. H_i, W_i, and C_i represent the height, width, and channel of the corresponding features, respectively. The spatial attention maps $\mathrm{SA}_i^{RGB} \in \mathbb{R}^{H_i \times W_i \times 1}$ and $\mathrm{SA}_i^{Freq} \in \mathbb{R}^{H_i \times W_i \times 1}$ can be computed by

$$\mathrm{SA}_i^{RGB}, \mathrm{SA}_i^{Freq} = s[f_{1\times1}(\mathbf{F}_i^{RGB}), f_{1\times1}(\mathbf{F}_i^{Freq})], \quad (3)$$

where $s[\cdot,\cdot]$ denotes the element-wise *softmax* operation for its two inputs. $f_{1\times1}(\cdot)$ represents a 1×1 convolutional layer followed by a batch normalization layer. It reduces the feature channel to one to facilitate the attention map generation. The channel attention maps $\mathrm{CA}_i^{RGB} \in \mathbb{R}^{1\times1\times C_i}$ and $\mathrm{CA}_i^{Freq} \in \mathbb{R}^{1\times1\times C_i}$ are defined as

$$\mathrm{CA}_i^{RGB}, \mathrm{CA}_i^{Freq} = s[\mathrm{GAP}(\mathbf{F}_i^{RGB}), \mathrm{GAP}(\mathbf{F}_i^{Freq})]. \quad (4)$$

The final enhanced features $\mathbf{M}_i^{RGB} \in \mathbb{R}^{H_i \times W_i \times C_i}$ and $\mathbf{M}_i^{Freq} \in \mathbb{R}^{H_i \times W_i \times C_i}$ can be computed by

$$\mathbf{M}_i^k = (\mathrm{SA}_i^k \otimes \mathrm{CA}_i^k) \otimes \mathbf{F}_i^k, k \in \{RGB, Freq\}, \quad (5)$$

where the spatial and channel attention maps are multiplied together to generate the final weights for each domain.

The RGB and frequency attention maps focus on different regions by using *softmax* operations on spatial and channel dimensions. Each region corresponds to specific information in its domain. Finally, the output RGB and frequency features concentrate on the prominent information in their own domain, and their complementarity are enhanced. Inspired by [26], we insert RFCE modules after the 2nd, 4th, and 7th layers to enhance the features in low-, middle-, and high-level stages, respectively.

3.4 Frequency Guided Attention (FGA)

Inspired by [4, 27], we utilize an FGA module to fuse complementary features in different domains and generate discriminative features for detection. Specifically, FGA regards frequency features as additional guidance to guide RGB features focusing on unseen forgery clues in frequency domain. The outputs of the RFCE module after the 7th layer, denoted as $\mathbf{M}^{RGB} \in \mathbb{R}^{H_7 \times W_7 \times C_7}$ and $\mathbf{M}^{Freq} \in \mathbb{R}^{H_7 \times W_7 \times C_7}$, will be sent into the FGA module. The FGA module is formulated as

$$\begin{aligned} \mathbf{M} &= \mathbf{M}^{RGB} \otimes \sigma(f_{1\times1}(\mathbf{M}^{Freq})), \\ \mathbf{F}_{FGA} &= \mathbf{M}^{RGB} \oplus (\mathbf{M} \otimes s'(f_{1\times1}(\mathrm{GAP}(\mathbf{M})))), \end{aligned} \quad (6)$$

where $\mathbf{M} \in \mathbb{R}^{H_7 \times W_7 \times C_7}$ is an intermediate variable. $\oplus$ denotes the element-wise addition, and $s'(\cdot)$ is a *softmax* function along the channel dimension. $\mathbf{F}_{FGA} \in \mathbb{R}^{H_7 \times W_7 \times C_7}$ is the output of the FGA module, which will be fed into a fully connected layer for classification.

4 Experiments

4.1 Datasets and Implementation Details

We evaluate the proposed FAClue on three popular fake video datasets, namely FaceForensics++ (FF++) [1], Celeb-DF [28], and WildDeepfake [29]. Each original video in FF++ is manipulated by four forgery types, i.e., Deepfakes (DF), FaceSwap (FS), Face2Face (F2F), and NeuralTextures (NT). Following [31], we use FF++ C23 version (compression factor of 23) in all experiments. We extract frames from videos, and utilize DLIB [30] to detect face regions from the frames with the help of the provided masks. Following the settings in [1], we enlarge face regions by a factor of 1.3

Table 1: Comparison of intra-testing results (AUC and Acc).

Method	Celeb-DF		WildDeepfake	
	AUC	Acc	AUC	Acc
Meso4 (WIFS'18) [2]	0.6617	0.6753	0.6650	0.6447
Reccurent (AVSS'18) [32]	0.8652	0.7120	0.6735	0.6687
FWA (CVPRW'19) [33]	0.6016	0.6473	0.5792	0.5546
Xception (ICCV'19) [1]	0.8975	0.9034	0.8089	0.7526
F3-Net (ECCV'20) [6]	0.9893	0.9595	0.8753	0.8066
Add-Net (MM'20) [29]	0.9955	0.9693	0.8617	0.7625
RFM (CVPR'21) [34]	0.9994	0.9796	0.8392	0.7738
MADD (CVPR'21) [26]	0.9994	0.9792	**0.9071**	0.8286
FTTS (TCSVT'22) [35]	0.8667	0.8074	0.6809	0.6878
FInfer (AAAI'22) [31]	0.9330	0.9047	0.8138	0.7588
FAClue (Ours)	**0.9995**	**0.9883**	0.8991	**0.8298**

Table 2: Comparison of cross-testing results (AUC). "-" denotes the results are unavailable.

Method	FF++ (DF)	Celeb-DF	WildDeepfake
Meso4 (WIFS'18) [2]	0.8470	0.5480	0.5974
Reccurent (AVSS'18) [32]	0.9013	0.6356	0.6703
Multi-task (BTAS'19) [13]	0.7630	0.5430	-
FWA (CVPRW'19) [33]	0.8010	0.5690	0.6735
Xception (ICCV'19) [1]	0.9970	0.6530	0.6054
Capsule (ICASSP'19) [3]	0.9660	0.5750	-
EfficientB4 (ICML'19) [25]	0.9970	0.6429	-
F3-Net (ECCV'20) [6]	0.9810	0.6517	-
Two-branch (ECCV'20) [36]	0.9318	0.7341	-
MADD (CVPR'21) [26]	0.9980	0.6744	-
SPSL (CVPR'21) [37]	0.9691	**0.7688**	-
LTW (AAAI'21) [38]	0.9850	0.6410	-
FTTS (TCSVT'22) [35]	0.9247	0.6556	0.5982
FInfer (AAAI'22) [31]	0.9567	0.7060	0.6946
FAClue (Ours)	**0.9994**	0.7146	**0.7132**

around the center of the tracked faces, and then resize the faces to 224×224. We perform random horizontal and vertical flipping on training data. The patch size in the local part of the FAE is 16×16. Our model is implemented with PyTorch and trained on an NVIDIA RTX 3090 GPU. We adopt EfficientNet-B0 [25] as the backbone network. We utilize SGD optimizer with initial learning rate of 1e-2, weight decay of 1e-4, and momentum of 9e-1. We use cosine annealing scheduler to adjust our learning. The batch size is set as 32. The cross-entropy loss is utilized to supervise the learning process. The model achieving the best performance on validation set is chosen to be evaluated on test set. Following previous works [26, 31], we use area under the receiver operating characteristic curve (AUC) and accuracy score (Acc) for evaluation.

4.2 Comparison with State-of-the-art Methods

4.2.1 Intra-testing

Following the evaluation protocol in [31], we conduct the intra-testing on Celeb-DF and WildDeepfake datasets, and 20 frames per video are sampled to calculate frame-level AUC and Acc. The results indicate that our FAClue achieves competitive performance compared with other methods on the intra-testing, as shown in Table 1.

4.2.2 Cross-testing

To evaluate the generalization capacity of the FAClue, we conduct cross-testing experiments. We train our model on FF++, and test on Celeb-DF and WildDeepfake. We also test on DF category of FF++ in the second column of Table 2. Although Two-branch [36] and SPSL [37] achieve better performance on Celeb-DF, our performance is comparable and is better than theirs on FF++. It can be observed that the FAClue achieves competitive performance compared with most of state-of-the-art methods. The possible reason is that FAClue adaptively highlights useful frequency bands to complement spatial domain features, and the extracted features are more generalized and discriminative for detection.

Table 3: Ablation studies for FAE, RFCE module, and FGA module.

Model	Celeb-DF		WildDeepfake	
	AUC	Acc	AUC	Acc
w/o FAE	0.9983	0.9867	0.8838	0.7895
w/o FAE-G	0.9991	0.9866	0.8791	0.7968
w/o FAE-L	0.9985	0.9830	0.8839	0.8011
FAE $\rightarrow$ FB [7]	0.9990	0.9866	0.8735	0.7878
FAE $\rightarrow$ BG [10]	0.9987	0.9877	0.8822	0.8027
w/o RFCE	0.9993	0.9881	0.8939	0.8033
RFCE $\rightarrow$ RFAM [7]	0.9984	0.9879	0.8861	0.7981
RFCE $\rightarrow$ MEM [10]	0.9985	0.9874	0.8975	0.8092
w/o FGA	0.9990	0.9871	0.8792	0.8000
FGA $\rightarrow$ RGA	0.9972	0.9830	0.8535	0.7762
FAClue (Ours)	**0.9995**	**0.9883**	**0.8991**	**0.8298**

4.3 Ablation Study

In this subsection, we perform ablation studies on Celeb-DF and WildDeepfake following the same evaluation protocol as the intra-testing to verify the effectiveness of each component of the FAClue.

4.3.1 Effectiveness of Frequency-Attention Extractor (FAE)

The results of ablation studies on the FAE are shown in the first five rows of Table 3. In Table 3, "w/o FAE" denotes the FAClue without the FAE; "w/o FAE-G" denotes the FAE without the global perspective; "w/o FAE-L" dentoes the FAE without the local perspective; "FAE $\rightarrow$ FB" represents using the filter-based operation proposed in [7] instead of the FAE; "FAE $\rightarrow$ BG" denotes adopting the band gather operation proposed in [10] instead of the FAE. Experimental results demonstrate the effectiveness of each part in the FAE. It can be observed that the FAE outperforms band gather and filter-based operations proposed in [10] and [7], respectively. The possible reason is that the FAE can adaptively highlight prominent frequency bands and extract frequency features from both global and local perspectives comprehensively, which integrates the advantage of both band gather and filter-based methods.

For the FA module, it is noticed that CNN has inductive biases, i.e., locality and translation equivariance, while DCT coefficients do not have these properties required by convolution. Intuitively, it might be appropriate to use 1×1 convolution because it only interacts information along channel dimension. Surprisingly, the experimental results indicate the 7×7 convolution achieves excellent performance, which can be seen in Table 4. In Table 4, $f_{k\times k} + f_{k\times k}$ denotes a $k \times k$ convolution followed by a $k \times k$ convolution sequentially. The reason might be that the adjacent values of DCT coefficients correspond to similar frequencies. Con-

Table 4: Ablation studies for FA module. We replace the 7×7 convolution in FA module with other structures.

Model	Celeb-DF		WildDeepfake	
	AUC	Acc	AUC	Acc
$f_{1\times1}$	0.9993	0.9882	0.8718	0.7949
$f_{3\times3}$	0.9978	0.9870	0.8832	0.7990
$f_{5\times5}$	0.9993	**0.9884**	0.8723	0.7873
$\boldsymbol{f_{7\times7}}$ **(Ours)**	**0.9995**	0.9883	**0.8991**	**0.8298**
$f_{9\times9}$	0.9989	0.9874	0.8623	0.7752
$f_{1\times1} + f_{1\times1}$	0.9994	0.9880	0.8774	0.7913
$f_{3\times3} + f_{3\times3}$	0.9979	0.9858	0.8933	0.8080
$f_{3\times3} + f_{3\times3} + f_{3\times3}$	0.9994	0.9876	0.8825	0.7990

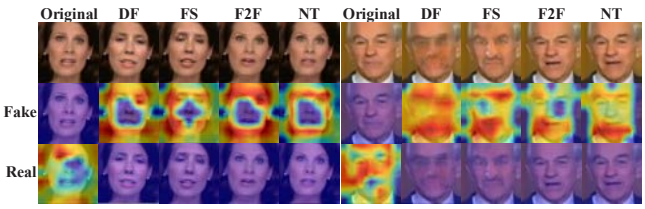

Fig. 2: Grad-CAM visualization of the proposed FAClue. The warm color represents the regions that respond strongly to the prediction of fake (the second row) or real (the third row). The heatmaps highlight the areas around the face boundaries, especially around eyes and mouths.

volution with large kernel may generate weights based on a close range of frequencies rather than just a single frequency, which may use more frequency information to predict more accurate weights for each frequency band.

4.3.2 Significance of RGB-Frequency Complementary Enhancement (RFCE) and Frequency Guided Attention (FGA)

The ablation study results of RFCE and FGA modules are shown in the 6th to 8th and 9th to 10th rows of Table 3, respectively. In Table 3, "w/o RFCE" denotes the FAClue without RFCE modules; "RFCE $\rightarrow$ RFAM" denotes using RFAM [7] instead of the RFCE module; "RFCE $\rightarrow$ MEM" denotes adopting MEM [10] instead of the RFCE module; "w/o FGA" denotes replacing the FGA module with the element-wise addition; "FGA $\rightarrow$ RGA" represents using RGB features to guide frequency features in the guided attention module. Experimental results demonstrate the significance of RFCE and FGA modules. It can be observed that the performance is degraded when the RFCE module is replaced with the RFAM [7] or MEM [10]. It can be explained that the RFCE module makes RGB and frequency features focus on prominent information in their domains, which increases the complementarity of different domain features.

4.4 Grad-CAM Visualization

As shown in Fig. 2, we utilize Grad-CAM [39] to visualize the feature maps extracted by the FAClue. The feature maps are the outputs of the FGA module. Our FAClue is trained on FF++. The warm color represents the regions that respond strongly to the prediction of the corresponding category. It can be observed that the heatmaps highlight the boundary areas of the faces, especially around eyes and mouths. It indicates the FAClue utilizes foreground (facial areas) and background (areas outside the face) information together. The FAClue focuses on the inconsistency between the boundaries of foreground and background regions to detect forgeries. The heatmaps have similar highlighted regions for four manipulation types, which demonstrates the forgery features extracted by the FAClue are generalized. From the above two visualizations and analysis, utilizing facial and background information together can improve the generalization capacity, which is consistent with the conclusion in [40].

As shown in the first and sixth columns of Fig. 2, for original category of FF++, our FAClue shows no response to fake class while shows high response to real class. It indicates that the features extracted by the FAClue are relatively discriminative for deepfake detection.

5 Conclusion

In this paper, we propose FAClue, a dual-stream network for end-to-end deepfake detection. Specifically, the FAE learns discriminative frequency features comprehensively, and RFCE modules make RGB and frequency features complementary in an explicit manner. The FGA module fuses two domain features and generates discriminative forgery features for detection. Our FAClue is compared with several state-of-the-art methods on three benchmark datasets, and achieves competitive performance on both intra-testing and cross-testing. Ablation studies demonstrate the effectiveness of the three main components. Grad-CAM visualizations show that the proposed FAClue can extract generalized and discriminative features for deepfake detection. In the future, we will try to capture more fine-grained frequency features and integrate features from other domains to enhance the generalization capacity.

References

[1] A. Rössler, D. Cozzolino, L. Verdoliva, et al., FaceForensics++: Learning to Detect Manipulated Facial Images, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019: 1-11.

[2] D. Afchar, V. Nozick, J. Yamagishi, et al., MesoNet: A Compact Facial Video Forgery Detection Network, in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2018: 1-7.

[3] H. H. Nguyen, J. Yamagishi, and I. Echizen, Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019: 2307-2311.

[4] J. Cao, C. Ma, T. Yao, et al., End-to-End Reconstruction-Classification Learning for Face Forgery Detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 4103-4112.

[5] S. Jia and C. Ma, T. Yao, et al., Exploring Frequency Adversarial Attacks for Face Forgery Detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 4093-4102.

[6] Y. Qian, G. Yin, L. Sheng, et al., Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues, in *Proceedings of the European Conference on Computer Vision*, 2020: 86-103.

[7] S. Chen, T. Yao, Y. Chen, et al., Local Relation Learning for Face Forgery Detection, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021: 1081-1088.

[8] J. Wang, Z. Wu, W. Ouyang, et al., M2TR: Multi-Modal Multi-Scale Transformers for Deepfake Detection, in *Proceedings of the International Conference on Multimedia Retrieval*, 2022: 615-623.

[9] J. Li, H. Xie, J. Li, et al., Frequency-aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 6454-6463.

[10] Q. Gu, S. Chen, T. Yao, et al., Exploiting Fine-Grained Face Forgery Clues via Progressive Enhancement Learning, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022: 735-743.

[11] N. Ahmed, T. Natarajan, and K. Rao, Discrete Cosine Transform, *IEEE Transactions on Computers*, 100(1): 90-93, 1974.

[12] S. Woo, J. Park, J.-Y. Lee, et al., CBAM: Convolutional Block Attention Module, in *Proceedings of the European Conference on Computer Vision*, 2018: 3-19.

[13] H. H. Nguyen, F. Fang, J. Yamagishi, et al., Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos, in *Proceedings of the IEEE International Conference on Biometrics Theory, Applications and Systems*, 2019: 1-8.

[14] L. Li, J. Bao, T. Zhang, et al., Face X-Ray for More General Face Forgery Detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 5000-5009.

[15] H. Dang, F. Liu, J. Stehouwer, et al., On the Detection of Digital Face Manipulation, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2020: 5780-5789.

[16] T. Mittal, U. Bhattacharya, R. Chandra, et al., Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affective Cues, in *Proceedings of the ACM International Conference on Multimedia*, 2020: 2823-2832.

[17] Y. Zhou and S.-N. Lim, Joint Audio-Visual Deepfake Detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 14780-14789.

[18] X. Wu, Z. Xie, Y. Gao, et al., SSTNet: Detecting Manipulated Faces through Spatial, Steganalysis and Temporal Features, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020: 2952-2956.

[19] Y. Luo, Y. Zhang, J. Yan, et al., Generalizing Face Forgery Detection with High-Frequency Features, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 16312-16321.

[20] P. Chen, J. Liu, T. Liang, et al., FSSpotter: Spotting Face-Swapped Video by Spatial and Temporal Clues, in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2020: 1-6.

[21] Y. Zheng, J. Bao, D. Chen, et al., Exploring Temporal Coherence for More General Video Face Forgery Detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 15024-15034.

[22] Z. Gao, C. Sun, Z. Cheng, et al., TBNet: A Two-Stream Boundary-Aware Network for Generic Image Manipulation Localization, *IEEE Transactions on Knowledge and Data Engineering*, 1-16, 2022.

[23] Y. Zhong, B. Li, L. Tang, et al., Detecting Camouflaged Object in Frequency Domain, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 4504-4513.

[24] J. Hu, L. Shen, and G. Sun, Squeeze-and-Excitation Networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 7132-7141.

[25] M. Tan and Q. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, in *Proceedings of the International Conference on Machine Learning*, 2019: 6105-6114.

[26] H. Zhao, T. Wei, W. Zhou, et al., Multi-Attentional Deepfake Detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 2185-2194.

[27] H. Li, G. Chen, G. Li, et al., Motion Guided Attention for Video Salient Object Detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019: 7273-7282.

[28] Y. Li, X. Yang, P. Sun, et al., Celeb-DF: A Large-Scale Challenging Dataset for Deepfake Forensics, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 3204-3213.

[29] B. Zi, M. Chang, J. Chen, et al., WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection, in *Proceedings of the ACM International Conference on Multimedia*, 2020: 2382-2390.

[30] Y. Cheng, Q. Liu, C. Zhao, et al., *Design and Implementation of Mediaplayer based on FFmpeg*. Berlin Heidelberg: Springer-Verlag, 2012.

[31] J. Hu, X. Liao, J. Liang, et al., FInfer: Frame Inference-Based Deepfake Detection for High-Visual-Quality Videos, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022: 951-959.

[32] D. Güera and E. J. Delp, Deepfake Video Detection using Recurrent Neural Networks, *IEEE International Conference on Advanced Video and Signal based Surveillance*, 2018: 1-6.

[33] Y. Li and S. Lyu, Exposing DeepFake Videos by Detecting Face Warping Artifacts, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[34] C. Wang and W. Deng, Representative Forgery Mining for Fake Face Detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 14918-14927.

[35] J. Hu, X. Liao, W. Wang, et al., Detecting Compressed Deepfake Videos in Social Networks using Frame-Temporality Two-Stream Convolutional Network, *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1089-1102, 2022.

[36] I. Masi, A. Killekar, R. M. Mascarenhas, et al., Two-Branch Recurrent Network for Isolating Deepfakes in Videos, in *Proceedings of the European Conference on Computer Vision*, 2020: 667-684.

[37] H. Liu, X. Li, W. Zhou, et al., Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 772-781.

[38] K. Sun, H. Liu, Q. Ye, et al., Domain General Face Forgery Detection by Learning to Weight, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021: 2638-2646.

[39] R. R. Selvaraju, M. Cogswell, A. Das, et al., Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017: 618-626.

[40] D.-K. Kim and K. Kim, Generalized Facial Manipulation Detection with Edge Region Feature Extraction, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022: 2784-2794.