Exposing Face Forgery Clues via Retinex-based Image Enhancement

Han Chen[1−4[0000−0002−9439−9133]], Yuzhen Lin[1−4[0000−0001−7788−2054]], and Bin Li[1−4[0000−0002−2613−5451]⋆]

[1] Guangdong Key Laboratory of Intelligent Information Processing
[2] Shenzhen Key Laboratory of Media Security
[3] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[4] Shenzhen University, Shenzhen 518060, China.
{2016130205, linyuzhen2020}@email.szu.edu.cn; libin@szu.edu.cn

Abstract. Public concerns about deepfake face forgery are continually rising in recent years. Existing deepfake detection approaches typically use convolutional neural networks (CNNs) to mine subtle artifacts under high-quality forged faces. However, most CNN-based deepfake detectors tend to over-fit the content-specific color textures, and thus fail to generalize across different data sources, forgery methods, and/or post-processing operations. It motivates us to develop a method to expose the subtle forgery clues in RGB space. Herein, we propose to utilize multi-scale retinex-based enhancement of RGB space and compose a novel modality, named MSR, to complementary capture the forgery traces. To take full advantage of the MSR information, we propose a two-stream network combined with salience-guided attention and feature re-weighted interaction modules. The salience-guided attention module guides the RGB feature extractor to concentrate more on forgery traces from an MSR perspective. The feature re-weighted interaction module implicitly learns the correlation between the two complementary modalities to promote feature learning for each other. Comprehensive experiments on several benchmarks show that our method outperforms the state-of-the-art face forgery detection methods in detecting severely compressed deepfakes. Besides, our method also shows superior performances on cross-datasets evaluation.

Keywords: Deepfake detection · Multi-scale retinex · Generalization.

1 Introduction

Deepfake techniques [1–3] refer to a series of deep learning-based facial forgery techniques that can swap or reenact the face of one person in a video to another. While deepfake technology is very popular in the entertainment and film industries, it is also notorious for its unethical applications that can threaten

⋆ Han Chen and Yuzhen Lin contributed equally to this work. Bin Li is the corresponding author.

politics, economics, and personal privacy. Over the past few years, a large number of deepfake videos (called deepfakes) uploaded to the Internet with potential harms have been reported. Accordingly, the countermeasures being desired to identify deepfakes become an urgent topic in social security.

To prevent malicious deepfake media from threatening the credibility of human society, deepfake detection (i.e, face forgery detection) is becoming an urgent topic that has attracted widespread attention. Early works leverage handcrated features(e.g., eyes-blinking[4] or visual artifacts [5]) or semantic features extracted by universal CNNs [6, 7] to identify real and fake images/videos. These methods achieve promising performance when the training and testing data are sampled from the same distribution. However, deepfakes in the real-world are different from those contained in a training set in terms of the data source, forgery method, and post-processing. Due to these mismatched domain gaps, most deepfake detection methods suffer from severe performance drops in practical applications. Therefore, generalization capability is one of the major concerns for existing deepfake detection systems.

In general, there are two typical manners for addressing the generalizing problem have been explored. On the one hand, some works train the deepfake detector with synthetic data that artificially simulates the forgery traces (e.g., visual resolution[8] or blending boundary [9]), which encourages models to learn generic features for face forgery detection. However, these methods suffer severe performance drop when facing post-process distortions (e.g. video compression). On the other hand, some works utilize two-stream networks that introduce information from other domains, such as DCT[10] and SRM[11] features. These methods either simply concatenate RGB and other features at the end of the network or fuse them with at a shallow layer, which rarely considers the relation and interaction between the additional information and regular color textures. This makes it difficult for them to fully utilize the additional information.

As pointed in [12, 13], the poor generalization in CNN-based deepfake detection can attribute to the fact that deep CNN models tend to easily capture the content-specific texture patterns in the RGB space. Thus, designing a deepfake detector with good generalization should consider suppressing the content-specific color textures and exposing discrepancies between forged and real regions. With this simple but powerful insight, herein, we utilize a multi-scale retinex enhancement inspired by the illumination-reflection model [14, 15] and compose a novel modality, named MSR, to complementary capture the forgery traces. To take full advantage of the MSR information, we propose a two-stream network combined with salience-guided attention and feature re-weighted interaction modules. The salience-guided attention module guides the RGB feature extractor to concentrate more on forgery traces from the MSR perspective at multi-scale level. The feature re-weighting module leverages the correlation between the two complementary modalities to promote feature learning for each other. Extensive experiments demonstrate that the proposed framework achieves consistent performance improvement compared to state-of-the-art methods. The main contributions of our work are summarized as follows.

- To expose the forgery traces, we perform retinex-based enhancement and propose a multi-scale retinex (MSR) feature as the complementary modality for RGB images.
- To take full advantage of MSR information, we devise a novel two-stream framework to collaboratively learn comprehensive representation. We design two functional modules to promote the correlation and interaction between the MSR and RGB components, i.e., the feature re-weighted interaction module and salience-guided attention module.
- Comprehensive experiments are presented to reveal the robustness and generalization of our proposed method compared to several state-of-the-art competitors.

2 Related Works

2.1 Face Forgery Detection

The past four years have witnessed a wide variety of methods proposed for defending against the malicious usage of deepfakes. Early works focus on handcrafted features such as eyes-blinking [4] and visual artifacts [5]. Due to the tremendous success of deep learning, convolutional neural networks (CNNs) [6, 7, 16] is widely used to deepfake detection task and achieved better performance. As have been criticized, most of the methods suffer from severe over-fitting to the training data and cannot be effectively used in many practical scenarios. There are methods trying to cope with the over-fitting issue of deepfake detectors. One of the effective approaches to address this problem is training models with synthetic data. For instance, Li *et al.* [8] noticed the quality gap between GAN-synthesized faces and natural faces, and proposed FWA (Face Warping Artifacts) to simulate the fake images by blurring the facial regions of real images. BI (Blending Image) [9] and I2G (Inconsistency Image Generator) [17] were introduced to generate blended faces which can simulate the blending artifacts from of some pristine image pairs with similar facial landmarks. In addition, it is also a common idea to use multi-modality (e.g., frequency domain) learning framework and auxiliary supervisions (e.g., forgery mask) to further mine the heuristic forgery clues and improve the robustness of the model. Qian *et al.* first employed the global and local frequency information for deepfake detection task. Luo *et al.* [11] employed SRM filter that extract the high-frequency noise to guide RGB features. Wang *et al.* [18] amplified implicit local discrepancies from RGB and frequncy domain with a novel multi-modal contrastive learning framework. Kong *et al.* [19] introduced PRNU noise information to guide the RGB features, and proposed a novel two-stream network for not only identifying deepfakes but also localizing the forgery regions.

In this work, we utilize a novel MSR modality based on the Retinex theory [20] that exposes the forgery traces in RGB space. Furthermore, we devise a novel two-stream network that combines the MSR and RGB information to collaboratively learn comprehensive representation for detecting deepfakes.

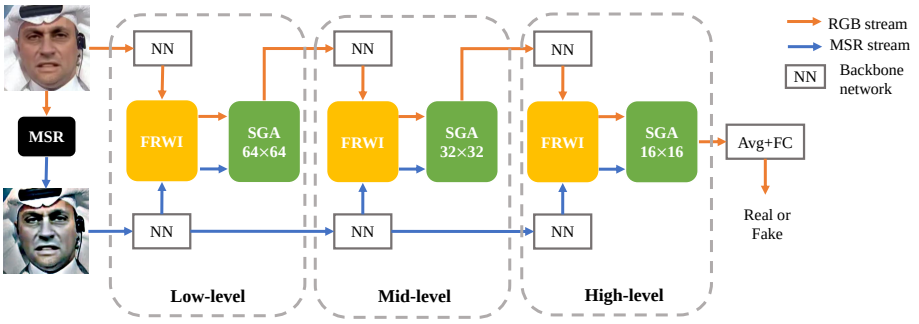


Fig. 1. Overall framework of our proposed method.

2.2 Retinex-based Methods

Retinex theory [20] models the color perception of human vision on natural scenes. It assumes that the observed images can be decomposed into two components, i.e., reflectance and illumination, which can be mathematically formulated as:

$$S(x,y) = R(x,y) \otimes I(x,y) \tag{1}$$

where x and y are image pixel coordinates. $R(x,y)$ represents reflectance, $I(x,y)$ represents illumination and $\otimes$ represents element-wise multiplication.

Retinex-based methods are widely accepted among image enhancement methodologies [14, 21] due to their robustness. Besides, it can also be viewed as a fundamental theory for the intrinsic image decomposition problem, which aims at disentangling an image into two independent components, such as the structure and texture [22].

As for image forensics task, Chen *et al.* [15] proposed to use retinex-based information for face anti-spoofing task and achieve great generalization performances.

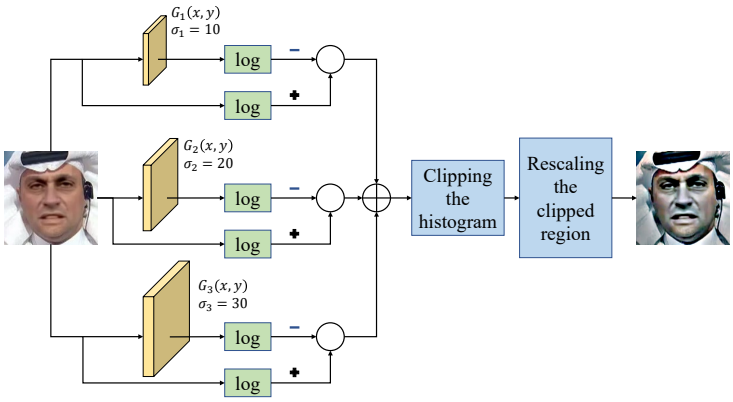
In this work, we apply the retinex-based information as the complementary of RGB modality, which aims to improve the generalization performance of deepfake detection.

3 Methodology

3.1 Overall Framework

In this work, we propose a novel two-stream framework that utilizes MSR information for face forgery detection. Specifically, the original RGB images are first converted to MSR images. Following that, to learn comprehensive feature representation, the RGB and MSR information are integrated at three semantic levels (low, mid and high) through a two-stream network combined with salience-guided attention and feature re-weighted interaction modules.

(a) MSR modality can attenuate content-specific colors and enhance forgery clues



(b) Pipeline of MSR extraction in this work

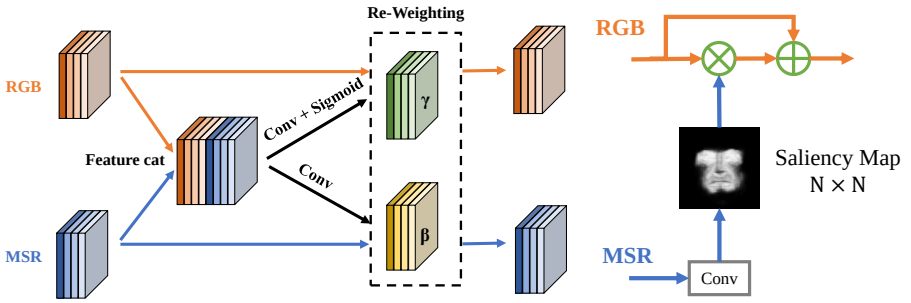
Fig. 2. Illustration of multi-scale retinex (MSR) enhancement. (a) Inspired by [15], we adopt a MSR-based algorithm for deepfake detection task. (b) Red boxes mark blending traces that are hard to recognize in the RGB space but distinctive in the MSR space.

According to the resolutions of output feature maps, we abstractly divide the whole network into three semantic layers. As for CNN, the low-resultion feature maps at the end of the network contain high semantic information, and vice versa. Thus, we denote these semantic layers as $l \in \{low, mid, high\}$ for simplicity. H^l, W^l, C^l are the height, width, and channel of the feature map of the corresponding layer. Formally, we define the feature map from the RGB and MSR stream at l-th layers of network as $F_R^l \in \mathbb{R}^{H^l \times W^l \times C^l}$ and $F_M^l \in \mathbb{R}^{H^l \times W^l \times C^l}$, respectively. The overall framework of our proposed approach is illustrated in Figure 1, and several components are elaborated in more detail as follows.

3.2 Multi-Scale Retinex Extraction

For the retinex theory, Eq.(1) is usually transformed into the logarithmic domain as:

$$\log[S(x,y)] = \log[R(x,y)] + \log[I(x,y)] \tag{2}$$



(a) Feature Re-weighted Interaction (b) Salience-Guided Attention

Fig. 3. The pipeline of our proposed method. We design a novel two-stream architecture, which aims to collaboratively learn comprehensive representations from the RGB and MSR information with Feature Re-weighted Interaction and Salience-Guided Attention modules.

We respectively represent $\log[S(x,y)]$ and $\log[R(x,y)]$ by $s(x,y)$ and $r(x,y)$ for convenience. Summarizing the previous work, the illumination image can be generated from the source image using the center/surround etinex.

$$r(x,y) = s(x,y) - \log[S(x,y) * F(x,y)] \tag{3}$$

where $F(x,y)$ denotes the surround function, and $*$ is the convolution operation, and it is the so called Single Scale Retinex (SSR) model. To overcome the highly dependency on the parameter of $F(x,y)$, Jobson *et al.* [14] proposed a Multi-scale Retinex (MSR) model, which weights the outputs of several SSRs with different $F(x,y)$.

As shown in Figure 2(a), the MSR modality can attenuate content-specific colors and enhance forgery clues. Thus, we utilize the multi-scale retinex enhancement and compose a novel modality, named MSR, for deepfake detection task. As shown in Figure 2(b), the pipeline of MSR extraction in this work can be formulated as:

$$r_{MSR}(x,y) = \sum_{i=1}^{3} w_i \left\{\log[S(x,y)] - \log\left[S(x,y) * G_i(x,y)\right]\right\} \tag{4}$$

where $G_i(x,y)$ denotes three Gaussian filters with $\sigma_i = 10, 20, 30$. We also add color restoration operations for eliminating color shifts after employing the above MSR pipeline.

3.3 Feature Re-weighted Interaction

To collaboratively align and integrate the feature maps from two domains, we proposed a novel Feature Re-weighted Interaction (FRWI) module inspired by

the mechanisms of SPADE [23]. The computation block of FRWI is described in Fig. 3(a).

Firstly, we concatenate feature maps from RGB and MSR streams with $F^l_{Concat} = \mathcal{C}\left(F^l_R, F^l_M\right)$, where $\mathcal{C}\left(\cdot,\cdot\right)$ being feature concatenation in channel dimension. In order to align and aggregate the feature maps from two domains, we respectively generate the weight γ^l and bias β^l by utilizing F^l_{Concat} as:

$$\gamma^l = \delta\left(f_1\left(F^l_{Concat}\right)\right) \tag{5}$$

$$\beta^l = f_2\left(F^l_{Concat}\right) \tag{6}$$

Specifically, γ^l is learned through a 3×3 convolution layer (denoted as f_1) and a sigmoid activation (denoted as $\delta\left(\cdot\right)$) while β^l learned through another 3×3 convolutional layer (denoted as f_2). The outputs of the FRWI module can be formulated as:

$$\widetilde{F}^l_R = \gamma^l \otimes F^l_R + \beta^l \tag{7}$$

$$\widetilde{F}^l_M = \gamma^l \otimes F^l_M + \beta^l \tag{8}$$

where the $\widetilde{F}^l_R$ and $\widetilde{F}^l_M$ represents the aligned feature maps of RGB and MSR streams, respectively.

3.4 Salience-Guided Attention

Utilizing the forgery mask as auxiliary supervision is a universal trick to improve the performance of face forgery detection. Inspired by this, we further adopt the forgery mask as the saliency map to highlight the manipulation traces. In particular, we introduce the spatial attention mechanism and design a Salience-Guided Attention (SGA) which guides a feature learning in the RGB modality with MSR information at different semantic layer. The computation block of SGA is described in Fig. 3(b).

Specifically, we predict the saliency map (denoted as $\hat{\mathcal{M}}^l$) of l-th semantic layer as:

$$\hat{\mathcal{M}}^l = \delta\left(f^l_3\left(\widetilde{F}^l_M\right)\right) \tag{9}$$

where f^l_3 represents a 1×1 convolution layer to transform the channels of $\widetilde{F}^l_M$ with 1. We respectively set $N = 64, 32, 16$ for predicting the $\hat{\mathcal{M}}^l$ in the low, mid and high level layer.

The output of SGA module is formulated as:

$$\widetilde{F}^l_{out} = \widetilde{F}^l_R + \hat{\mathcal{M}}^l \otimes \widetilde{F}^l_R \tag{10}$$

3.5 Training Details and Loss Functions

We employ the Efficient-B4 (EN-b4) as the backbone of our work. In order to capture more artifacts at higher resolutions, we change the stride of the first convolution layer at the backbone model from 2 to 1. The whole end-to-end

training process involves the supervision of binary classification and salience prediction task, and the overall loss function consists of two components:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{SM} \tag{11}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{SM}$ represents the binary cross-entropy loss and saliency map loss, respectively. λ is the balance weight.

Specifically, the cross-entropy binary classification loss $\mathcal{L}_{cls}$ is formulated as:

$$\mathcal{L}_{cls} = \mathbf{y}_t \log \mathbf{y}_p + (1 - \mathbf{y}_t) \log (1 - \mathbf{y}_p) \tag{12}$$

where $\mathbf{y}_t$ and $\mathbf{y}_p$ represents ground-truth label and the prediction logits, respectively.

The saliency map loss $\mathcal{L}_{SM}$ consits of the l_2 loss in of three semantic layers, which can be formulated as:

$$\mathcal{L}_{SM} = \sum_{l=1}^{3} \frac{1}{\Omega\left(\mathcal{M}^l\right)} \left\| \mathcal{M}^l - \hat{\mathcal{M}}^l \right\|_2 \tag{13}$$

where $\Omega(\cdot)$ repents the total number of elements. $\mathcal{M}^l$ is the ground truth forgery mask of the l-th semantic layer. We employ DSSIM [24] algorithm, which compute the paired face manipulation images and their corresponding source face pristine images with threshold, to get the original forgery mask $\mathcal{M}_{gt}$ with size of 256×256. Besides, we use bi-linear interpolation to down-sample $\mathcal{M}_{gt}$ by $\{4\times, 8\times, 16\times\}$, respectively obtain ground-truth saliency maps for the low, mid and high semantic layer (i.e., $\mathcal{M}^l, i = 1, 2, 3$).

4 Experiments

4.1 Experimental Setup

Datasets and Pre-processing In this paper, we mainly conducted experiments on the challenging *FaceForensics++ (FF++)* [7] dataset. FF++ contains 1000 Pristine (PT) videos (i.e., the real sample) and 4000 fake videos forged by five manipulation methods, i.e., Deepfakes (DF), Face2Face (F2F) [25], FaceSwap (FS), NeuralTextures (NT) [26]. Besides, FF++ provides three quality levels controlled by the constant rate quantization parameter (QP) in compression for these videos: raw (QP=0), HQ (high-quality, QP=23) and LQ (low-quality, QP=40). Considering the deployment in real-world application scenarios, we conduct our experiments on both HQ videos and LQ videos. The samples were split into disjoint training, validation, and testing sets at the video level follows the official protocol [7].

As for pre-processing, we utilized MTCNN [27] to detect and crop the face regions (enlarged by a factor of 1.3) from each video frame, and resized the them to 256×256 as the input images.

Table 1. Detection performances (%) on FF++ dataset. HQ and LQ denote the high-quality and low-quality data. * indicate the model is trained by us implementing the official code. The best results are in bold.

Methods	HQ		LQ	
	ACC	AUC	ACC	AUC
MesoNet[6]	83.10	-	70.47	-
Xception[7]	95.73	-	81.00	-
PRRNet[29]	96.15	-	86.13	-
SPSL[30]	91.50	95.32	81.57	82.82
MTA[31]	97.60	99.29	**88.69**	90.40
MC-LCR[18]	97.89	99.65	88.07	90.28
D&L[19]	**98.40**	**99.77**	84.84	87.10
Xception*	96.06	98.89	86.35	90.25
RGB baseline	96.08	98.98	86.36	91.00
Ours	96.94	99.32	88.39	**92.98**

Implementation Details and Evaluation Metrics The proposed framework is implemented by PyTorch on an NVIDIA Tesla A100 GPU (40GB). We use Efficient-B4 (EN-b4) [28] as the backbone network and initialized with the weights pre-trained on ImageNet. We employed an Adam optimizer with a cosine learning rate scheduler and set the training hyper-parameters by: the mini-batch size as 12, the initial learning rate as 0.0002, the weight decay as 0.05, $\beta_1 = 0.9$, $\beta_2 = 0.999$. We implemented the training stage with 500 epochs. We set $\lambda = 10$ for the loss function.

Following most existing face forgery detection methods, we mainly utilize the Accuracy rate (ACC) and the Area Under Receiver Operating Characteristic Curve (AUC) as our evaluation metrics. We take AUC as the key evaluation metric and reports the frame-level performances.

4.2 Comparison with Previous Methods

In this part, we compare the proposed method with several stat-of-the-art face forgery detection methods. Since following the official data splitting settings[7], we directly cite the results of previous methods from the corresponding papers. We also report performance for the RGB baseline that removes the MSR stream and the proposed FRWI and SGA in our proposed framework.

Detection on different video qualities In the real-world situation, the videos spread in the social medias are always compressed by popular algorithms such as H.264. Therefore, we evaluate our models on two video qualities, i.e., FF++(HQ and LQ). Table 1 reports the comparison results with previous methods. For FF++(HQ), our method achieves comparable high performances (nearly 100%AUC) compared to state-of-the-art methods. Detecting low-quality manipulated face images is a challenging task as severe compression

Table 2. AUC(%) performance of binary detection on the FF++(LQ) dataset with each four manipulation methods. * indicate the model is trained by us implementing the official code. The best results are in bold.

Methods	DF	F2F	FS	NT
MesoNet[6]	89.52	84.44	83.56	75.74
Xception[7]	94.28	**91.56**	93.70	82.11
PRRNet[29]	95.63	90.15	94.93	80.01
SPSL[30]	93.48	86.02	92.26	76.78
MC-LCR[18]	**97.23**	91.08	94.44	82.13
Xception*[7]	95.60	89.76	93.33	78.87
RGB baseline	95.67	88.48	93.50	80.10
Ours	97.14	91.37	**94.94**	**82.54**

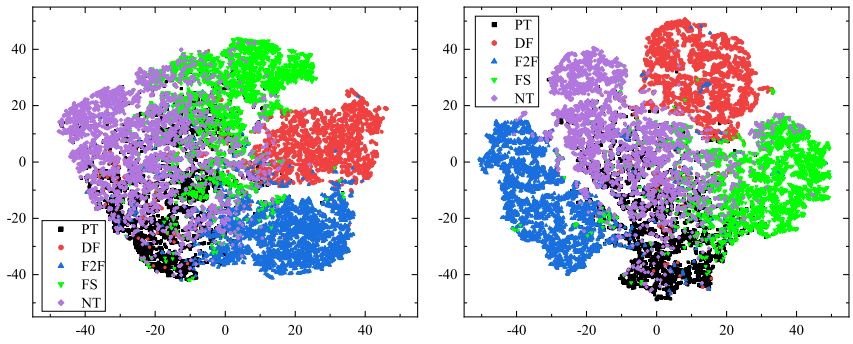
erases much detailed information from the original faces. For FF++(LQ), our method achieves the remarkable performance. Comparing the very recent work D&L[19], our method improves ACC and AUC in 3.55% and 5.88%, respectively. Furthermore, our method achieve better than comparing with the RGB baseline. It demonstrates that introducing the MSR information and joint learning with RGB feartures can effectively improve the detection performance.

Table 3. Recall rate(%) of multi-class classification on the FF++(LQ) dataset with each four manipulation methods. * indicate the model is trained by us implementing the official code. The best results are in bold.

Methods	DF	F2F	FS	NT	PT	Avg
MesoNet[6]	62.45	40.37	28.89	63.35	40.93	47.20
Xception[7]	86.61	78.88	83.16	52.94	75.55	75.43
SPSL[30]	91.16	78.31	88.75	58.97	**77.49**	78.94
D&L[19]	**95.28**	**86.96**	**93.24**	71.66	63.80	82.19
Xception*[7]	93.13	79.24	85.93	66.74	67.25	78.46
RGB baseline	89.38	80.41	86.01	70.36	59.94	77.22
Ours	92.97	84.47	91.13	**74.61**	71.22	**82.88**

Detection on specific manipulation methods Although identifying the authenticity of input faces is of great importance, specifying the manipulation method is also a non-trivial problem. We evaluate the proposed method against different manipulation methods in FF++(LQ). The models were trained and tested on the FF++(LQ) for each manipulation method. Comparing with previous detection methods, the proposed model achieves the best detection accuracy on all four manipulation methods.

Furthermore, multi-classification is more challenging and significant than binary classification.We further evaluate the proposed model on this five-way (pristine and four respective manipulation methods) classification task. As reported



(a) RGB baseline (b) Our proposed method

Fig. 4. A t-SNE [32] visual comparison of embedding spaces between RGB baseline (left) and our proposed method (right) on FF++(LQ) in the multi-classification task.

in Table 3, our method achieve the best average recall performance. As for the recent work D&L[19], although the performance of identifying DF, F2F, and FS is slightly better than our method, the accuracy of distinguishing NT and real samples is very poor. This reveals that the D&L method tends to over-fit deepfake samples with large forgery artifacts, and perform a high false alarm rate.

We also show the t-SNE [32] feature spaces of data in FF++(LQ) with the multi-class classification task. As shown in Figure 4, the RGB baseline is more likely to confuse pristine faces with NT-based fake faces because this manipulation method modifies very limited pixels in the spatial domain. In particular, NT-based images, which just slightly tampered with lip, are very similar to pristine images causing almost indistinguishable in the RGB domain. Conversely, our proposed method can split up all classes in the embedding feature spaces. These improvements may benefit from the introduction of the MSR information.

Cross-datasets evaluations Most existing detection models always suffer a significant performance drop when applied to unseen datasets. To comprehensively evaluate the generalization ability of the proposed model, we conduct extensive cross-dataset experiments in this paper. We train our model on the FF++/DF and Pristine (HQ) data and test it on the unseen Deepfake-TIMIT(DT-HQ/LQ)[33], CelebDF [24], DFD-HQ[5] and DFDC-p [34] datasets.

As shown in Table 4, the proposed method achieves the best generalization performances under all cross-dataset settings. For trained on FF++/DF and tested on CelebDF, which is a common protocol that indicates the generalization performance, our method outperforms at least 2.8% at the AUC metric compared with other methods. The cross-dataset experiment demonstrates that the proposed model is capable of achieving high generalization capability.

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

Table 4. AUC(%) performance of cross-dataset evaluations. * indicate the model is trained by us implementing the official code. The results for within-dataset settings are shown in gray and the best results are in bold.

Methods	FF++/DF	DT-LQ	DT-HQ	DFD-HQ	DFDC-P	CelebDF
Xception[7]	99.70	95.90	94.40	85.90	72.20	65.50
Nirkin *et al.*[35]	99.70	-	-	-	-	66.00
MC-LCR[18]	99.84	-	-	-	-	71.61
D&L[19]	**99.85**	56.08	47.20	76.23	-	70.65
Xception*[7]	99.74	92.83	90.49	88.21	71.66	59.09
Ours	99.67	**97.69**	**94.58**	**89.44**	**76.70**	**74.46**

Table 5. Ablation studies on the FF++(LQ) dataset with identifying specific manipulation methods.

RGB	MSR	FRWI	SGA	DF	FF	FS	NT
√	√	√	√	97.14	91.37	94.94	82.54
√	√		√	↓ 0.65	↓ 0.71	↓ 1.07	↓ 0.51
√	√	√		↓ 1.13	↓ 1.27	↓ 1.46	↓ 0.91
√				↓ 1.47	↓ 2.89	↓ 1.44	↓ 2.44
	√			↓ 3.43	↓ 3.45	↓ 2.94	↓ 5.79

4.3 Ablation Studies and Visualizations

To explore the influence of each component, we evaluated the proposed model and its variants by identifying the specific manipulation method on FF++(LQ). The results are present in Table 5. From these experiments we get the following observations. In the single-stream setting, using only the RGB or MSR data as input leads to poor results. In the two-stream setting, combining the original two stream with the proposed FRWI or SGA can improve the performance, which verifies that the MSR input is distinct and complementary to the RGB data. The performance can be further improved by both adding the proposed FRWI and SGA, reaching the peak when using the overall proposed framework. This shows the effectiveness of each module: MSR exposes fake clues in the RGB space as supplementary information, and the FRWI and SGA enhance the above information by integrating them.

Furthermore, we presented the visualization of MSR and predicted salicency maps with SGA in Figure 5 and qualitatively analyzed the results. We enlarged the predicted saliency maps $\hat{\mathcal{M}}^l$ to the same size as the ground-truth forgery mask $\mathcal{M}_{gt}$. We can observe that the MSR image can attenuate content-specific colors and enhance forgery clues. As for predicted salicency maps, the predicted saliency map can accurately localize the forgery region at all three semantic levels. It demonstrates that SGA can promote the RGB feature to capture the subtle forgery clues with the help of multi-scale spatial guidance.

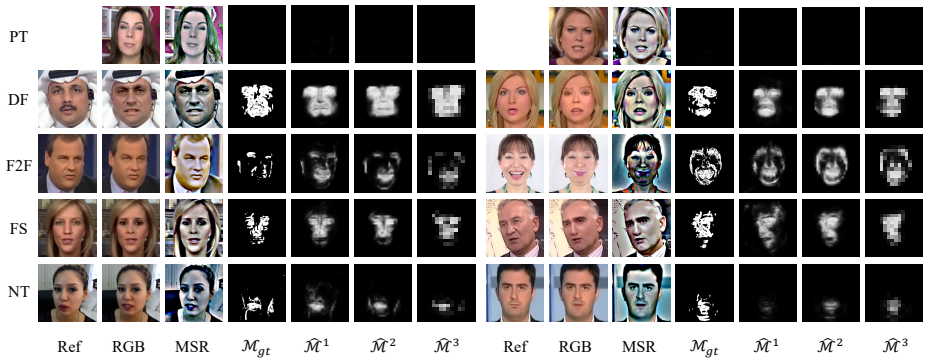


Fig. 5. Qualitative results. For the fake faces, we also display the corresponding real faces for reference.

5 Conclusions

In this work, we present a novel two-stream framework for deepfake detection. In particular, we utilize a multi-scale retinex enhancement inspired by the illumination-reflection model and compose a novel modality, named MSR, to complementary capture the forgery traces. To take full advantage of the MSR information, we propose a two-stream network combined with multi-scale salience-guided and feature re-weighting modules. The multi-scale salience-guided attention module guides the RGB feature extractor to concentrate more on forgery traces from the MSR perspective at multi-scale level. The feature re-weighting module leverages the correlation between the two complementary modalities to promote feature learning for each other. Extensive experiments demonstrate that the proposed framework achieves consistent performance improvement compared to state-of-the-art methods. Future studies can focus on extending this work at a video level so that multiple types of manipulated facial videos can be identified by using a general model.

Acknowledgments This work was supported in part by NSFC (Grant 61872244), Guangdong Basic and Applied Basic Research Foundation (Grant 2019B151502001), Shenzhen R&D Program (Grant JCYJ20200109105008228).

References

1. Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing High Fidelity Identity Swapping for Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5074–5083, 2020.
2. Le Minh Ngo, Christian aan de Wiel, Sezer Karaoglu, and Theo Gevers. Unified Application of Style Transfer for Face Swapping and Reenactment. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

3. Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. HifiFace: 3D Shape and Semantic Prior Guided High Fidelity Face Swapping. In *Twenty-Ninth International Joint Conference on Artificial Intelligence*, volume 2, pages 1136–1142, 2021.
4. Y. Li, M. Chang, and S. Lyu. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018.
5. F. Matern, C. Riess, and M. Stamminger. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92, 2019.
6. 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)*, pages 1–7, 2018.
7. Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019.
8. Yuezun Li and Siwei Lyu. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 46–52, 2019.
9. Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-Ray for More General Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020.
10. Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In *ECCV 2020*, pages 86–103, 2020.
11. Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing Face Forgery Detection With High-Frequency Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16317–16326, 2021.
12. Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.
13. Zhengzhe Liu, Xiaojuan Qi, and Philip H. S. Torr. Global Texture Enhancement for Fake Face Detection in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8060–8069, 2020.
14. D.J. Jobson, Z. Rahman, and G.A. Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing*, 6(7):965–976, 1997.
15. H. Chen, G. Hu, Z. Lei, Y. Chen, N. M. Robertson, and S. Z. Li. Attention-Based Two-Stream Convolutional Networks for Face Spoofing Detection. *IEEE Transactions on Information Forensics and Security*, 15:578–593, 2020.
16. Jian Han and Theo Gevers. MMD based Discriminative Learning for Face Forgery Detection. In *Proceedings of the Asian Conference on Computer Vision*, 2020.
17. Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning Self-Consistency for Deepfake Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15023–15033, 2021.

18. Gaojian Wang, Qian Jiang, Xin Jin, Wei Li, and Xiaohui Cui. MC-LCR: Multimodal contrastive classification by locally correlated representations for effective face forgery detection. *Knowledge-Based Systems*, 250:109114, 2022.
19. Chenqi Kong, Baoliang Chen, Haoliang Li, Shiqi Wang, Anderson Rocha, and Sam Kwong. Detect and Locate: Exposing Face Manipulation by Semantic- and Noise-Level Telltales. *IEEE Transactions on Information Forensics and Security*, 17:1741–1756, 2022.
20. Edwin H. Land and John J. McCann. Lightness and retinex theory. *J. Opt. Soc. Am.*, 61(1):1–11, Jan 1971.
21. Qianting Ma, Yang Wang, and Tieyong Zeng. Retinex-based variational framework for low-light image enhancement and denoising. *IEEE Transactions on Multimedia*, pages 1–9, 2022.
22. Jun Xu, Yingkun Hou, Dongwei Ren, Li Liu, Fan Zhu, Mengyang Yu, Haoqian Wang, and Ling Shao. STAR: A Structure and Texture Aware Retinex Model. *IEEE Transactions on Image Processing*, 29:5022–5037, 2020.
23. Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic Image Synthesis With Spatially-Adaptive Normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.
24. Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020.
25. Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016.
26. Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics*, 38(4):66:1–66:12, 2019.
27. K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.
28. Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.
29. Zhihua Shang, Hongtao Xie, Zhengjun Zha, Lingyun Yu, Yan Li, and Yongdong Zhang. PRRNet: Pixel-Region relation network for face forgery detection. *Pattern Recognition*, 116:107950, 2021.
30. Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 772–781, 2021.
31. Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-Attentional Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2185–2194, 2021.
32. Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
33. P. Korshunov and S. Marcel. Vulnerability assessment and detection of Deepfake videos. In *2019 International Conference on Biometrics (ICB)*, pages 1–6, 2019.

34. Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The Deepfake Detection Challenge (DFDC) Preview Dataset. *arXiv:1910.08854 [cs]*, 2019.
35. Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. DeepFake Detection Based on Discrepancies Between Faces and their Context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.