Local Relation Learning for Face Forgery Detection

Shen Chen[1,2*], **Taiping Yao**[2*], **Yang Chen**[2], **Shouhong Ding**[2†], **Jilin Li**[2], **Rongrong Ji**[1,3†]
[1] Media Analytics and Computing Lab, Department of Artificial Intelligence,
School of Informatics, Xiamen University
[2] YouTu Lab, Tencent
[3] Institute of Artificial Intelligence, Xiamen University
chenshen@stu.xmu.edu.cn, rrji@xmu.edu.cn,
{taipingyao, wizyangchen, ericshding, jerolinli}@tencent.com

Abstract

With the rapid development of facial manipulation techniques, face forgery detection has received considerable attention in digital media forensics due to security concerns. Most existing methods formulate face forgery detection as a classification problem and utilize binary labels or manipulated region masks as supervision. However, without considering the correlation between local regions, these global supervisions are insufficient to learn a generalized feature and prone to overfitting. To address this issue, we propose a novel perspective of face forgery detection via local relation learning. Specifically, we propose a Multi-scale Patch Similarity Module (MPSM), which measures the similarity between features of local regions and forms a robust and generalized similarity pattern. Moreover, we propose an RGB-Frequency Attention Module (RFAM) to fuse information in both RGB and frequency domains for more comprehensive local feature representation, which further improves the reliability of the similarity pattern. Extensive experiments show that the proposed method consistently outperforms the state-of-the-arts on widely-used benchmarks. Furthermore, detailed visualization shows the robustness and interpretability of our method.

1 Introduction

Recent studies have shown rapid progress in facial manipulation, which enables an attacker to manipulate or forge the facial area of human faces, such as Deepfakes (Tora 2018) and FaceSwap (Kowalski 2018). With the remarkable success in synthesizing realistic faces, it becomes infeasible even for humans to distinguish whether an image has been manipulated. At the same time, these forged images may be abused for malicious purposes, causing severe trust issues and security concerns in our society. Therefore, it is of paramount importance to develop effective methods for detecting face forgery.

Early works (Afchar et al. 2018; Nguyen, Yamagishi, and Echizen 2019; Rössler et al. 2019) treat this challenge as a binary classification problem and develop CNNs to model the decision boundary between real and forged faces. However, this setting is known to easy to overfit and lacks interpretability. To relieve this problem, recent works (Nguyen et al. 2019; Stehouwer et al. 2020) introduce the manipulated region mask as supervision to assist networks in locating specific forged regions. Although these methods have achieved remarkable performance on high-quality images, they are vulnerable to disturbances like image compression or noises. Moreover, the generalization issue is still not well addressed. When applied to forgery generated by unseen face manipulation methods, they experience a significant performance drop. Face X-ray (Li et al. 2020a) has noticed this generalization problem and designs new supervision named face X-ray, which focuses on the artifacts caused by image blending between two images. However, it can not be applied to forgery methods without using blending and also be greatly affected by image noises. This means these methods have not learned the intrinsic characteristics in face forgery detection. In a word, most current methods use a kind of global supervision by utilizing binary labels or manipulated region masks at the end of the network, but still fail to learn robust and generalized features.

To address the above issues, we propose a novel perspective of face forgery detection via local relation learning to focuses on the relation of local regions, which is a generalized descriptor that can be used to effectively capture the forged trace, such as abnormal texture (Liu et al. 2020) and high frequency noise (Qian et al. 2020). Specifically, we design a Multi-scale Patch Similarity Module (MPSM) to explicitly model a second-order relationship between different local regions and build a similarity pattern with pair-wise cosine measurement. In this similarity pattern, real regions are similar to each other, so as the forged regions. But real regions and forged regions are not similar. By encoding local forgery semantics which is more robust against disturbances and less sensitive to specific manipulation methods, the similarity pattern further connects them and builds a global face forgery representation, which is more robust and generalized for forgery detection.

Moreover, to further improve the reliability of the similarity pattern, we propose an RGB-Frequency Attention Module (RFAM) to both consider RGB and frequency information and collaboratively learn a comprehensive local representation. Concretely, we first transform the image into frequency domain and amplify the artifacts hidden in high fre-

[*] Equal contribution.
[†] Corresponding authors.

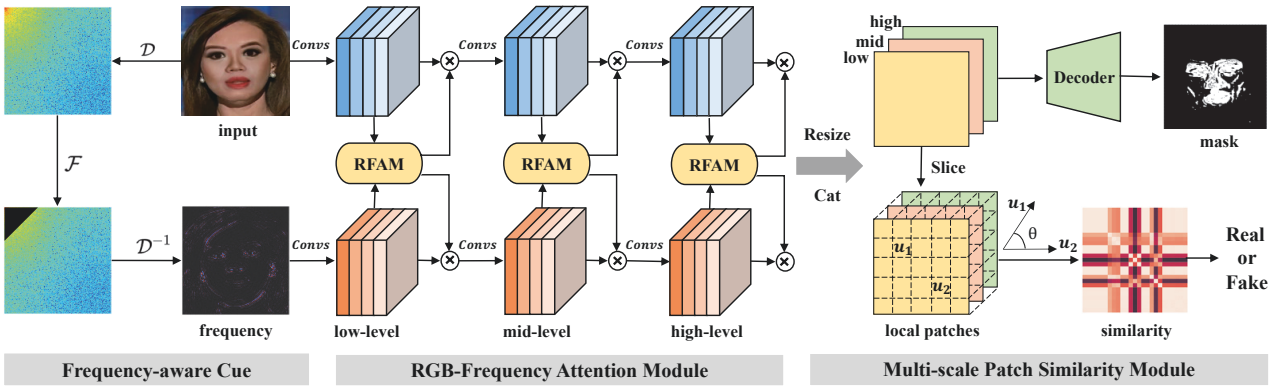


Figure 1: The framework of the proposed Local Relation Learning for Face Forgery Detection.

quency based on *Discrete Cosine Transform* (DCT), then a two-stream network with attention modules in different blocks is designed to fuse RGB and frequency features, improving the integrity of local representation.

The main contributions of this work are summarized as follows:

- We address face forgery detection via local relation learning, and propose a similarity pattern which is a generalized descriptor that can be used to effectively capture the forged trace.
- We combine RGB and frequency information based on the attention module to collaboratively learn comprehensive representation, which further improves the reliability of the similarity pattern.
- Extensive experiments and visualizations are presented to reveal the robustness and generalization of the proposed similarity pattern, which demonstrates the effectiveness of our method against the state-of-the-art competitors.

2 Related Works

Over the past several years, forgery creation has recently gained significant attention. With the complementary property of forgery creation and detection, face forgery detection also becomes an increasingly emerging research area. In this section, we briefly review prior image forgery methods including face forgery to which our method belongs.

Face forgery detection is mostly regarded as merely a binary (real or forgery) classification problem. With the tremendous success of deep learning (Chollet 2017; Tan and Le 2019; Lin et al. 2020), some works (Afchar et al. 2018; Nguyen, Yamagishi, and Echizen 2019; Rössler et al. 2019) adopt neural networks to automatically extract discriminative features for forgery detection. However, they are easy to overfit and lacks interpretability. Besides classification, there are methods focusing on localizing the manipulated region. Some works (Bappy et al. 2017; Nguyen et al. 2019; Salloum, Ren, and Kuo 2018) use multi-task learning to simultaneously classify the manipulated images and locate the manipulated region. Instead of simply using a multi-task learning strategy, Stehouwer *et al.* (Stehouwer et al. 2020) highlights the informative regions through an attention mechanism where the attention map is guided by the ground truth manipulation mask. Face X-ray (Li et al. 2020a) localizes the forged boundary in a self-supervised mechanism, achieving remarkable performance at high-quality data. Although these works have achieved considerable performance, they are easily affected by image compression or noise, leading to poor generalizability.

Recently, several attempts have been made to solve forgery detection using frequency cues. For example, Durall *et al.* (Durall et al. 2019) extracts frequency-domain information using DFT transform and averages the amplitudes of different frequency bands. Two-branch RN (Masi et al. 2020) amplifies multi-band frequencies using a Laplacian of Gaussian (LoG) as a bottleneck layer. F3 Net (Qian et al. 2020) takes advantage of frequency-aware decomposed image components and local frequency statistics to deeply mine the forged patterns. Although these methods extract features more comprehensively, they still cannot guarantee strong generalization. To learn more robust and generalized features, we address face forgery detection via local relation learning, which focuses on the relation of local regions.

3 Approach

The framework of the proposed approach is illustrated in Fig. 1. We will present the extraction strategy for frequency information in Sec. 3.1, and the proposed RFAM for co-learning RGB and frequency domain information in Sec.3.2. We also describe the MPSM for learning local region associations in Sec. 3.3. All the supervised loss functions are summarized in Sec. 3.4.

3.1 Frequency-aware Cue

As stated in (Zhang, Karaman, and Chang 2019), most existing face manipulated methods are based on GAN, where the up-sampling operation causes anomalies in the frequency statistical properties of the forged faces. To this end, we introduce frequency information to assist the network in mining the essential difference between real and forged regions.

Without loss of generality, let $\mathbf{x}_1 \in \mathbb{R}^{H\times W\times 3}$ denote the input RGB image, where H and W denote the height and

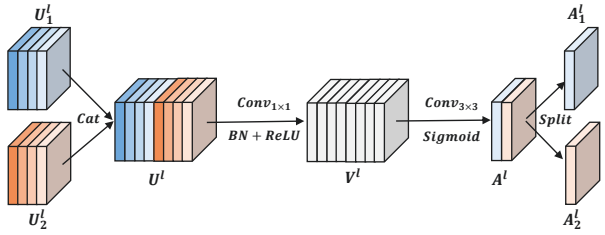


Figure 2: RGB-Frequency Attention Module.

width of image, respectively. We first transform $\mathbf{x}_1$ from RGB domain to frequency domain. That is:

$$\mathbf{x}_1^d = \mathcal{D}(\mathbf{x}_1), \tag{1}$$

where $\mathbf{x}_1^d \in \mathbb{R}^{H\times W\times 1}$, and $\mathcal{D}$ denotes the *Discrete Cosine Transform* (DCT), according to its wide applications in image processing, and its nice layout of the frequency distribution, *i.e.*, low frequency responses are placed in the top-left corner, and high frequency responses are located in the bottom-right corner.

We then suppress the image content by filtering out the low frequency information to amplify subtle artifacts at high frequencies. That is:

$$\mathbf{x}_1^f = \mathcal{F}(\mathbf{x}_1^d, \alpha), \tag{2}$$

where $\mathcal{F}$ denotes high pass filtering and α controls the low frequency component to be filtered out. Specifically, $\mathcal{F}$ sets the triangle with length α on the top-left corner of $\mathbf{x}_1^d$ to 0.

Since the frequency domain does not match the shift-invariance and local consistency owned by natural images, we invert $\mathbf{x}_1^f$ back into RGB color space via $\mathcal{D}^{-1}$ to obtain the desired representation at frequency domain. The entire process can be summarized as follows:

$$\mathbf{x}_2 = \mathcal{D}^{-1}(\mathcal{F}(\mathcal{D}(\mathbf{x}_1), \alpha)), \tag{3}$$

where $\mathbf{x}_2 \in \mathbb{R}^{H\times W\times 1}$. We will explain in detail how frequency domain information aids attention to subtle artifacts in the visualization sec. 4.4.

3.2 RGB-Frequency Attention Module

In this section, we develop a two-stream network whose input contains RGB data $\mathbf{x}_1$ and frequency data $\mathbf{x}_2$. For forged images, RGB information is useful for locating anomalous textures, while frequency information amplifies the subtle manipulated artifacts. Fusing them contributes to a more comprehensive feature representation. To take full advantage of this information, we design an RGB Frequency Attention Module (RFAM) that collaboratively fuse RGB and frequency information at different semantic layers, facilitating the learning of local region features.

As shown in Fig. 2, $\mathbf{U}_1^l \in \mathbb{R}^{H^l\times W^l\times C^l}$ and $\mathbf{U}_2^l \in \mathbb{R}^{H^l\times W^l\times C^l}$ represent the feature map of the frequency stream and the RGB stream at the l-th layer of network, respectively. To simplify, we let $l \in \{low, mid, high\}$ denote the different semantic layers of the network, and H^l, W^l, and C^l be the height, width, and channel of the feature map of the corresponding layer.

We first concatenate $\mathbf{U}_1^l$ and $\mathbf{U}_2^l$ in the channel dimension to get $\mathbf{U}^l \in \mathbb{R}^{H\times W\times 2C}$. Then $\mathbf{U}^l$ is through a 1×1 convolution with BatchNormalization (Ioffe and Szegedy 2015) and ReLU non-linearity to comprehensively fuse the RGB and frequency information, which is formulated as:

$$\mathbf{V}^l = \delta(\mathcal{B}(Conv_{1\times1}(\mathbf{U}^l))), \tag{4}$$

where $\mathbf{V}^l \in \mathbb{R}^{H\times W\times 2C}$, $\mathcal{B}$ denote BatchNormalization and δ is the ReLU function.

Furthermore, an attention map $\mathbf{A}^l \in \mathbb{R}^{H\times W\times 2}$ is created to highlight the regions of interest in $\mathbf{U}_1^l$ and $\mathbf{U}_2^l$, respectively. That is:

$$\mathbf{A}^l = \sigma(Conv_{3\times3}(\mathbf{V}^l)), \tag{5}$$

where $Conv_{3\times3}$ is a 3×3 convolution, σ denotes the *Sigmoid* function. Finally, we split $\mathbf{A}^l$ into $\mathbf{A}_1^l \in \mathbb{R}^{H^l\times W^l\times 1}$ and $A_2^l \in \mathbb{R}^{H^l\times W^l\times 1}$, which correspond to the attention map of RGB stream and frequency stream, respectively. The most significant spatial regions are highlighted via dot products between the feature map and attention map.

The proposed RFAM collaboratively fuses the forged cues under the RGB and frequency domains, facilitating the learning of the correlation between local regions.

3.3 Multi-scale Patch Similarity Module

Unlike existing methods that discriminate based on the global pooling of feature map, we propose a Multi-scale Patch Similarity Module (MPSM), which measures the similarity of local regions based on the learned RGB and frequency features.

Specifically, we first fuse the results from l-th layer of the two-stream network, as follows:

$$\widetilde{\mathbf{U}}^l = \mathbf{A}_1^l \odot \mathbf{U}_1^l + \mathbf{A}_2^l \odot \mathbf{U}_2^l, \tag{6}$$

where $\widetilde{\mathbf{U}}^l \in \mathbb{R}^{H^l\times W^l\times C^l}$ denotes the fused feature map of l-th layer. Through summing over the spatial dimension, we can directly merge the regions of interest in the RGB and frequency domains.

To address the problem of different forged region sizes, we introduce the multi-scale features, where high-resolution low-level features facilitate localization and high-level features rich in semantic information are used for identification. Specifically, we resize $\widetilde{\mathbf{U}}^{low}$ and $\widetilde{\mathbf{U}}^{mid}$ to the same size as $\widetilde{\mathbf{U}}^{high}$, and then concatenate them together to get the multi-scale features $\widetilde{\mathbf{U}} \in \mathbb{R}^{\widetilde{H}\times\widetilde{W}\times\widetilde{C}}$, where $\widetilde{H} = H^{high}$, $\widetilde{W} = W^{high}$, and $\widetilde{C} = \sum_{l\in\{low,mid,high\}} C^l$.

To construct associations between local regions, we partition $\widetilde{\mathbf{U}}$ spatially into $k\times k$ patches $\widetilde{\mathbf{u}}_i \in \mathbb{R}^{h\times w\times\widetilde{C}}$, where $i \in \{1, 2, \ldots, k^2\}$, $h = Ceil(\frac{\widetilde{H}}{k})$ and $w = Ceil(\frac{\widetilde{W}}{k})$. Then $\widetilde{\mathbf{u}}_i$ is flattened into a one-dimensional vector $\mathbf{u}_i \in \mathbb{R}^{hw\widetilde{C}}$, and the similarity between patch u_i and u_j is calculated based on cosine distance. That is:

$$\hat{\mathbf{s}}_{i,j} = \frac{\left\langle \frac{u_i}{\|u_i\|_2}, \frac{u_j}{\|u_j\|_2} \right\rangle + 1}{2}, \tag{7}$$

Methods	ACC (Raw)	AUC (Raw)	ACC (HQ)	AUC (HQ)	ACC (LQ)	AUC (LQ)
Steg.Features (Fridrich and Kodovský 2012)	97.63%	-	70.97%	-	55.98%	-
LD-CNN (Cozzolino, Poggi, and Verdoliva 2017)	98.57%	-	78.45%	-	58.69%	-
C-Conv (Bayar and Stamm 2016)	98.74%	-	82.97%	-	66.84%	-
CP-CNN (Rahmouni et al. 2017)	97.03%	-	79.08%	-	61.18%	-
MesoNet (Afchar et al. 2018)	95.23%	-	83.10%	-	70.47%	-
Xception (Rössler et al. 2019)	99.26%	-	95.73%	-	86.86%	-
Face X-ray (Li et al. 2020a)	-	-	-	87.40%	-	61.60%
Two-branch RN (Masi et al. 2020)	-	-	96.43%	88.70%	86.34%	86.59%
F3-Net (Qian et al. 2020)	**99.95%**	99.80%	97.52%	98.10%	90.43%	93.30%
Ours	99.87%	**99.92%**	**97.59%**	**99.46%**	**91.47%**	**95.21%**

Table 1: Quantitative results in terms of ACC and AUC on FaceForensics++ dataset with all quality settings, *i.e.*, low quality (LQ), high quality (HQ), and raw videos without compression (Raw).

where $\hat{\mathbf{s}}_{ij}$ ranges from 0 and 1, and $\hat{\mathbf{s}} \in \mathbb{R}^{k^2 \times k^2}$. A lower $\hat{\mathbf{s}}_{ij}$ indicates a larger discrepancy between patches, *i.e.*, one is the real and the other is forged, and vice versa.

We introduce the manipulated region mask and transform it into a kind of second-order supervision to guide the learning of similarity $\hat{\mathbf{s}}$. To construct the mask, we pair forged images with their corresponding source images, compute the absolute pixel-wise difference in the RGB channels, convert into grayscale, and then divide by 255 to produce a map in the range of [0, 1]. We empirically determine the threshold of 0.15 to obtain the desired binary mask $\mathbf{M}$.

Without loss of generality, let $\mathbf{M} \in \mathbb{R}^{H \times W}$ denote the mask, where the value of the real region is 0 and the value of the forged region is 1. That is:

$$\mathbf{M}_{ij} = \begin{cases} 0 & \text{if } \mathbf{x}_{ij} \text{ is real} \\ 1 & \text{if } \mathbf{x}_{ij} \text{ is forged} \end{cases}. \tag{8}$$

Subsequently, we divide $\mathbf{M}$ into $k \times k$ patches $\mathbf{m}_i \in \mathbb{R}^{h \times w}$, where $i \in \{1, 2, \ldots, k^2\}$, $h = \mathit{Ceil}(\frac{H}{k})$ and $w = \mathit{Ceil}(\frac{W}{k})$. Then the forged probability $\mathbf{p}_i \in [0, 1]$ for each patch $\mathbf{m}_i$ is obtained by averaging all the pixel values of $\mathbf{m}_i$. Finally, based on the Euclidean distance between the forged probabilities, we obtain the desired second-order supervision, *i.e.*, the relation of local regions. That is:

$$\mathbf{s}_{ij} = 1 - (\mathbf{p}_i - \mathbf{p}_j)^2, \tag{9}$$

where $\mathbf{s}_{ij}$ ranges from 0 and 1, and $\mathbf{s} \in \mathbb{R}^{k^2 \times k^2}$ guides the learning of local relation. Formally, we formulate the **Similarity Loss** function as follows:

$$\mathcal{L}_{sim} = \|\mathbf{s} - \hat{\mathbf{s}}\|_2. \tag{10}$$

3.4 Loss Functions

We have learned a second-order pattern based on the association of local regions, which is used to identify whether a face is forged or not. To this end, we flatten $\hat{\mathbf{s}}$ and successively pass it through the fully connected layer and *Sigmoid* function to obtain the final predicted probability $\hat{\mathbf{y}}$. And the **Cross-Entropy Loss** function is defined as:

$$\mathcal{L}_{\text{ce}} = -\left[\mathbf{y} \log \hat{\mathbf{y}} + (1 - \mathbf{y}) \log(1 - \hat{\mathbf{y}})\right], \tag{11}$$

where $\mathbf{y}$ is set to 1 if the face image has been manipulated, otherwise it is set to 0.

To compensate for the loss of edge information due to averaging operations in $\mathbf{m}_i$, we further introduce a decoder module to locate the specific forged region $\hat{\mathbf{M}} \in \mathbb{R}^{H \times W}$, and formulate the **Segmentation Loss** function as:

$$\mathcal{L}_{\text{seg}} = \sum_{i,j} -\left[\mathbf{M}_{ij} \log \hat{\mathbf{M}}_{ij} + (1 - \mathbf{M}_{ij}) \log\left(1 - \hat{\mathbf{M}}_{ij}\right)\right]. \tag{12}$$

The total loss functions of the proposed framework are described as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \lambda_1 \mathcal{L}_{\text{sim}} + \lambda_2 \mathcal{L}_{\text{seg}}, \tag{13}$$

where λ_1 and λ_2 are the hyper-parameters used to balance these loss functions. The parameters of network are updated via back-propagation.

Methods	DF	FF	FS	NT
Steg.Features	67.00%	48.00%	49.00%	56.00%
LD-CNN	75.00%	56.00%	51.00%	62.00%
C-Conv	87.00%	82.00%	74.00%	74.00%
CP-CNN	80.00%	62.00%	59.00%	59.00%
MesoNet	90.00%	83.00%	83.00%	75.00%
Xception	96.01%	93.29%	94.71%	79.14%
F3-Net	97.97%	95.32%	96.53%	83.32%
Ours	**98.84%**	**95.53%**	**97.53%**	**89.31%**

Table 2: Quantitative results (ACC) on FaceForensics++ (LQ) with four manipulation methods, *i.e.* DeepFakes (DF), Face2Face (FF), FaceSwap (FS) and NeuralTextures (NT).

Methods	Training dataset	Test datset					
		Celeb-DF		DFDC		DFD	
		AUC	EER	AUC	EER	AUC	EER
Xception	FF++	36.19	59.64	48.98	50.45	87.86	21.49
Ours	FF++	**78.26**	**29.67**	**76.53**	**32.41**	**89.24**	**20.32**

Table 3: Benchmark results in terms of AUC and EER for our framework and Xception (Rössler et al. 2019) on unseen datasets.

RGB	Freq	RFAM	MPSM	ACC	AUC
√				89.15%	93.03%
√	√			89.99%	93.13%
√	√	√		90.05%	93.78%
√			√	91.06%	94.96%
√	√	√	√	**91.47%**	**95.21%**

Table 4: Abalation study on the influence of different model components on FaceForensics++ (LQ) dataset.

4 Experiments

In this section, we first experimentally evaluate the effectiveness of the proposed algorithm against state-of-the-art techniques and investigate its robustness under unseen manipulation methods in Sec. 4.2. Subsequently, we conduct an ablation study to explore the influence of proposed components in Sec. 4.3. Finally, we demonstrate the interpretability of our approach through visualization analysis in Sec. 4.4.

4.1 Experimental Setup

Datasets. Following the convention, we adopt the widely-used benchmark dataset FaceForensics++ (FF++) (Rössler et al. 2019) for training. FF++ is a face forgery detection dataset consisting of 1000 original videos with real faces, in which 720 videos are used for training, 140 videos are reserved for validation and 140 videos for testing. Then each video is forged by four state-of-the-art face manipulation methods, *i.e.*, DeepFakes (DF) (Tora 2018), Face2Face (F2F) (Thies et al. 2019), FaceSwap (FS) (Kowalski 2018), and NeuralTextures (NT) (Thies, Zollhöfer, and Nießner 2019). Output videos are generated with different quality levels to create a realistic setting for manipulated videos, *i.e.*, raw, high quality (HQ) and low quality (LQ), respectively. To evaluate the robustness of our method, we also conduct experiments on the recent proposed large-scale face manipulated dataset, *i.e.*, Deepfake Detection Challenge (DFDC) (Dolhansky et al. 2019), Celeb-DF (Li et al. 2020b) and DeepfakeDetection[1] (DFD).

Implementation. We implement the proposed framework via open-source PyTorch (Paszke et al. 2017). For the frequency-aware cue, the α in Equ. 3 is empirically set to 0.33. And the number of patches k is set to 5. To enhance the learning of local region relations, we set λ_1 and λ_2 in Equ. 13 to 10 and 1, respectively. Following FaceForensics++ (Rössler et al. 2019), we resize the input image to 299×299, and train the network using Adam optimizer (Kingma and Ba 2015) with a learning rate of 2e-4, a batch size of 32, betas of 0.9 and 0.999, and weight decay equal to 1e-5. The total number of training epochs is set to 50, and the learning rate is reduced to half every 10 epochs.

[1]https://ai.googleblog.com/2019/09/contributing-data-to-deepfake- detection.html

4.2 Results and Discussions

In this section, we apply the Accuracy score (ACC) and Area Under the Receiver Operating Characteristic Curve (AUC) as our evaluation metrics. For a fair comparison, the results of all comparison methods are obtained from their paper.

We first evaluate the model performance under different quality settings (Raw, HQ and LQ) and the comparison with the state-of-the-art is showed in Tab. 1. We can observe that: 1) Our proposed method consistently outperforms all compared opponents by a considerable margin. For example, compared with the state-of-the-art F3 Net, the AUC of our method exceeds it by 0.12%, 1.36%, and 1.91% at all the three quality settings, and this performance gain is also obtained under ACC. Different from F3 Net which only utilizes the frequency information, our model combines both RGB and frequency information for a more comprehensively feature presentation, so that all kinds of artifacts of the forged face can be captured. 2) Compared with Two-branch RN which also takes both RGB and frequency information into consideration, our method demonstrates superior ACC and AUC performance. This proves the effectiveness of our proposed RFAM. 3) Besides, our method has a significantly higher AUC on LQ than Face X-ray by 33.61%. To explain, Face X-ray relies on blending boundary that is susceptible to noise, whereas our method learns a content-independent local pattern through the proposed MPSM and is therefore more robust against disturbances including but not limited to image noise. This is important as the face forgery detection on low quality images is most challenging and measures the model's generalization ability to a large extent.

Furthermore, we evaluate the proposed model on four different face manipulation methods listed in the FF++ dataset. In the experiment setting, we follow the common practice by only using low-quality (LQ) videos in each manipulation method for training and test. The results are shown in Tab. 2. It can be observed that our method consistently outperforms all comparisons on the four manipulation methods. In particular, on the most challenging NeuralTextures (NT) manipulation method which does not produce noticeable forged artifacts, our model exceeds the state-of-the-art F3 Net by 5.99% under the ACC metrics, which further illustrates the

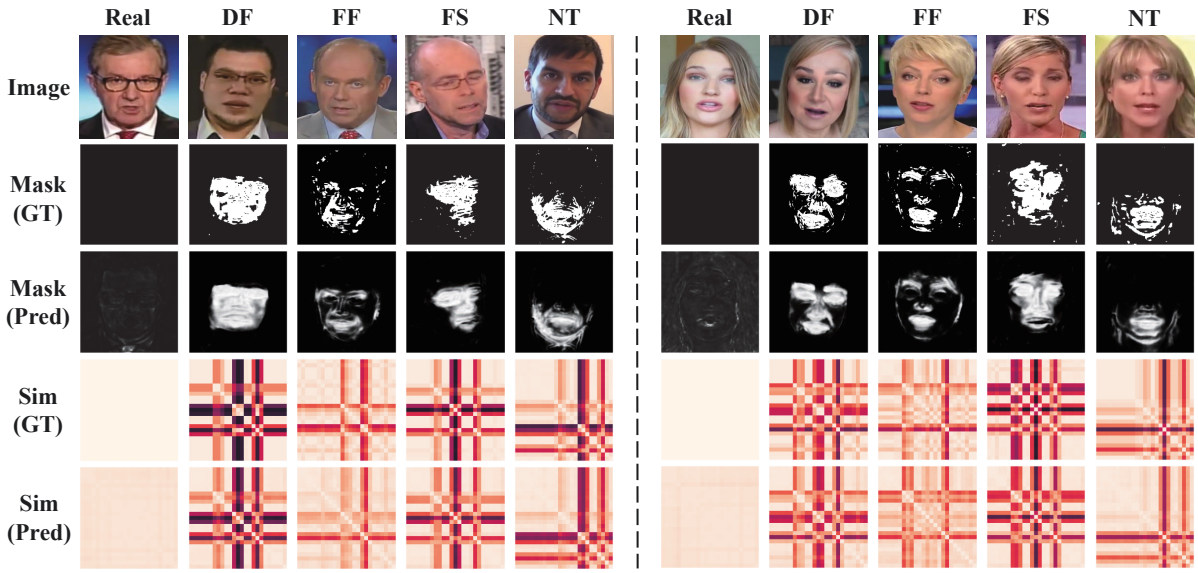


Figure 3: Visual examples of mask and local similarity patterns on various types of faces, *i.e.*, Real, Deepfakes (DF), Face2Face (FF), FaceSwap (FS) and NeuralTextures (NT).

effectiveness of our proposed local relation learning.

To demonstrate the generalization capabilities of our proposed model, we perform an inter-test by training on the FF++ dataset while testing on Celeb-DF, DFDC and DFD datasets. Since Face X-ray introduces an additional dataset *BI* in their experiment setting, we only include the results of Xception for a fair comparison. As shown in Tab. 3, our method significantly outperforms Xception on all unseen datasets. The gain mainly benefits from that our method focuses on the intrinsic differences between real and forged regions, which are commonly present under all manipulated faces with various manipulation techniques. And our model carefully designs the similarity pattern in MPSM to describe this intrinsic difference which improves the model's generalizability to unknown scenarios.

4.3 Ablation Study

To evaluate the effectiveness of the components of our proposed model, we develop the following variants: 1) the baseline model which contains only RGB input and is supervised by classification (Equ. 11) and segmentation (Equ. 12), 2) our method w/o RFAM and MPSM, which directly concatenate the final feature maps of two-stream network, 3) our method w/o MPSM, 4) our method w/o frequency and RFAM, where the input of MPSM is only RGB stream.

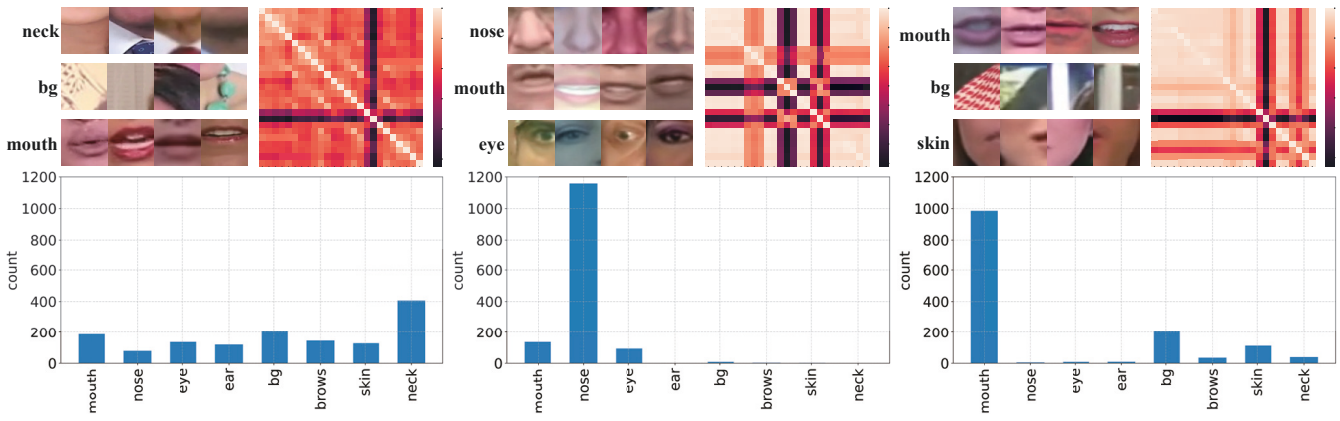
The quantitative results are listed in Tab. 4. By comparing variant 1 and variant 2, the introduction of frequency information consistently improves the ACC and AUC. The performance is further improved by fusing the two-stream information through the proposed RFAM. It is worth noting that by using only MPSM, the ACC and AUC are substantially increased to 91.06% and 94.96%, which fully demonstrates the efficiency of the local relation features. Combining all the proposed modules, our method achieves the best performance, 91.47% and 95.21% for ACC and AUC.

4.4 Visualization

Visual examples. Our framework makes predictions about the forgery based on the local similarity. The visual examples of mask and local similarity patterns on various types of faces are shown in Fig. 3. It can be seen that the predicted similarity and mask are the same as the ground-truth, which well captures the local forged regions generated by different face manipulated algorithms. For real faces, the features are similar between local regions, while forged faces show different similarity patterns depending on the specific manipulated regions. Therefore, such discrepancy can be effectively served as a basis of classification.

Interpretability of local relation. As shown in the Fig. 4, we statistically estimate the similarity patterns, and histograms of the most predictive patches on various types of faces, *i.e.*, Real, Deepfakes and NeuralTextures. Specifically, for each face type, we randomly sample 1000 images and average the similarity of local regions extracted by our method to obtain the corresponding similarity pattern (the top-right in Fig. 4). We can observe that the similarity of real faces tend to be consistent across regions, while forged faces show specific patterns due to different manipulated regions, which indicates the local relations is a generalized pattern for the Face Forgery Detection task. Besides, we select the most predictive region of each image based on the activation of the final feature map. For real face, the attention areas are evenly distributed, while fake face is mostly concentrated in forged regions such as the nose or mouth, which further proves the effectiveness and interpretability of our method.

Complementarity of RGB and frequency information. Fig. 5 presents the visualization of feature maps extracted by baseline and our method. We can observe that: 1) The baseline has a higher response in the middle region of any



(a) Real (b) Deepfakes (c) NeuralTextures

Figure 4: The local similarity patterns and histograms of the most predictive patches on various types of faces.

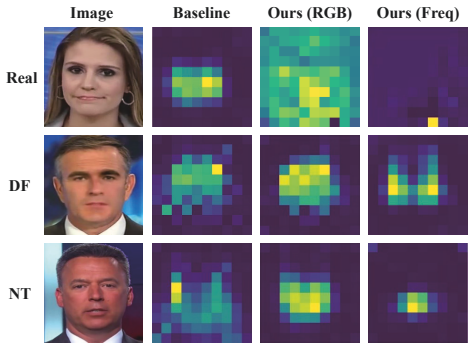


Figure 5: The feature maps of various types of faces extracted by baseline and our method, where the third and fourth columns are the feature map corresponding to RGB and frequecy, respectively.

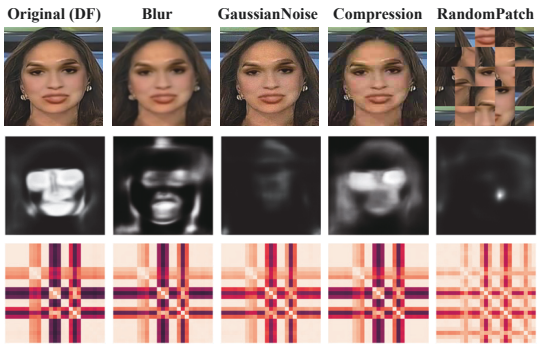


Figure 6: The visualization of the baseline (second row) and our method (third row) under different noises.

type of face, while our method exhibits a different focus between real and forgery. For example, in real faces, all regions of feature map have almost the same activation values, while in forged faces the activation values are larger only in the manipulated regions. This is because our method imposes constraints on local features to ensure that the network learns a content-independent feature and focuses more on artifacts. 2) The frequency and RGB information of the forged faces are located in different regions, *i.e.*, the former captures subtle forged boundary, while the latter shows higher response in facial areas such as eyes and mouth. Through the proposed RFAM, the information can complement each other to further facilitate the learning of local features. 3) Finally, since the real face has not been manipulated, there is no anomalous response in the feature map of frequency stream.

Robustness toward noises. Fig. 6 shows the visualization of the baseline and our method under different noises. Specifically, we apply different augmentations to the forged faces, *i.e.*, blur, gaussian noise, compression, and random patch. From Fig. 6, we can observe that the baseline has a significant error in the predicted masks under noises, especially under the gaussian noise and random patch. This is because the baseline classifies based on global information and is sensitive to unseen patterns. Conversely, the local similarity pattern of our method remains consistent under different noises. Even if the face structure is broken, *i.e.*, the random patch, the result is still robust.

5 Conclusion

In this paper, we introduce a new perspective for face forgery detection that models the relation of local regions. A novel architecture based on RGB-Frequency Attention Module is proposed with Multi-scale Patch Similarity supervision, which both considers RGB and frequency information and collaboratively learns comprehensive local relations. And the relations are further used for forgery detection. Extensive experiments and visualization demonstrate the robustness and generalizability of the proposed method on widely-used face forgery detection datasets.

Acknowledgements

This work is supported by the National Science Fund for Distinguished Young (No.62025603), the National Natural Science Foundation of China (No.U1705262, No.62072386, No.62072387, No.62072389, No.62002305, No.61772443, No.61802324 and No.61702136) and Guangdong Basic and Applied Basic Research Foundation (No.2019B1515120049).

References

Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* 1–7.

Bappy, J. H.; Roy-Chowdhury, A.; Bunk, J.; Nataraj, L.; and Manjunath, B. S. 2017. Exploiting Spatial Structure for Localizing Manipulated Image Regions. *2017 IEEE International Conference on Computer Vision (ICCV)* 4980–4989.

Bayar, B.; and Stamm, M. C. 2016. A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer. In *IH&MMSec '16*.

Chollet, F. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1800–1807.

Cozzolino, D.; Poggi, G.; and Verdoliva, L. 2017. Recasting Residual-based Local Descriptors as Convolutional Neural Networks: an Application to Image Forgery Detection. *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security* .

Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; and Canton-Ferrer, C. 2019. The Deepfake Detection Challenge (DFDC) Preview Dataset. *ArXiv* abs/1910.08854.

Durall, R.; Keuper, M.; Pfreundt, F.; and Keuper, J. 2019. Unmasking DeepFakes with simple Features. *ArXiv* abs/1911.00686.

Fridrich, J.; and Kodovský, J. 2012. Rich Models for Steganalysis of Digital Images. *IEEE Transactions on Information Forensics and Security* 7: 868–882.

Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv* abs/1502.03167.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980.

Kowalski, M. 2018. FaceSwap. https://github.com/marekkowalski/faceswap. Accessed: 2020-08-01.

Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; and Guo, B. 2020a. Face X-Ray for More General Face Forgery Detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 5000–5009.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020b. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 3204–3213.

Lin, M.; Ji, R.; Wang, Y.; Zhang, Y.; Zhang, B.; Tian, Y.; and Shao, L. 2020. HRank: Filter Pruning Using High-Rank Feature Map. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 1526–1535.

Liu, Z.; Qi, X.; Jia, J.; and Torr, P. 2020. Global Texture Enhancement for Fake Face Detection in the Wild. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 8057–8066.

Masi, I.; Killekar, A.; Mascarenhas, R. M.; Gurudatt, S. P.; and Abd-Almageed, W. 2020. Two-branch Recurrent Network for Isolating Deepfakes in Videos.

Nguyen, H. H.; Fang, F.; Yamagishi, J.; and Echizen, I. 2019. Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos. *ArXiv* abs/1906.06876.

Nguyen, H. H.; Yamagishi, J.; and Echizen, I. 2019. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2307–2311.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; Devito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch. In *NeurIPS Autodiff Workshop*.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues. *ArXiv* abs/2007.09355.

Rahmouni, N.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2017. Distinguishing computer graphics from natural images using convolution neural networks. *2017 IEEE Workshop on Information Forensics and Security (WIFS)* 1–6.

Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* 1–11.

Salloum, R.; Ren, Y.; and Kuo, C.-C. J. 2018. Image Splicing Localization using a Multi-task Fully Convolutional Network (MFCN). *ArXiv* abs/1709.02016.

Stehouwer, J.; Dang, H.; Liu, F.; Liu, X.; and Jain, A. 2020. On the Detection of Digital Face Manipulation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 5780–5789.

Tan, M.; and Le, Q. V. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *ICML*.

Thies, J.; Zollhöfer, M.; and Nießner, M. 2019. Deferred Neural Rendering: Image Synthesis using Neural Textures. *arXiv: Computer Vision and Pattern Recognition* .

Thies, J.; Zollhöfer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2019. Face2Face: real-time face capture and reenactment of RGB videos. *Commun. ACM* 62: 96–104.

Tora, M. 2018. Deepfakes. https://github.com/deepfakes/faceswap/tree/v2.0.0. Accessed 2020-08-01.

Zhang, X.; Karaman, S.; and Chang, S.-F. 2019. Detecting and Simulating Artifacts in GAN Fake Images. *2019 IEEE International Workshop on Information Forensics and Security (WIFS)* 1–6.