Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.036688
Article

Mining Fine-Grain Face Forgery Cues with Fusion Modality

Shufan Peng, Manchun Cai*, Tianliang Lu and Xiaowen Liu

People's Public Security University of China, Beijing 100038, China
*Corresponding Author: Manchun Cai. Email: caimanchun@ppsuc.edu.cn

Received: 09 October 2022; Accepted: 08 February 2023

Abstract: Face forgery detection is drawing ever-increasing attention in the academic community owing to security concerns. Despite the considerable progress in existing methods, we note that: Previous works overlooked fine-grain forgery cues with high transferability. Such cues positively impact the model's accuracy and generalizability. Moreover, single-modality often causes overfitting of the model, and Red-Green-Blue (RGB) modal-only is not conducive to extracting the more detailed forgery traces. We propose a novel framework for fine-grain forgery cues mining with fusion modality to cope with these issues. First, we propose two functional modules to reveal and locate the deeper forged features. Our method locates deeper forgery cues through a dual-modality progressive fusion module and a noise adaptive enhancement module, which can excavate the association between dual-modal space and channels and enhance the learning of subtle noise features. A sensitive patch branch is introduced on this foundation to enhance the mining of subtle forgery traces under fusion modality. The experimental results demonstrate that our proposed framework can desirably explore the differences between authentic and forged images with supervised learning. Comprehensive evaluations of several mainstream datasets show that our method outperforms the state-of-the-art detection methods with remarkable detection ability and generalizability.

Keywords: Face forgery detection; fine-grain forgery cues; fusion modality; adaptive enhancement

1 Introduction

Recent studies have shown rapid advances in face forgery techniques [1–4], which allow attackers to perform facial area manipulation at a much lower cost. With the remarkable success represented by Deepfakes, the subtle differences between authentic and forged images are indistinguishable. Face forgery's malicious usage may cause serious social problems and political threats. Therefore, developing high-performance detection methods has become a popular research direction.

Face forgery detection technology intends to prevent the harm caused when Deepfakes technology is abused. Such as preventing Deepfakes technology from manipulating elections [5], interfering with media messages [6], creating pornography featuring female celebrities, creating fake accounts, and financial fraud. Some forgeries involving politics are likely to have unpredictable consequences [7]; in December 2020, DeepFake videos featuring Vladimir Putin and Kim Jong-un appeared on social media, exciting discussions about elections and democracy in the United States. Hence high-performance face forgery detection methods have become a hot research concern. The ideal detection model can be applied to most forgery data in the first place and has good detection ability and generalizability to deal with unseen forgeries.

Researchers have developed various methods to detect face forgery employing distinct traces, such as apparent visual artifacts [8–10], temporal inconsistencies [11–15], and multimodal conflicts [16–18]. These traces are not universal. Existing methods are more demanding on data and less meaningful in real scenarios. In the real Internet, the vast majority of face forgery is presented as images. Thus we put our research on the image-based face forgery detection method. Previous image-based research focused on modifying the network structure or extracting various features. These methods are based on single-modality or physiological features, which could otherwise be more satisfactory in terms of accuracy and generalization. Spatial-based detection methods generally apply modified visual network architectures to face forgery detection, such as capsule networks [19], Xception [20], vision transformers [21], etc. The above methods' robustness is susceptible to image post-processing, such as video compression and smoothing. Some image processing methods, such as frequency analysis, have been introduced for highly compressed datasets to face forgery detection. Durall et al. [22] utilized the unnatural spectral distribution generated by the prevalent generative models for detection. Frank et al. [23] found that the generative adversarial network (GAN)-generated images exhibited severe artifacts in the frequency domain. These methods are still single-modal-only, and the upper limit of performance achieved on different datasets is somewhat constrained. Moreover, these single modality-based detection methods fail to explore forgery patterns. These forgery traces extracted depend heavily on the training data and may fail on unseen forgeries.

Supervised face forgery detection methods rely on neural network fitting capabilities for learning. With a narrow gap between network architectures, how to uncover more critical and more generalizable forgery features becomes a problem worth investigating. From how face forgery images are generated [24,25], a forgery face often blends two existing faces or is synthesized by deep neural networks (DNNs). This mode has some similarities with image splicing. Both are similar to the blending of two types of images. However, there are obvious signs of tampering at the boundary between the manipulated region and the genuine region of the spliced image.

In contrast, face forgery images represented by Deepfakes tend to have fine-grain forgery cues, such as visual artifacts and unusual noise, resulting in an anomaly in high-frequency regions. For face forgery detection tasks, local cues play a more critical role than global semantics. Unlike image splicing detection, which utilizes boundary information, several advanced manipulation methods [26,27] generate local forgery traces, leading to global facial features' discriminability suffering from small-scale tampering. Therefore, exploring the universal local forgery traces is the key to the face forgery detection task.

We observe that if only the RGB modality is employed, detailed local properties are prone to be overlooked as the perceptual field increases. Moreover, we assume that the key to exploring the critical local forgery traces is to exploit the inconsistency in details between authentic and forged images. Several works have proposed solutions in response to this phenomenon. Dang et al. [28] utilized attention maps to locate manipulated parts, Chai et al. [29] segmented images into local patches, and Zhao et al. [30] employed multiple spatial attention heads to focus on the image's different regions. Although the above approaches emphasize local features, local forgery features that rely only on color space are fragile for image post-processing. Future solutions need to be more robust and practical in real scenarios. The noise modality, on the other hand, due to its local properties, its introduction helps the model to learn some local anomalies or local forgery traces. Previous works utilizing image noise modality still intrinsically treat the noise modality as an independent complementary feature to enhance the model's accuracy. Zhou et al. [31] leveraged the complementary properties of RGB and noise streams to detect and locate tampered images efficiently. Luo et al. [32] observed that current convolutional neural network (CNN)-based detectors tend to over-fit color textures and proposed introducing multi-scale noise features to improve generalization across multiple benchmark datasets. Fei et al. [33] proposed a learnable adaptive spatial rich model (ASRM) filter to compensate for conventional noise features' shortcomings in adaptive. Previous work ignored the correspondence between noise features and RGB features in the spatial domain. We expect to exploit the spatial commonality of the two modal features to guide the model's perception of local forgery cues.

In contrast to the above approaches, which utilize two modalities of global image features to complement each other, our method is expected to learn more about generalizable forgery patterns. We design a novel fusion enhancement method to introduce the noise modality and employ a particular chunking learning approach to enhance the sensitivity to fine-grain face forgery cues. We design a novel fusion enhancement method to introduce the noise modality and employ a particular chunking learning approach to enhance the sensitivity to fine-grain face forgery cues.

Based on these observations of face forgery image properties, the main motivations behind our work are: (1) In this work, we focus on capturing forgery traces from the perspective of fine-grain face forgery cues. Such local semantics with high transferability have better detectability and generalization. In contrast, learning the global features of images is less important. (2) Specifically, unlike the previous view of frequency information as a separate feature stream, we note that noise features contain some fine-grained local anomalies that are often not easily detectable in RGB features. As an inherent property of images, noise features also correspond to RGB features spatially, and the two features can somewhat complement each other. Therefore, we want to fuse noise features to guide the network to notice such local anomalies of forged images and use them as forgery cues for subsequent sensitive block mining. (3) Since deeper features correspond to larger perceptual fields, a deep network is challenging to learn fine-grained noise features adequately. We design a novel adaptive enhancement method for noise features in the fusion modality that can adaptively adjust the magnitude of the enhancement according to data. (4) We employ a novel chunking learning approach to enhance the network's learning of fine-grained face forgery cues. Specifically, given a face image, we select the sensitive blocks that are most important for detection results by aggregating deep feature descriptors. Unlike previous patch-wise learning methods, our approach adaptively learns vital local forgery patterns while ignoring the less critical features and does not require external annotation.

Our contributions can be summarized as follows:

We propose a novel perspective to address the face forgery detection task, aiming at mining fine-grain face forgery cues to learn the difference between authentic and forged images. To end this, we introduce and adaptively enhance the image noise modality utilizing sensitive blocks to ensure the discrimination between genuine and manipulated regions in deep local features.

We propose two functional modules to reveal and locate the deeper forged features. A dual-modality progressive fusion module (DPFM) is designed to explore dual-modal correlations in spatial and channel dimensions in shallow features and fuse them on this basis. Furthermore, a noise adaptive enhancement module (NAEM) is designed to excavate the artifact hidden in the noise feature adaptively.

We design a sensitive patch branch (SPB) shared with the main network parameters to isolate vital subtle forgery traces. SPB selects as input the sensitive blocks corresponding to the most critical windows for classifiers. SPB effectively enhances the learning of the network for forgery cues, which gives the network remarkable detectability and generalization.

We performed a comprehensive evaluation of mainstream face forgery datasets. The experimental results demonstrate our proposed method's effectiveness with the most advanced competitors.

2 Related Work

Spatial-based forgery detection methods. Early face forgery techniques tend to generate obvious forged signs. Many manual feature-based methods explore forgery image anomalies in the spatial domain. These manual features include image noise residual [34,35], face warping artifacts (FWA) [8], visual artifacts [9], etc. Face X-ray [36] is based on the property of blending boundaries when faces are swapped, significantly generalizing local forged images. However, this method achieved undesirable results in highly compressed or entire synthetic datasets. Given the excellent performance of deep neural networks in computer vision, DNN-based methods have gradually become the mainstream of research. Some works directly applied existing classification networks [37–39]. MesoNet [40] utilized a shallow CNN architecture for forgery detection based on mid-level semantics. Bayer et al. [41] developed a new convolutional layer capable of adaptively learning manipulation detection features. Current state-of-the-art methods explore and learn about forged features. Dang et al. [28] decomposed the face forgery detection task into the localization of manipulated regions and detection. Zhao et al. [30] modeled the detection task as a fine-grain image classification task, leading to learning the proposed Multi-attentional Deepfake Detection (MAT) framework for local forgery traces and shallow texture features in manipulated images.

The above spatial-based detection methods tend to modify the network structure. Their detection performance varies widely across diverse datasets. RGB-modal-only Methods can render the detector overfit to method-specific color texture and thus fail to generalize. Furthermore, these methods are highly impacted by image post-processing, such as compression and smoothing masks—lack applicability in real scenarios.

Frequency-based forgery detection methods. The spatial-based detection methods are susceptible to compression rate. There are anomalies, such as distribution differences of high-frequency components and checkerboard artifacts in the synthetic images. Furthermore, researchers have applied many traditional mathematical methods to practical tasks [42–45] with impressive results in recent years. Thus, frequency analysis is introduced into the detection task and achieves desired results in highly compressed datasets. Some works utilized digital image processing methods such as Discrete Fourier Transform (DFT) [46,47] and Discrete Cosine Transform (DCT) [48] to obtain frequency domain features and detect anomalies. Frequency in Face Forgery Network (F^3-Net) [49] proposed frequency-aware decomposition and local frequency statistics to obtain forgery information in the frequency domain. Fake Generated Painting Detection via Frequency Analysis (FGPD-FA) [50] performs forgery detection by fusing three distinct frequency domain features.

However, since different forgery generation methods vary dramatically in the frequency domain space, we observe that the accuracy of the frequency-based detection method alone is substantially reduced on unseen datasets. Most existing frequency analysis-based methods directly convert the entire image into a spectrum. Locally tampered faces [27,51] show indistinct anomalies in the global frequency domain. Thus, these methods also suffer from subtle local forgery traces.

Forgery detection methods combine spatial and frequency features. Due to single-modality limitations, dual-modality-based detection methods are becoming mainstream research [11,12,32,52]. Spatial-Phase Shallow Learning (SPSL) [53] employed a shallow network to capture the pixel differences in the phase spectrum of the synthetic images. The shallow network makes it difficult for the method to detect subtle forgery traces. Frequency-aware Discriminative Feature Learning (FDFL) [54] designed a single-center loss to improve intra-class compactness and inter-class separability in the embedding space with dual-modal features. Multimodal Contrastive Classification by Locally Correlated Representations (MC-LCR) [55] proposed a novel perspective that aims to amplify the implicit local discrepancies between authentic and forged face images from dual-modality features.

Previous works treat spatial and frequency domain features as two separate feature streams but neglect the existence of correspondence between the two features in terms of location. In essence, they need to explore the forgery cues on dual-modal features further.

Forgery detection methods utilize local receptive fields. Previous methods of patch-wise training tend to perform even chunking [29,55]. We note that existing patch-wise detection methods ignore the variation in the forged features between patches and lack of adaptivity. To end this, we introduce more flexible activation map-based sensitive patches, which can extract vital features of arbitrary size. The sensitive patch-based detection method can improve our framework's accuracy and generalization. Sensitive patches can enhance the learning of manipulated patterns rather than global features, such as visual artifacts. In addition, such local semantics makes the detector less susceptible to high-level facial image features, achieving better generalization.

3 The Proposed Method

Owing to post-processing with compression or smoothing in mainstream datasets, the difference between authentic and forged images is difficult to discriminate by RGB features alone. Face forgery images usually consist of authentic areas as well as forged areas. Noise features are inherent to the specificity of images. The noise features of the post-generated forged region and the genuine region are difficult to match. We, therefore, carefully design two modules to integrate RGB features with noise features fully. Furthermore, extract important local features employing sensitive patches to guide the learning of our framework for fine-grained forgery cues. Our framework is illustrated in Fig. 1.

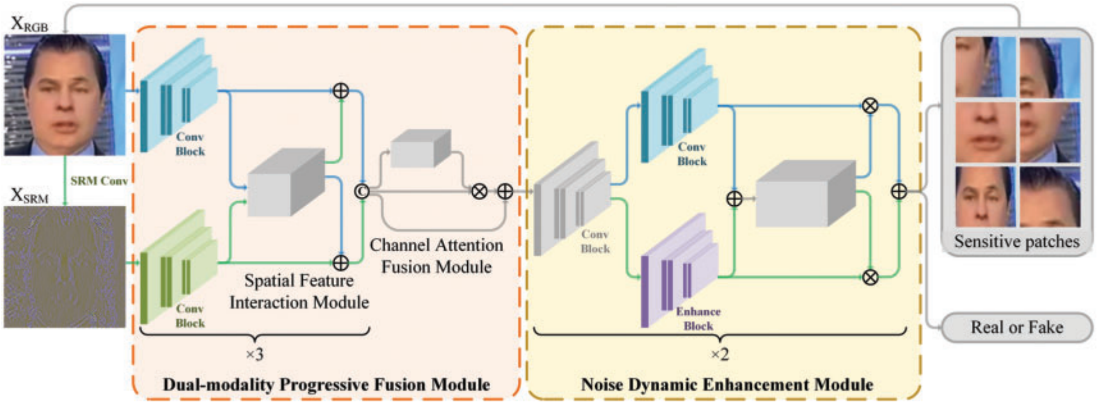

X_{RGB}
X_{SRM}
SRM Conv
Conv Block
Conv Block
Spatial Feature Interaction Module
Channel Attention Fusion Module
×3
Dual-modality Progressive Fusion Module
Conv Block
Conv Block
Enhance Block
×2
Noise Dynamic Enhancement Module
Sensitive patches
Real or Fake

Figure 1: In the dual-modality progressive fusion module, dual-modal spatial and channel correlations are mined separately using the spatial feature interaction module and channel attention fusion module. Different levels of noise features are enhanced adaptively in the noise adaptive enhancement module

Our framework's overall training and testing process is illustrated in Fig. 2. Our framework contains a novel modal fusion-enhancement process and an adaptive sensitive patch mining-learning process. Specifically, in the training phase, we obtain critical sensitive patches based on fusion modality and input them into a sensitive patch branch shared with the main network parameters. In the testing phase, we directly use the main network for testing.

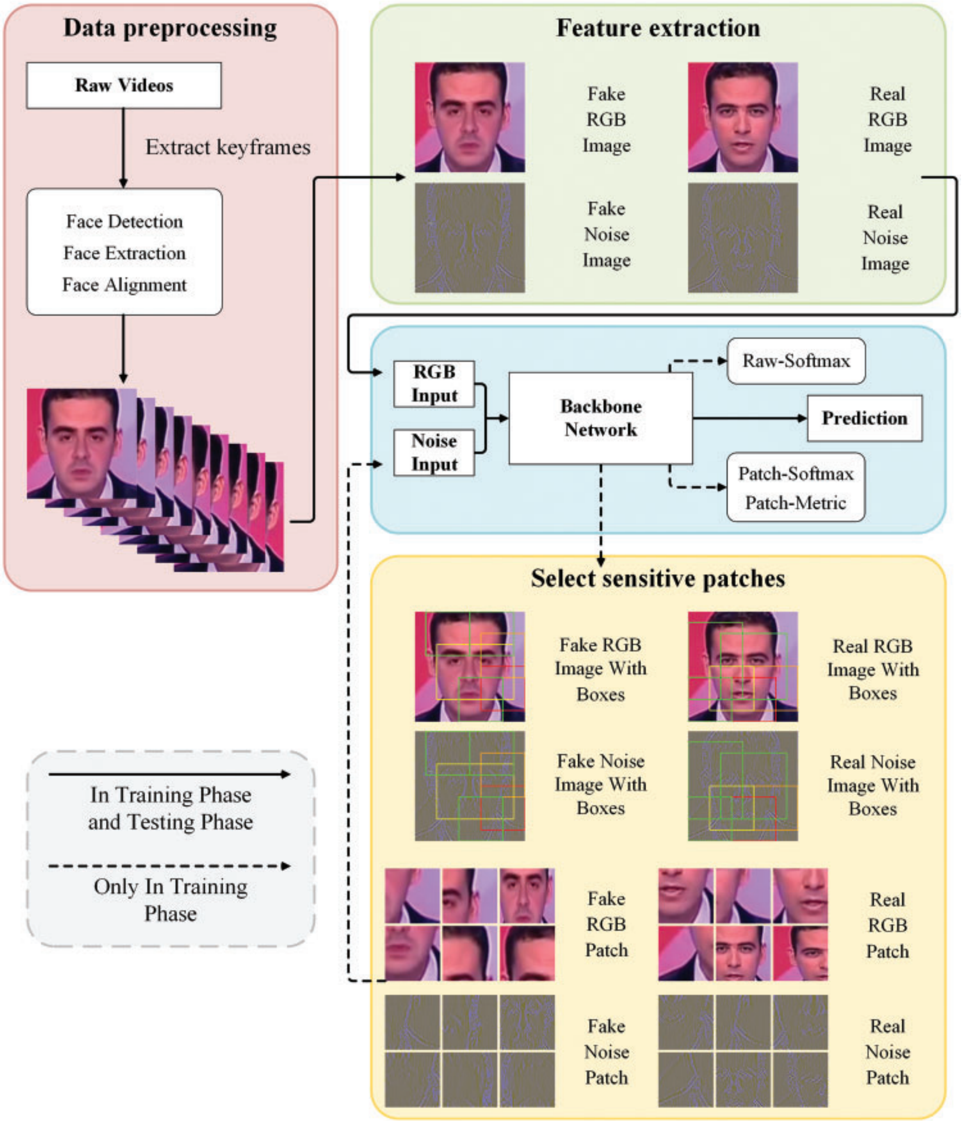


Figure 2: The overall training and testing process of our framework

3.1 Dual-Modality Progressive Fusion Module

Most of the existing face forgery generators are based on GAN, where the up-sampling causes anomalies in the frequency domain features of the face forgery images. As an inherent property of the image, the noise features of the manipulated region are often inconsistent with those of the genuine region. The image's post-forged part leaves a unique trace in the noise space, and this location information can correspond well to the RGB space. This property helps our proposed framework to locate high discrepancy regions between authentic and forged images. We thus introduce noise features to guide our framework in mining the differences between authentic and forged images.

In previous work, RGB and noise features were often treated as two separate feature streams and concatenated directly in the high-level features on the channel dimension. However, the RGB and noise features are not wholly unrelated features. There is quite a lot of shared information in these two modalities. Two-Stream networks may weaken the correlation between spatial features and cause redundancy in the network structure. Therefore, we propose a progressive fusion method to fully obtain the spatial and channel features of the two modalities.

As the network deepens, semantic information increases as the reception field increases, and spatial information diminish as the resolution decreases. We need to retain the spatial information in both modalities for sensitive patch mining. We use a progressive fusion strategy on feature maps of different resolutions in the fusion process. To this end, we propose a dual-modality progressive fusion module in the shallow layer to fully fuse the dual-modal features at different scales. Our proposed module consists of two sub-modules: a spatial feature interaction module and a channel attention fusion module (See Fig. 3), which explore the spatial and channel correlations between the two modalities separately.

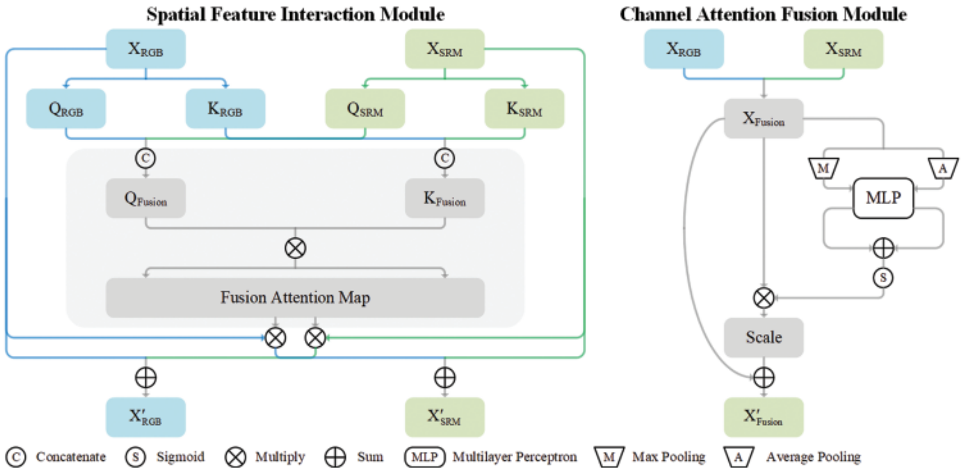


Figure 3: The spatial feature interaction module and the channel attention fusion module

In the spatial feature interaction module, let $\mathrm{X}_{\mathrm{RGB}}, \mathrm{X}_{\mathrm{SRM}} \in \mathbb{R}^{C\times H\times W}$ denote the input of RGB features and noise features. First, we obtain the information within each modality by $\mathrm{Q}_{\mathrm{RGB}}, \mathrm{Q}_{\mathrm{SRM}} \in \mathbb{R}^{C\times W\times H}$ and $\mathrm{K}_{\mathrm{RGB}}, \mathrm{K}_{\mathrm{SRM}} \in \mathbb{R}^{C\times H\times W}$. Then, stitching in spatial dimensions to obtain the fused modalities $\mathrm{Q}_{\mathrm{Fusion}} \in \mathbb{R}^{C\times W\times 2H}$ and $\mathrm{K}_{\mathrm{Fusion}} \in \mathbb{R}^{C\times 2H\times W}$. On this basis, the fusion attention map $\mathrm{M} \in \mathbb{R}^{C\times W\times W}$ is obtained:

$$\mathrm{M} = \mathrm{softmax}\left(\mathrm{Q}_{\mathrm{Fusion}} \otimes \mathrm{K}_{\mathrm{Fusion}}\right) \tag{1}$$

where $\otimes$ denotes the multiply operator, the final output is $\mathrm{X}'_{\mathrm{RGB}}, \mathrm{X}'_{\mathrm{SRM}} \in \mathbb{R}^{C\times H\times W}$:

$$\mathrm{X}'_{\mathrm{RGB}} = \mathrm{X}_{\mathrm{RGB}} + BN\left(Conv\left(\mathrm{X}_{\mathrm{SRM}} \otimes \mathrm{M}\right)\right) \tag{2}$$

$$\mathrm{X}'_{\mathrm{SRM}} = \mathrm{X}_{\mathrm{SRM}} + BN\left(Conv\left(\mathrm{X}_{\mathrm{RGB}} \otimes \mathrm{M}\right)\right) \tag{3}$$

where BN denotes batch normalization, + denotes the sum operator. $\mathrm{X}'_{\mathrm{RGB}}$ and $\mathrm{X}'_{\mathrm{SRM}}$ can interact with the features of another modality effectively.

The channel attention fusion module utilizes the attention mechanism to facilitate inter-channel interactions. This module combines information from all channels and determines each channel's significance. Let $\mathrm{X}'_{\mathrm{Fusion}}, \mathrm{X}'_{\mathrm{Fusion}} \in \mathbb{R}^{C\times H\times W}$ denote the fusion modality feature maps of the input and output:

$$\mathrm{X}_{\mathrm{Fusion}} = Conv\left(Cat\left(\mathrm{X}_{\mathrm{RGB}}, \mathrm{X}_{\mathrm{SRM}}\right)\right) \tag{4}$$

$$\mathrm{X}'_{\mathrm{Fusion}} = \mathrm{X}_{\mathrm{Fusion}} + \mathrm{X}_{\mathrm{Fusion}} \otimes \sigma\left(MLP\left(GAP\left(\mathrm{X}_{\mathrm{Fusion}}\right) + GMP(\mathrm{X}_{\mathrm{Fusion}})\right)\right) \tag{5}$$

where Cat is the concatenate operator in the channel dimension, MLP is a multi-layer perceptron, GAP denotes the global average pooling, GMP denotes the global max pooling, and σ is the sigmoid activation function. We adopt this progressive fusion method to fully obtain the information of both modalities, using this fusion modality as a basis for mining the differences between authentic and forged images.

3.2 Noise Adaptive Enhancement Module

For datasets with insignificant forgery features, subtle noise will play a critical role in forgery cue mining. The neural network has a low learning priority spectral bias for high frequencies. Furthermore, the sizeable perceptual field of high-level features is challenging to extract local noises. To this end, we design a noise adaptive enhancement module (See Fig. 4) to excavate the artifact hidden in the noise feature.

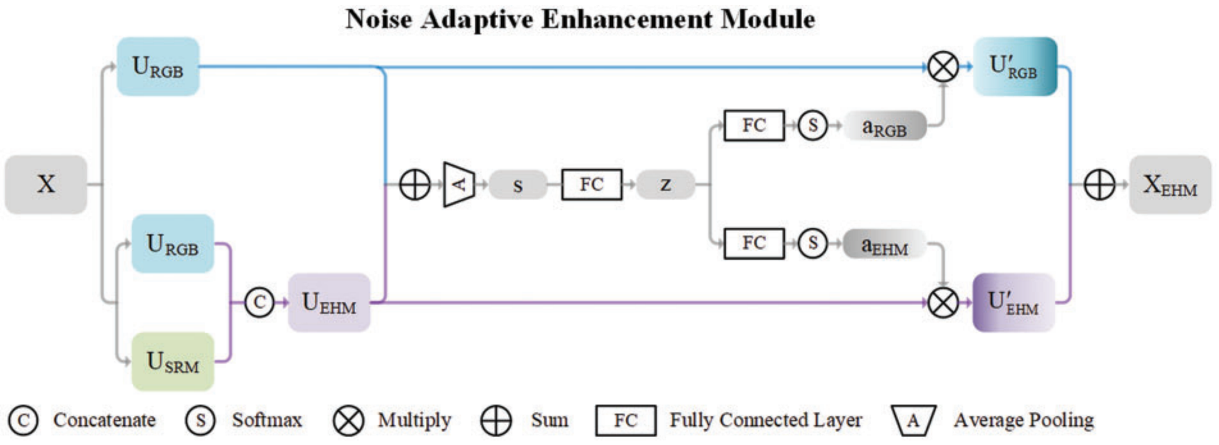


Figure 4: The noise adaptive enhancement module

In the noise adaptive enhancement module, for the input feature map $\mathrm{X} \in \mathbb{R}^{C\times H\times W}$, we first conduct two transformations:

$$\mathrm{U}_{\mathrm{RGB}} = Conv\left(\mathrm{X}\right) \tag{6}$$

$$\mathrm{U}_{\mathrm{EM}} = Conv\left(Cat\left(Conv\left(X\right), f_{\mathrm{SRM}}\left(X\right)\right)\right) \tag{7}$$

where $\mathrm{U}_{\mathrm{RGB}}, \mathrm{U}_{\mathrm{EM}} \in \mathbb{R}^{C\times H\times W}$. Corresponding to Xception, we use the depthwise separable convolution with a 3×3 kernel. We generate channel-wise statistics $\mathrm{s} \in \mathbb{R}^{C}$ by fusing the RGB branch and enhanced noise branch via element-wise summation and global average pooling:

$$\mathrm{s} = f_{\mathrm{GAP}}\left(\mathrm{U}_{\mathrm{RGB}} + \mathrm{U}_{\mathrm{EHM}}\right) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(\mathrm{U}_{\mathrm{RGB}}\left(i,j\right) + \mathrm{U}_{\mathrm{EHM}}\left(i,j\right)\right) \tag{8}$$

Then, the squeeze-excitation operation is used to learn the feature associations on the channel dimension. We obtain the compact feature descriptor $\mathrm{z} \in \mathbb{R}^{d}$ and normalize the soft attention vectors $\mathrm{a}_{\mathrm{RGB}}, \mathrm{a}_{\mathrm{EHM}} \in \mathbb{R}^{C}$ of the corresponding channels of two branches via a softmax operator.

$$\mathrm{z} = f_{\mathrm{Squeeze}}\left(\mathrm{s}\right) \tag{9}$$

$$\left[\mathrm{a}_{\mathrm{RGB}}, \mathrm{a}_{\mathrm{EHM}}\right] = \mathrm{softmax}\left(\left[f_{\mathrm{Excitation-RGB}}\left(\mathrm{z}\right), f_{\mathrm{Excitation-EHM}}\left(\mathrm{z}\right)\right]\right) \tag{10}$$

The final feature map is obtained through the attention weights of two branches:

$$\mathrm{X}_{\mathrm{EHM}} = \mathrm{a}_{\mathrm{RGB}} \cdot \mathrm{U}_{\mathrm{RGB}} + \mathrm{a}_{\mathrm{EHM}} \cdot \mathrm{U}_{\mathrm{EHM}} \tag{11}$$

where $\mathrm{X}_{\mathrm{EHM}} \in \mathbb{R}^{C\times H\times W}$ is the fusion modality feature map after adaptive enhancement, two noise adaptive enhancement modules are inserted between the blocks of the backbone network, preserving and amplifying subtle noise in low-level and high-level features, respectively. More helpful information is provided for subsequent sensitive patch mining.

3.3 Sensitive Patch Branch

Across different forgery face generators, local forgery cues tend to have higher generalizability than global structure. The global structure may vary among advanced semantic information about different faces. Nevertheless, local cues have better commonality across different manipulation methods, such as shared visual artifact features. We, therefore, hypothesize that local cues are more conducive to face forgery detection. In previous work, the features of interest to these networks were scattered, with relatively poor generalizability. Some work uses image chunking or masking of specific regions to limit the network's perception field for learning. Such methods may lose some of the features of the raw image or lack adaptivity. In particular, for partial forgery images, the network may have difficulty converging during the training process if many non-forged regions are included in the patches labeled as fake.

We propose a fine-grain forgery cue mining method (See Fig. 5) to precisely locate forged regions and use them as sensitive patches to enhance network learning for different forgery cues. Specifically, we use adaptive-sized sliding windows for extracting these sensitive patches. Furthermore, input these patches into the sensitive patch branch that shares parameters with the main network during the training phase. Our framework is ultimately more oriented towards the learning of crucial manipulated patterns.

We aggregate the high-level feature maps $\mathrm{X}_{\mathrm{H}} \in \mathbb{R}^{C\times H\times W}$ in the channel dimension to obtain the corresponding activation maps $\mathrm{A} \in \mathbb{R}^{1\times H\times W}$:

$$\mathrm{A}\left(x,y\right) = \sum_{i=0}^{C-1}\mathrm{x}_{i}\left(x,y\right) \tag{12}$$

where x_i is the ith feature map of X_{H}, (x,y) denotes the coordinates of the feature descriptor in space. The value of $\mathrm{A}(x,y)$ reflects the contribution of the descriptor to the result. To select patches containing more sensitive information, we use the average activation value within the sliding window to define the *score* of these patches:

$$score = \frac{1}{W \times H} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} \mathrm{A}\left(x, y\right) \tag{13}$$

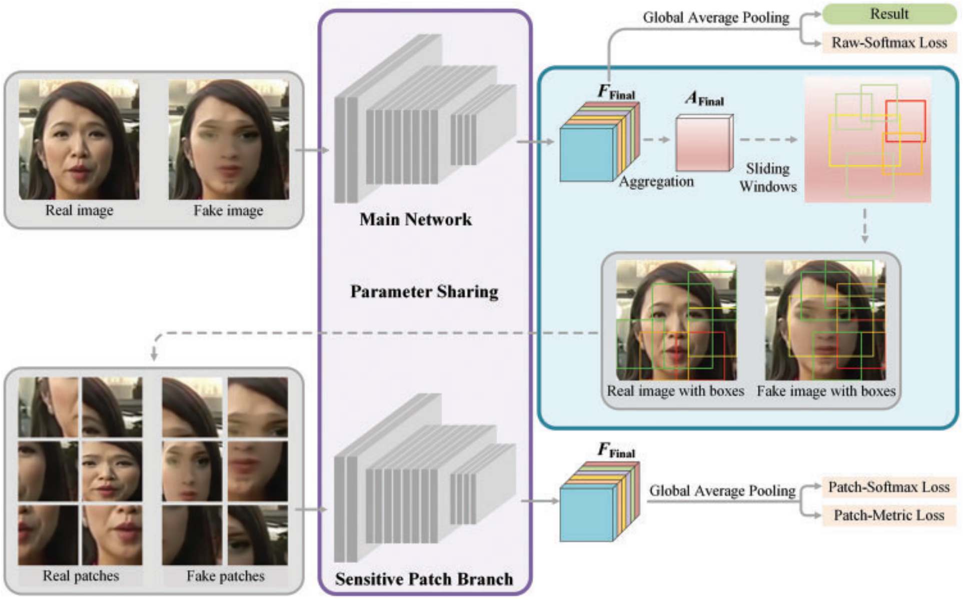


Figure 5: The sensitive patch branch. The network parameters in the purple box are shared. The blue box shows the extraction process of sensitive patches

where H and W are the height and width of windows, we sort by the *score* of all windows and select the first few windows as sensitive patches. We adopt a non-maximum suppression (NMS) between the extracted sensitive patches to mine more forgery cues. The sensitive patches with the highest *score* values are eventually used as input to the branch. To thoroughly learn the subtle forgery traces within sensitive patches, we adopt the Circle Loss as the metric loss to improve the intra-class compactness and inter-class separability in the embedding space.

Our objective function contains the cross-entropy loss of the main network and patch branch and the metric loss between sensitive patches:

$$L_{\text{Raw-Softmax}} = -\log\left(P_{\text{raw}}\left(label\right)\right) \tag{14}$$

$$L_{\text{Patch-Softmax}} = -\sum_{n=0}^{N-1} \log\left(P_{\text{patch}(n)}\left(label\right)\right) \tag{15}$$

$$L_{\text{Patch-Metric}} = \sum_{n=0}^{N-1} L_{\text{Circle}}\left(PE_{\text{patch}(n)}\right) \tag{16}$$

where *label* is the raw image's ground truth label, and P_{raw} is the raw image's category probability. $P_{\text{patch}(n)}$ is the softmax layer's output of the sensitive patch branch corresponding to the *n*th sensitive patch. $PE_{\text{patch}(n)}$ is the feature embedding of the nth sensitive patch. Since the unstable accuracy of the activation map in the first epoch, the total loss is as follows:

$$L_{\text{total}} = \begin{cases} L_{\text{Raw-Softmax}} & \text{if epoch} = 1 \\ L_{\text{Raw-Softmax}} + L_{\text{Patch-Softmax}} + L_{\text{Patch-Metric}} & \text{otherwise} \end{cases} \tag{17}$$

4 Experiments

4.1 Settings

Datasets. We adopt five widely-used public datasets in our experiments, i.e., FaceForensics++ (FF++) [56], FaceShifter (FSR) [57], DeepfakeDetection (DFD) [58], Celeb-DF [26], DeeperForensics-1.0 (DF1.0) [59], and WildDeepfake [60] (See Table 1). We uniformly set the ratio of the training and testing sets to 7:3. We take the high-quality version (c23) and the low-quality version (c40) of FF++ and FSR. FF++ contains four manipulation methods: Deepfakes (DF), Face2Face (F2F) [51], FaceSwap (FSP) [61], and NeuralTextures (NT) [27]. As the level of compression increases, detection of forgery cues can become increasingly challenging.

Table 1: Specifications of benchmark databases

Database	Video scale	Manipulation algorithm
FF++(c23/c40)	1000 real, 4000 fake	DF, FSP, NT, F2F
FaceShifter(c23/c40)	1000 fake	FSR
DFD	363 real, 3068 fake	Improved DF
Celeb-DF	590 real, 5639 fake	Improved DF
DF1.0	11,000 fake	DF-VAE
WildDeepfake	3805 real, 3509 fake	Improved DF

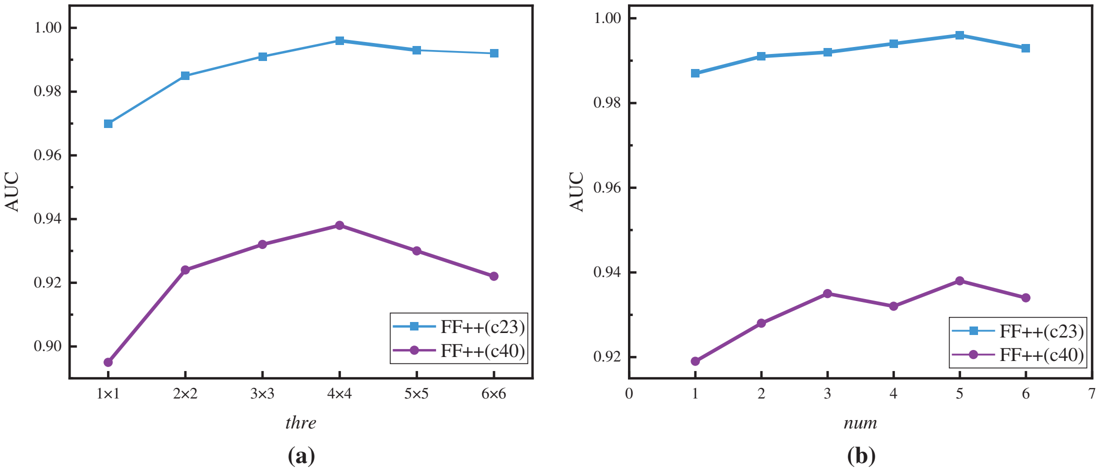
Data preprocessing. Since the dominant face forgery dataset today is in video format, some preprocessing is necessary for our task. We extract keyframes for each video every 10 s, for a total of 30. This process can avoid data redundancy while maintaining the richness of the sample images. For these keyframes, we use Retinaface [62] for face extraction and alignment and resize the aligned faces to 299 × 299. This processing has become a default standard in face forgery detection and facilitates comparing our method with other state-of-the-art works.

Implementation detail. Xception pre-trained on ImageNet is adopted as the backbone of our framework, which is trained with AdamW optimizer with a learning rate of 2×10^{-4}, weight decay of 1×10^{-5}, and batch size of 8. The cosine decay is with a total of 20 epochs. We obtain the comparison results from their paper and specify our implementation by † otherwise.

4.2 Ablation Study

Parameter influence. The number of sensitive patches tagged as *num*, and the minimum threshold of window size tagged as *thre* may affect the final result. Primarily, *thre* controls the hierarchy of mined forgery cues and flexibility. Furthermore, *num* controls our framework's ability to extract forgery cues. We conducted an empirical analysis based on the c23 version and the c40 version of FF++ to study the optimal values of the two hyper-parameters.

We observe that the sensitive patch's window size correlates with *score* negatively. The window size corresponding to the extracted sensitive patch tends to be around *thre*. If *thre* is too small, the percentage of sensitive patches is small and relatively concentrated, which is not conducive to mining forgery cues at different levels. Conversely, it may contain irrelevant regions that lack flexibility and affect the training phase. As *num* increases, initially sensitive patches may capture more forgery cues, but too large may introduce some authentic background regions to the detriment of the final result. As illustrated in Figs. 6a and 6b, we get the best results when setting *thre* to be 4 × 4 and *num* to be 5. Area Under Curve (AUC) reached 0.9961 on the c23 version and 0.9377 on the c40 version.



(a) (b)

Figure 6: The detection performances achieve by (a) varying *thre* when *num* is fixed as 5 and (b) varying *num* when *thre* is fixed as 4 × 4

Components. To demonstrate the effectiveness of each module, we evaluate the proposed framework and its variants by intra-testing (See Table 2) and cross-testing (See Table 3).

Table 2: Ablation study on FF++ (c40) and FaceShifter (c40). The metric is AUC

Model	DF	F2F	FSP	NT	FSR
RGB (Xception)	0.9791	0.9453	0.9441	0.7877	0.9658
RGB + SPB	0.9940	0.9654	0.9862	0.8232	0.9834
RGB + SRM (Baseline)	0.9865	0.9602	0.9778	0.8057	0.9798
Baseline + SPB	0.9956	0.9684	0.9887	0.8415	0.9861
Baseline + SPB + DPFM	0.9969	0.9740	0.9901	0.8443	0.9903
Baseline + SPB + DPFM + NAEM	**0.9982**	**0.9806**	**0.9943**	**0.8564**	**0.9937**

Taking Xception as the Baseline, RGB means RGB modality as the input, spatial rich model (SRM) means noise modality as the input, and Baseline means fusion modality as the input. SPB, DPFM, and NAEM represent the sensitive patch branch, dual-modality progressive fusion module, and noise adaptive enhancement module.

We obtain the following conclusions from this experiment. First, the sensitive patch branch can effectively improve the performance of our framework by fine-grain feature learning in both RGB and fusion modalities. In addition, the improvement is insignificant if sensitive patches are mined directly based on the fusion modality compared to the RGB-only modality. The experimental results demonstrate that our specifically designed functional modules can better capture and utilize noise information. As shown in the last three rows, the model's performance gradually improves as each module is added, demonstrating the effectiveness of each module.

Table 3: Ablation study from FF++(c23) to Celeb-DF. The metric is AUC

Model	Celeb-DF
RGB (Xception)	0.6527
RGB + SPB	0.7451
RGB + SRM (Baseline)	0.7255
Baseline + SPB	0.7629
Baseline + SPB + DPFM	0.7724
Baseline + SPB + DPFM + NAEM	**0.7868**

4.3 Comparison with Recent Works

Within-manipulation-method evaluation. We compare our method with previous detection methods using frequency domain features on FF++ and FSR datasets. The results are shown in Tables 4 and 5. Xception, F^3-Net, SPSL, and Generalizing Face Forgery Detection (GFF) are the most advanced methods for face forgery detection. (Since SPSL is not open source, we are unable to reproduce their method in our comparison experiments)

Table 4: Within-method evaluation of five manipulation techniques (c40). The metric is AUC

Model	DF	F2F	FSP	NT	FSR
Xception† [37]	0.9791	0.9453	0.9442	0.7878	0.9657
F^3-Net† [45]	0.9954	0.9713	0.9879	0.8264	0.9873
GFF† [32]	0.9945	0.9574	0.9592	0.7981	0.9735
SPSL [49]	0.9850	0.9462	0.9810	0.8049	–
Ours	**0.9982**	**0.9806**	**0.9943**	**0.8564**	**0.9937**

Table 5: Within-method evaluation of five manipulation techniques (c23). The metric is AUC

Model	DF	F2F	FSP	NT	FSR
Xception† [37]	0.9956	0.9951	0.9894	0.9523	0.9910
F^3-Net† [45]	0.9995	0.9989	0.9993	0.9884	0.9967
GFF† [32]	0.9942	0.9973	0.9969	0.9691	0.9942
SPSL [49]	–	–	–	–	–
Ours	**0.9999**	**0.9994**	**0.9998**	**0.9933**	**0.9976**

Our framework outperforms other advanced detection methods in terms of the AUC index on different versions of the five manipulation methods. Especially in the most challenging c40 version, the AUC metrics achieve more significant improvements. The experimental results show that our method can effectively capture the features of the manipulation method and achieve the desired detection results.

Within-database evaluation. We compared our method with several state-of-the-art methods. The results are shown in Table 6. Our framework can utilize fusion modality to focus more on manipulated patterns rather than global semantic information. It still has considerable advantages in large datasets composed of multiple manipulation methods.

Table 6: Within-database evaluation on the FF++ dataset with c23 and c40 versions and WildDeepfake dataset. The metric is accuracy (Acc) and AUC

Model	FF++ (c23)		FF++ (c40)		WildDeepfake	
	Acc	AUC	Acc	AUC	Acc	AUC
Fridrich et al. [34]	70.97%	–	55.98%	–	–	–
Cozzolino et al. [35]	78.45%	–	58.69%	–	–	–
Bayar and Stamm [41]	82.97%	–	66.84%	–	–	–
Rahmouni et al. [39]	79.08%	–	61.18%	–	–	–
DSP-FWA [8]	–	0.5750	–	0.6230	–	–
MesoNet [40]	83.10%	–	70.47%	–	–	–
Face X-ray [36]	–	0.8735	–	0.6160	–	–
SPSL [49]	91.50%	0.9532	81.57%	0.8282	–	–
Xception [37]	90.88%†	0.9347†	80.32%†	0.8176†	78.42%†	0.8677†
Add-Net [56]	96.78%	0.9774	87.50%	0.9101	77.01%†	0.8365†
F^3-Net [45]	97.52%	0.9810	**90.43%**	0.9330	80.78%†	0.8756†
FDFL [50]	96.69%	0.9930	89.00%	0.9240	–	–
MAT [30]	97.60%	0.9927	88.69%	0.9040	82.23%†	0.9098†
Ours	**97.83%**	**0.9961**	89.62%	**0.9377**	**83.76%**	**0.9130**

Cross-database evaluation. In the actual situation, we cannot predict the means of face image forgery. Generalizability is also an essential criterion for evaluating detection models. The generalizability of the model directly affects its practical application value. A comprehensive cross-database evaluation is performed in this section to check the generalizability of our method. Our framework is trained on different versions of the FF++ dataset and evaluated on DFD, WildDeepfake, Celeb-DF, and DF1.0. In this section, our method is compared with several advanced methods in terms of generalizability.

Specifically, cross-database evaluation is more challenging due to the difference in the distribution of the training and testing sets. The compression level of DF1.0 is c23. Thus, we do not use DF1.0 to evaluate the generalizability of the method when it is trained on the c40 version. From Table 7, we can observe that our method significantly outperforms the rest of the competitors in almost all datasets.

This advantage of generalizability can be further extended to about 5% to 6% with low-quality images as the training set, as shown in Table 8.

Our method achieves superior results on cross-database evaluation. We also note that supervised learning methods inevitably lead the framework to focus on textures generated by specific manipulation methods. It leads to difficulties in achieving the desired results in the cross-manipulation-method evaluation. Thus, how generalizing forgery cues among unknown manipulation methods is also a problem worth investigating in the future. MLDG [63] and LTW [64] represent Meta-Learning for Domain Generalization and Learning-To-Weight, respectively.

Table 7: Cross-database evaluation from FF++ (c23) to DFD, WildDeepfake, Celeb-DF, and DF1.0. The metric is AUC. Results in gray indicate the within-database performance

Model	FF++(c23)	DFD	WildDeepfake	Celeb-DF	DF1.0
Xception [37]	0.9347	0.8413	0.6617	0.6527	0.6824
EfficientNet-B4 [38]	0.9422	0.8737	0.6140	0.6852	–
Face X-ray [36]	0.8735	0.8560	–	0.7420	0.7230
MLDG [63]	0.9899	0.8814	0.6412	0.7456	–
F^3-Net [45]	0.9810	0.8610	0.6771	0.7121	–
MAT [30]	0.9927	0.8758	0.7015	0.7665	–
LTW [64]	0.9917	0.8856	0.6712	0.7714	–
Local-relation [65]	0.9946	0.8924	0.6876	0.7826	–
GFF [32]	0.9930	0.9190	–	**0.7940**	0.9380
Ours	**0.9961**	**0.9329**	**0.7213**	0.7868	**0.9421**

Table 8: Cross-database evaluation from FF++ (c40) to DFD, WildDeepfake, and Celeb-DF. The metric is AUC. Results in gray indicate the within-database performance

Model	FF++(c40)	DFD	WildDeepfake	Celeb-DF
Xception [37]	0.8176	0.6413	0.6059	0.6218
Add-Net [56]	0.9101	0.5736	0.5421	0.5603
F^3-Net [45]	0.9330	0.6988	0.6039	0.6877
MAT [30]	0.9040	0.7418	0.6549	0.6904
Ours	**0.9377**	**0.8134**	**0.7367**	**0.7830**

5 Visualizations

Class Activation Mapping (CAM). We use Grad-CAM to visualize the attention map and thus explore the regions of interest for the detection method, as shown in Fig. 7. We compare the differences between our method and Xception class activation mapping. We observe that Xception's attention is scattered and focused on a larger area, sometimes exceeding the scope of the forged area. Furthermore, it is unreasonable that Xception focuses on similar areas for different manipulation methods of images. Our framework can find key forgery cues and focus attention evenly on different manipulation methods separately. Our method reaches better detection capability and generalization.

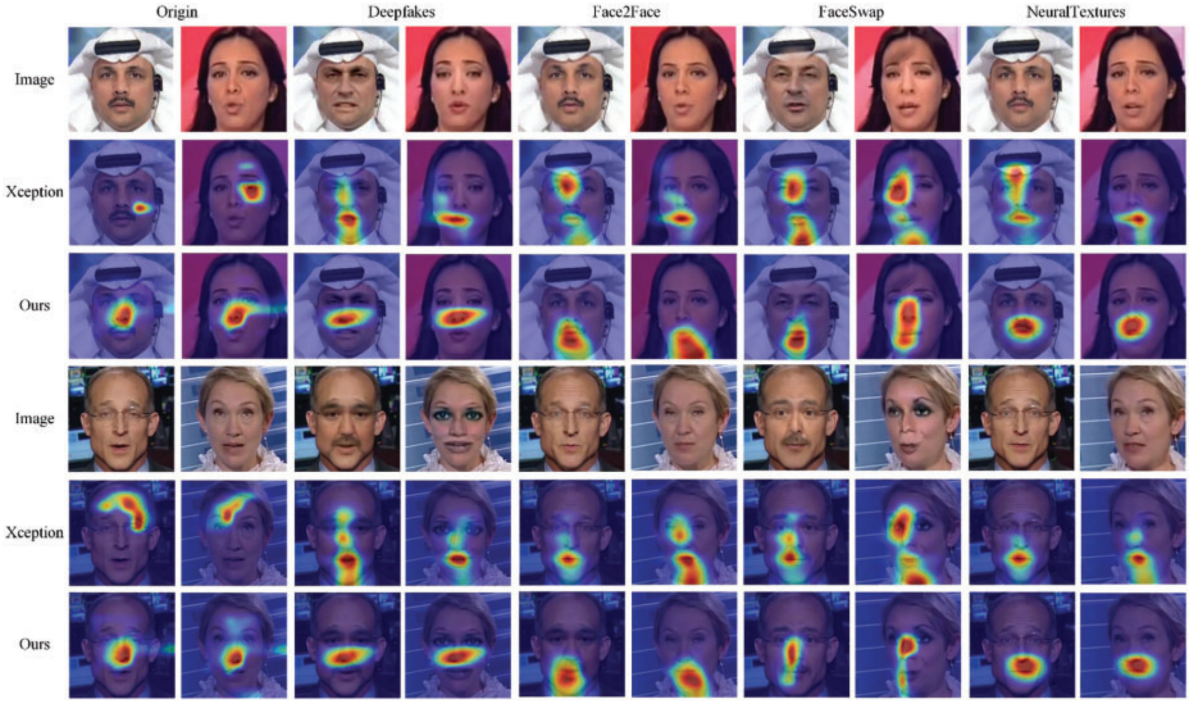


Figure 7: The attention maps of Xception and our framework for different kinds of faces

Adaptive Enhancement. Fig. 8 illustrates the enhanced region of different level feature maps via the noise adaptive enhancement module. We observe that for low-level feature maps, the adaptive enhancement regions are more discrete and vary considerably depending on the input. As for the high-level feature maps, the proposed module focuses on potentially manipulated regions, such as the nose and mouth. This noise adaptive enhancement mechanism can adaptively find and enhance helpful subtle forgery cues among different levels of feature maps.

Sensitive Patch. We visualize the mined sensitive patches in Fig. 9. The sensitive patches are represented as red, orange, yellow, and green. We observe that in RGB modality, the forged features are mainly contained in the low-frequency part, such as irregular color blocks. Furthermore, the high-frequency part plays a more prominent role in the noise modality, such as abnormal face detail features and visual artifacts. There is a spatial correspondence between the two modalities so that they can complement each other to some extent in the fusion modality.

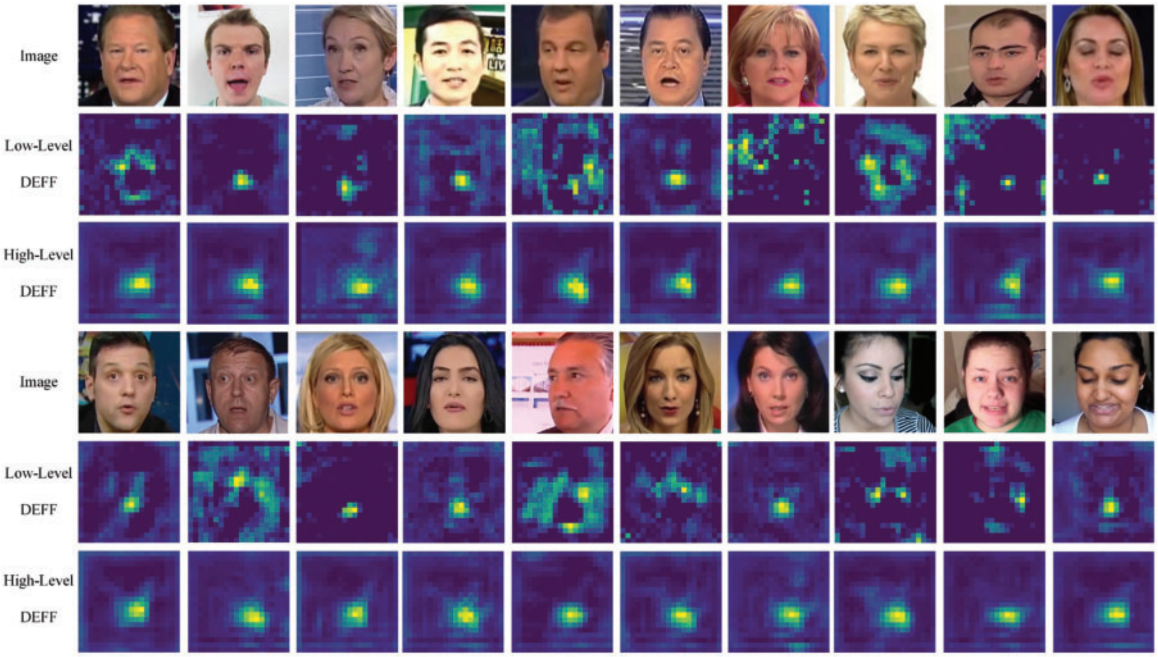


Figure 8: Adaptive enhancement of feature maps at different levels

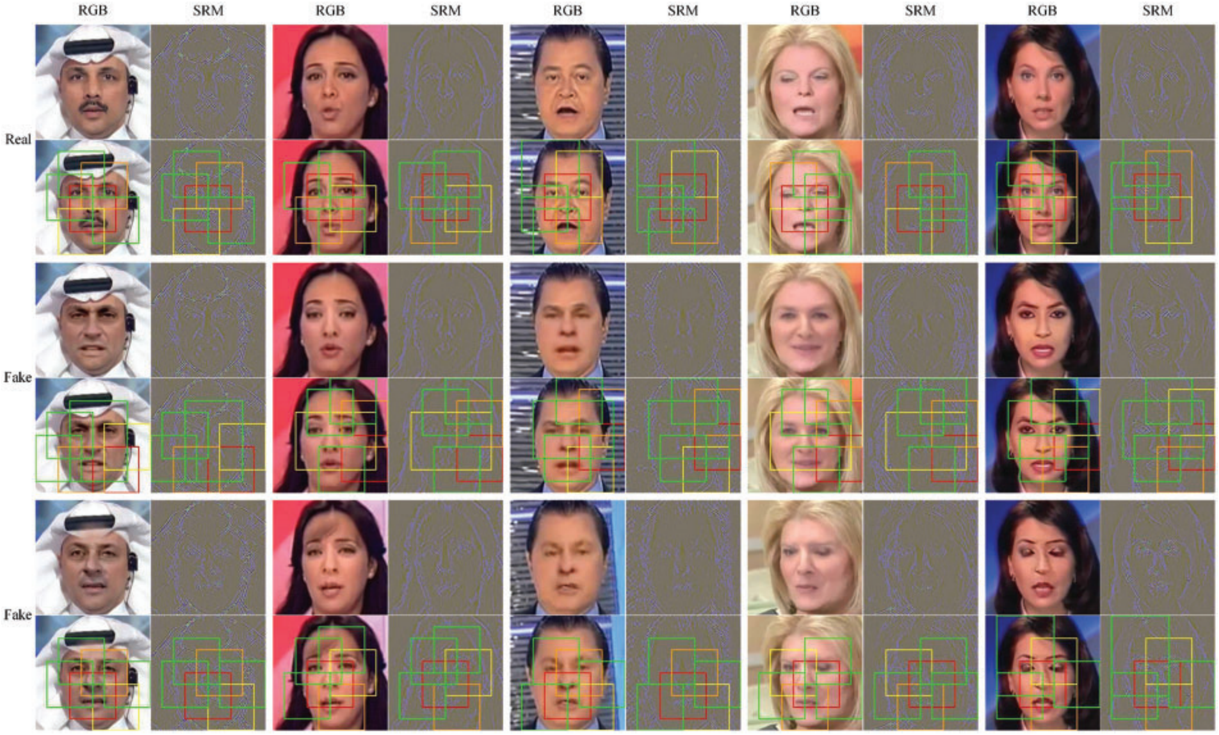


Figure 9: Sensitive patches on RGB modality and noise modality

6 Conclusion

We propose a novel perspective of fine-grain forgery cues mining with fusion modality to address the face forgery detection task. Firstly, a dual-modality progressive fusion module is designed to complement the single-modal features by interacting and fusing the RGB and noise modalities at different scales. A noise adaptive enhancement module is subsequently designed to enhance the subtle noise features in the feature maps of different levels. Learning key manipulated patterns is achieved by mining subtle forgery traces in the sensitive patch branch. Experiments demonstrate that our method has considerable accuracy and generalization advantages. The visualization of class activation mappings, feature maps, and sensitive patches reveals the intrinsic mechanism of our method and explains its effectiveness.

Funding Statement: This study is supported by the Fundamental Research Funds for the Central Universities of PPSUC under Grant 2022JKF02009.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study..

References

[1] I. Kemelmacher-Shlizerman, "Transfiguring portraits," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–8, 2016.

[2] M. R. Koujan, M. C. Doukas, A. Roussos and S. Zafeiriou, "Head2head: Video-based neural head synthesis," in *Proc. of the 15th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, Buenos Aires, Argentina, pp. 16–23, 2020.

[3] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 818–833, 2018.

[4] S. Suwajanakorn, S. M. Seitz and I. Kemelmacher-Shlizerman, "Synthesizing obama: Learning lip sync from audio," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.

[5] MIT TR, "Deepfake Putin is here to warn Americans about their self-inflicted doom," 2020. [Online]. Available: https://www.technologyreview.com/2020/09/29/1009098/ai-deepfake-putin-kim-jong-un-us-election/

[6] CNN, "Deepfake' Queen delivers alternative Christmas speech, in warning about misinformation," 2020. [Online]. Available: https://www.cnn.com/2020/12/25/uk/deepfake-queen-speech-christmas-intl-gbr

[7] BuzzFeed, "How to spot a deep-fake like the Barack Obama-Jordan Peele video," 2018. [Online]. Available: https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed

[8] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," arXiv:1811.00656, 2018.

[9] F. Matern, C. Riess and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proc. of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, Waikoloa Village, HI, USA, pp. 83–92, 2019.

[10] S. Schwarcz and R. Chellappa, "Finding facial forgery artifacts with parts-based detectors," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 933–942, 2021.

[11] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Glasgow, UK, pp. 667–684, 2020.

[12] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li *et al.,* "Local relation learning for face forgery detection," in *Proc. of the AAAI Conf. on Artificial Intelligence*, pp. 1081–1088, 2021.

[13] P. Chen, J. Liu, T. Liang, G. Zhou, H. Gao *et al.,* "Fsspotter: Spotting face-swapped video by spatial and temporal clues," in *Proc. of the 2020 IEEE Int. Conf. on Multimedia and Expo (ICME)*, London, UK, pp. 1–6, 2020.

[14] Y. Ru, W. Zhou, Y. Liu, J. Sun, Q. Li *et al.,* "Bita-Net: Bi-temporal attention network for facial video forgery detection," in *Proc. of the 2021 IEEE Int. Joint Conf. on Biometrics (IJCB)*, Shenzhen, China, pp. 1–8, 2021.

[15] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi *et al.,* "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces (GUI)*, vol. 3, no. 1, pp. 80–87, 2019.

[16] S. Agarwal, H. Farid, O. Fried and M. Agrawala, "Detecting deep-fake videos from phoneme-viseme mismatches," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, Seattle, WA, USA, pp. 660–661, 2020.

[17] K. Chugh, P. Gupta, A. Dhall and R. Subramanian, "Not made for each other-audio-visual dissonance-based deepfake detection and localization," in *Proc. of the 28th ACM Int. Conf. on Multimedia*, Seattle, WA, USA, pp. 439–447, 2020.

[18] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proc. of the 28th ACM Int. Conf. on Multimedia*, Seattle, WA, USA, pp. 2823–2832, 2020.

[19] H. H. Nguyen, J. Yamagishi and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. of the 2019 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, pp. 2307–2311, 2019.

[20] B. Chen, X. Ju, B. Xiao, W. Ding, Y. Zheng *et al.,* "Locally GAN-generated face detection based on an improved Xception," *Information Sciences*, vol. 572, no. 11, pp. 16–28, 2021.

[21] D. A. Coccomini, N. Messina, C. Gennaro and F. Falchi, "Combining efficientnet and vision transformers for video deepfake detection," in *Proc. of the Int. Conf. on Image Analysis and Processing*, Lecce, UK, pp. 219–229, 2022.

[22] R. Durall, M. Keuper and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, pp. 7890–7899, 2020.

[23] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa *et al.,* "Leveraging frequency analysis for deep fake image recognition," in *Proc. of the Int. Conf. on Machine Learning*, pp. 3247–3258, 2020.

[24] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.

[25] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, no. 1, pp. 131–148, 2020.

[26] Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF (v2): A new dataset for DeepFake Forensics," arXiv:1909.12962, 2019.

[27] J. Thies, M. Zollhöfer and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019.

[28] H. Dang, F. Liu, J. Stehouwer, X. Liu and A. K. Jain, "On the detection of digital face manipulation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 5781–5790, 2020.

[29] L. Chai, D. Bau, S. N. Lim and P. Isola, "What makes fake images detectable? understanding properties that generalize," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Glasgow, UK, pp. 103–120, 2020.

[30] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang *et al.,* "Multi-attentional deepfake detection," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 2185–2194, 2021.

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

[32] Y. Luo, Y. Zhang, J. Yan and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 16317–16326, 2021.

[33] J. Fei, Y. Dai, P. Yu, T. Shen, Z. Xia *et al.,* "Learning second order local anomaly for general face forgery detection," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 20270–20280, 2022.

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

[35] D. Cozzolino, G. Poggi and L. Verdoliva, "Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection," in *Proc. of the 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, PA, USA, pp. 159–164, 2017.

[36] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen *et al.,* "Face x-ray for more general face forgery detection," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 5001–5010, 2020.

[37] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 1251–1258, 2017.

[38] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. of the Int. Conf. on Machine Learning*, Long Beach, California, USA, pp. 6105–6114, 2019.

[39] N. Rahmouni, V. Nozick, J. Yamagishi and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in *Proc. of the 2017 IEEE Int. Workshop on Information Forensics and Security (WIFS)*, Rennes, France, pp. 1–6, 2017.

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

[41] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proc. of the 4th ACM Workshop on Information Hiding and Multimedia Security*, Vigo, Galicia, Spain, pp. 5–10, 2016.

[42] A. M. S. Mahdy, "A numerical method for solving the nonlinear equations of Emden-Fowler models," *Journal of Ocean Engineering and Science*, vol. 88, no. 16, pp. 3406, 2022.

[43] A. M. S. Mahdy, K. Lotfy and A. A. El-Bary, "Use of optimal control in studying the dynamical behaviors of fractional financial awareness models," *Soft Computing*, vol. 26, no. 7, pp. 3401–3409, 2022.

[44] A. M. S. Mahdy and M. Higazy, "Numerical different methods for solving the nonlinear biochemical reaction model," *International Journal of Applied and Computational Mathematics*, vol. 5, no. 6, pp. 1–17, 2019.

[45] A. M. S. Mahdy, K. Lotfy, A. El-Bary and I. M. Tayel, "Variable thermal conductivity and hyperbolic two-temperature theory during magneto-photothermal theory of semiconductor induced by laser pulses," *The European Physical Journal Plus*, vol. 136, no. 6, pp. 1–21, 2021.

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

[47] X. Zhang, S. Karaman and S. F. Chang, "Detecting and simulating artifacts in gan fake images," in *Proc. of the 2019 IEEE Int. Workshop on Information Forensics and Security (WIFS)*, Delft, The Netherlands, pp. 1–6, 2019.

[48] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa *et al.,* "Leveraging frequency analysis for deep fake image recognition," in *Proc. of the Int. Conf. on Machine Learning*, pp. 3247–3258, 2020.

[49] Y. Qian, G. Yin, L. Sheng, Z. Chen and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Glasgow, UK, pp. 86–103, 2020.

[50] Y. Bai, Y. Guo, J. Wei, L. Lu and R. Wang, "Fake generated painting detection via frequency analysis," in *Proc. of the 2020 IEEE Int. Conf. on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, pp. 1256–1260, 2020.

[51] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt and M. Niessner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 2387–2395, 2016.

[52] X. Wu, Z. Xie, Y. T. Gao and Y. Xiao, "Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features," in *Proc. of the 2020 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 2952–2956, 2020.

[53] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He *et al.,* "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 772–781, 2021.

[54] J. Li, H. Xie, J. Li, Z. Wang and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 6458–6467, 2021.

[55] G. Wang, Q. Jiang, X. Jin, W. Li, X. Cui *et al.,* "Multimodal contrastive classification by locally correlated representations for effective face forgery detection," *Knowledge-Based Systems*, vol. 250, no. 13, pp. 109088, 2022.

[56] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies *et al.,* "Faceforensics++: Learning to detect manipulated facial images," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Seoul, South Korea, pp. 1–11, 2019.

[57] L. Li, J. Bao, H. Yang, D. Chen and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," arXiv:1912.13457, 2019.

[58] Deepfakedetection, Accessed: 2020-05-10. *Available:* https://ai.googleblog.com/2019/09/contributing-data-to-deepfakedetection.html. 2020.

[59] L. Jiang, R. Li, W. Wu, C. Qian and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 2889–2898, 2020.

[60] B. Zi, M. Chang, J. Chen, X. Ma and Y. G. Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *Proc. of the 28th ACM Int. Conf. on Multimedia*, Seattle, WA, USA, pp. 2382–2390, 2020.

[61] Faceswap, Accessed: 2020-05-10. *Available:* https://github.com/MarekKowalski/FaceSwap. 2020.

[62] J. Deng, J. Guo, E. Ververas, I. Kotsia and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 5203–5212, 2020.

[63] D. Li, Y. Yang, Y. Z. Song and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proc. of the AAAI Conf. on Artificial Intelligence*, Phoenix, Arizona, USA, 32, 2018.

[64] K. Sun, H. Liu, Q. Ye, Y. Gao, J. Liu *et al.,* "Domain general face forgery detection by learning to weight," in *Proc. of the AAAI Conf. on Artificial Intelligence*, pp. 2638–2646, 2021.

[65] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li *et al.,* "Local relation learning for face forgery detection," in *Proc. of the AAAI Conf. on Artificial Intelligence*, pp. 1081–1088, 2021.