Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18836-2

Noise-aware progressive multi-scale deepfake detection

Xinmiao Ding[1] · Shuai Pang[1] · Wen Guo[1]

Received: 13 November 2022 / Revised: 17 January 2024 / Accepted: 28 February 2024

Abstract

The proliferation of fake images generated by deepfake techniques has significantly threatened the trustworthiness of digital information, leading to a pressing need for face forgery detection. However, due to the similarity between human face images and the subtlety of artefact information, most deep face forgery detection methods face certain challenges, such as incomplete extraction of artefact information, limited performance in detecting low-quality forgeries, and insufficient generalization across different datasets. To address these issues, this paper proposes a novel noise-aware multi-scale deepfake detection model. Firstly, a progressive spatial attention module is introduced, which learns two types of spatial feature weights: boosting weight and suppression weight. The boosting weight highlights salient regions, while the suppression weight enables the model to capture more subtle artifact information. Through multiple boosting-suppression stages, the proposed model progressively focuses on different facial regions and extracts multi-scale RGB features. Additionally, a noise-aware two-stream network is introduced, which leverages frequency-domain features and fuses image noise with multi-scale RGB features. This integration enhances the model's ability to handle image post-processing. Furthermore, the model learns global features from multi-modal features through multiple convolutional layers, which are combined with local similarity features for deepfake detection, thereby improving the model's robustness. Experimental results on several benchmark databases demonstrate the superiority of our proposed method over state-of-the-art techniques. Our contributions lie in the progressive spatial attention module, which effectively addresses overfitting in CNNs, and the integration of noise-aware features and multi-scale RGB features. These innovations lead to enhanced accuracy and generalization performance in face forgery detection.

Keywords Deepfake detection · Two-stream network · Spatial attention · Noise-awareness

✉ Wen Guo
wguo@nlpr.ia.ac.cn

Xinmiao Ding
dingxinmiao@126.com

Shuai Pang
plasson007@163.com

[1] School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai 264000, Shandong, China

Published online: 20 March 2024

1 Introduction

Recent years have witnessed rapid progress in facial manipulation [1, 2] based on Variational AutoEncoders and Generative Adversarial Networks. This manipulation enables attackers to change the identity or expression of a target subject to that of another and subsequently produce high-quality forged faces. These manipulation techniques are prone to misuse for malicious purposes, leading to serious security issues and even a crisis of confidence in our society [3]. Therefore, it is crucial to develop more effective and general methods for deepfake detection.

Most existing methods model deepfake detection as a straightforward binary classification problem and develop Convolutional Neural Networks (CNNs) to model the decision boundary between real and fake faces [4–11]. However, the inherent structure of CNN-based models leads to a bias towards emphasizing a single salient region of interest, which makes it challenging to capture other subtle yet distinguishable artifacts. This tendency limits the model's ability to generalize and increases the risk of overfitting [12]. Although some recent approaches that employ multitask learning have made progress in addressing this limitation [13, 14], they still remain susceptible to post-processing techniques such as compression.

Furthermore, some scholars have found that the facial forgery method will eliminate artifacts through post-processing so that they cannot be detected in the color domain, but still leaves tampering traces in high-frequency information. To improve the robustness of the detection model, noise information is introduced as a high-frequency information for forgery detection [15–18].Most of the noise features introduced in current forgery detection approaches are manually extracted, like the spatial rich model (SRM) filters [18] which lack some flexibility in handling some forged videos after complex processing. Moreover, it was also found that real faces are coherent in different local regions, while forged faces are mixed from different face sources and thus produce inconsistent information at certain locations. Therefore, the concept of consistency learning [19] was introduced into forgery detection [20–22], which usually measures the local similarity between individual patches of an image to capture the inconsistency between tampered and authentic regions. However, these methods tend to overlook the importance of global features, which encompass valuable discriminative information such as the colors of the artifacts in different facial regions and the contextual links between individual artifacts. In scenarios where captured artifacts are scattered across multiple local regions, the local inconsistency may not be prominent. Nevertheless, the combined statistical information from these artifacts can exhibit stronger discriminative properties at a global level. Consequently, incorporating global feature as a complement to local consistency information can improve the performance of the classifier.

To overcome the aforementioned challenges, we propose a noise-aware progressive multiscale network. Firstly, we address the issue of overfitting and capture more comprehensive features by suppressing features from salient regions during training and mining features from other regions. This is achieved through the design of a progressive spatial attention module, which incorporates a boosting-suppression mechanism. Secondly, we enhance the sensitivity of the high-frequency noise filter by using adaptive learning-based SRM filters (Conv-SRM) instead of fixed-parameter SRM filters. This adaptation makes the Conv-SRM filters more effective in detecting certain artifacts, as illustrated in Fig. 1. Additionally, considering that the phase spectrum [24] is also sensitive to artifacts resulting from up-sampling operations [23] just as shown in Fig. 1, we introduce the phase spectrum as a supplement of noise information. To leverage both RGB and noise information, we construct a two-branch network that learns a composite representation by combining the RGB features from various boosting regions

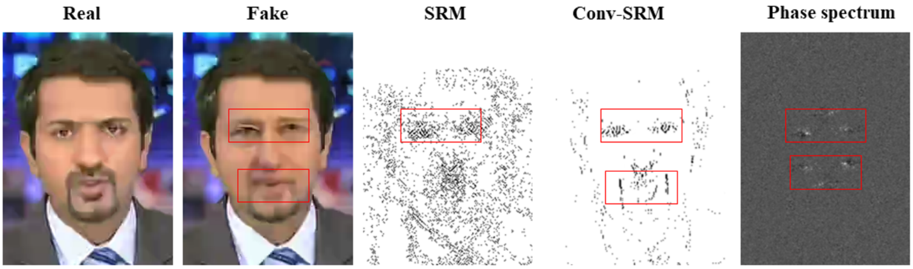


Fig. 1 Comparison of different filters on real and fake faces. Red boxes from fake faces mark artifacts that are biased to eyes and mouth. The Conv-SRM filter and the phase spectrum of the face can capture the eyes and mouth of the tampered face more obviously

with high-frequency features. Furthermore, we merge the local consistency information with global features learned from the multi-modal features for deepfake detection, enhancing the model's robustness.

The contributions are summarized as follows:

- A progressive spatial attention module is proposed to address the overfitting issue. It forces the model to explore more potential features and enables the network to focus on the salient region of faces.
- Noise features are extracted by adaptive learning instead of hand-crafting, and phase features which are more sensitive to forged faces after post-processing are introduced.
- Global information is added to collaborate with local consistency information to detect deepfake and improve the model's discriminatory ability.

2 Related work

Currently, various deep forgery detection methods are emerging, with early attempts focusing on detecting forged faces through manual features [4, 13, 25]. With the widespread use of deep learning methods [26–30] such as: a multitask manifold deep learning method effectively used to estimate face-pose [26], hierarchical deep click feature prediction for better fine-grained Image recognition [28], multimodal deep autoencoder and multi-view locality sensitive sparse retrieval successfully applied to human pose recovery [29, 30] and so on. Researchers are beginning to explore the use of deep neural networks to capture high-level semantic features in the spatial domain of forged images to improve detection. Most works [5, 7–10, 31] use a CNN to extract discriminative features for forgery detection. Nguyen et al. [5] designed a network model combining a VGG network and a capsule network to detect fake faces. Dang et al. [7] proposed a detection system based on CNN and attention mechanisms to process and improve the feature maps of the classifier model. Afchar et al. [8] proposed MesoNet combined with the Inception module to extract mesoscopic features and detect forged videos. Rossler et al. [9] constructed a deep forgery dataset FaceForensics++ and used seven convolutional networks to compare the classification performance of real and fake faces, among which XceptionNet gave the best results. Kohli et al. [31] extracted facial features from the frequency domain using a two-dimensional global discrete cosine transform (2D-GDCT) and used a three-layered frequency convolutional neural network (fCNN) to detect forged facial images. Although these works have achieved a significant performance level, they are easily overfitting and perform worse on some low-quality databases as the CNN-based method tends to emphasize more on a single salient region of interest.

Forgery detection with noise features. Several attempts [15–17] have been made to solve forgery detection based on noise features. Zhou et al. [15] used SRM filter based on steganalysis method to extract noise information as the input of noise stream together with RGB stream for forgery detection. Masi et al. [16] presented a two-branch network: one branch propagates the original information, while the other branch suppresses the face content yet amplifies multi-band frequencies using a Laplacian of Gaussian (LoG) as a bottleneck layer to assist in isolating the forged faces. Qian et al. [17] explored two complementary frequency-aware clues including frequency-aware decomposed image components and local frequency statistics to mine subtle forgery patterns. Although these studies extract features more comprehensively, the filters used in these works are often fixed, and the single type of high-frequency information cannot guarantee strong generalization.

Forgery detection based on consistency learning. Recent works [19–22] show that the manipulation methods typically disrupt the correlation between the local regions of the faces and attempt to utilize consistency learning to capture the local artifacts. Zhao et al. [20] extracted the middle layer features of the ResNet network and constructed patch similarity features based on them, which were used to assist in locating the forged regions and guide the model to detect the local inconsistencies of the forged faces. Chen et al. [21] constructed the similarity matrix of the frequency domain stream and the RGB stream to capture the local inconsistencies of the forged faces in both the spatial and frequency domains. Kuang et al. [22] proposed a dual-branch (spatial branch and temporal branch) neural network to detect the inconsistency in both spatial and temporal for DeepFake video detection. The spatial branch aims at detecting spatial inconsistency by the effective EfficientNet model. The temporal branch focuses on temporal inconsistency detection by a new network model. The softmax scores of two branches are finally combined with a binary-class linear SVM classifier. However the independent learning of each branch loses spatio-temporal contextual information, and the non-end-to-end leaning make it difficult to co-optimise the two branches. In addition, the above methods based on consistency learning only use local information for their final discrimination, ignoring the global information that is also discriminative.

3 Proposed method

3.1 Overview

The proposed framework is shown in Fig. 2. It is a two-stream network consisting of RGB stream and noise-aware stream and contains three main modules as follows:

Progressive Spatial Attention Module (PSAM). This module achieves feature extraction of the RGB stream, and its core is the Dual Feature Selective Module (DFSM). The DFSM can adaptively learn two weights of the spatial features, namely boosting and suppression weights. Multiplying by the original feature, the boosting and suppression features are respectively obtained. The boosting features prepare for the subsequent fusion with noise features, while the suppressed features are input again to the next stage of DFSM after convolution operations, forcing the model to mine potential features in regions other than the currently significant ones. Three different scales of boosting features are obtained from three DFSM.

Noise-aware Module (NAM). As for the noise stream, the adaptive noise features are extracted by the Conv-SRM filter and combined with the phase spectrum features as the input.

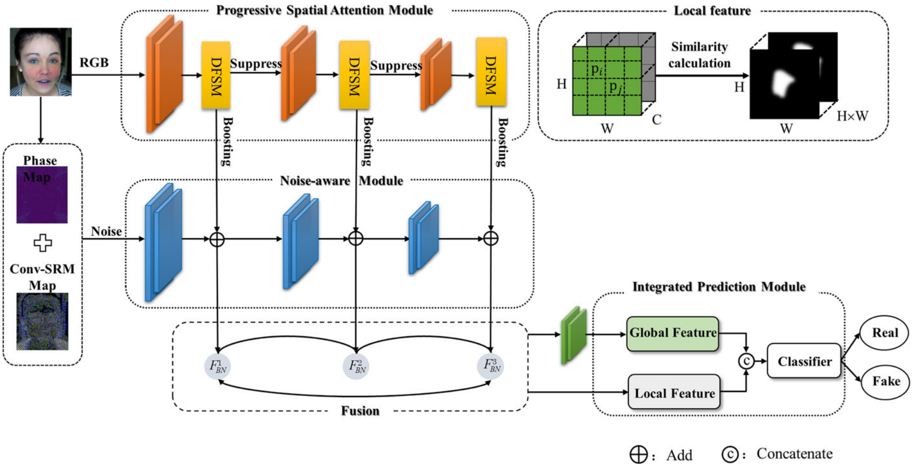


Fig. 2 Pipeline of the proposed Noise-aware Progressive Multi-scale Deepfake Detection

Then, the outputs of three convolutional layers are combined with boosting features obtained from the corresponding stage of the PSAM to produce dual-stream multi-scale features.

Integrated Prediction Module (IPM). The multi-scale features leaned from NAM are pairwise fused and then fed into the convolution block to obtain global features. Meanwhile they are divided equally into several local patches to produce local consistency information by calculating the similarity among the patches. Finally, the global features and local consistency information are combined to achieve the prediction.

3.2 Progressive spatial attention module

The pipeline of the core module DFSM is shown in Fig. 3. In DFSM, the boosting and suppression weights are computed to serve as the basis for selecting boosting and suppression features, respectively. More formally, denote the input face image of the RGB stream as I and the feature map extracted from the specific layer as X of dimensions $H \times W \times C$, where C is the number of channels, whereas H and W are the height and width of the feature map, respectively. The computation process of the boosting/suppression weights is as follows:

First, X is evenly divided into k parts along the width dimension denoted as $X_i \in \mathbb{R}^{H\times(W/k)\times C}, i \in [1, k]$. Second, these parts are input to the 1×1 convolution block for dimensionality reduction, and the output is activated through the normalization layer and ReLu to obtain the initial weight of each part as $S_i \in \mathbb{R}^{H\times(W/k)\times 1}$. Inspired by the Convolutional Block Attention Module (CBAM) [32], the normalized weight $S_{softmax} = (s_1, ..., s_L)$, $L = W/k$ are obtained by respectively fusing the results of average-pooling and max-pooling on S_i.

Then the boosting weight $W_{boost} = \left(w_{b_1}, ..., w_{b_L}\right)$ is calculated as follows:

$$w_{b_i} = \begin{cases} \alpha, \; if \; s_i \geq T \\ 0, \quad else \end{cases}, \tag{1}$$

where $i \in [1, L]$, T is the threshold taken from the n^{th} largest weight in $S_{softmax}$, and the default setting of n is $L/4$. If s_i is greater than or equal to the threshold, it can be regarded as a boosting spatial part and its weight is set to the hyper-parameter α which controls the extent

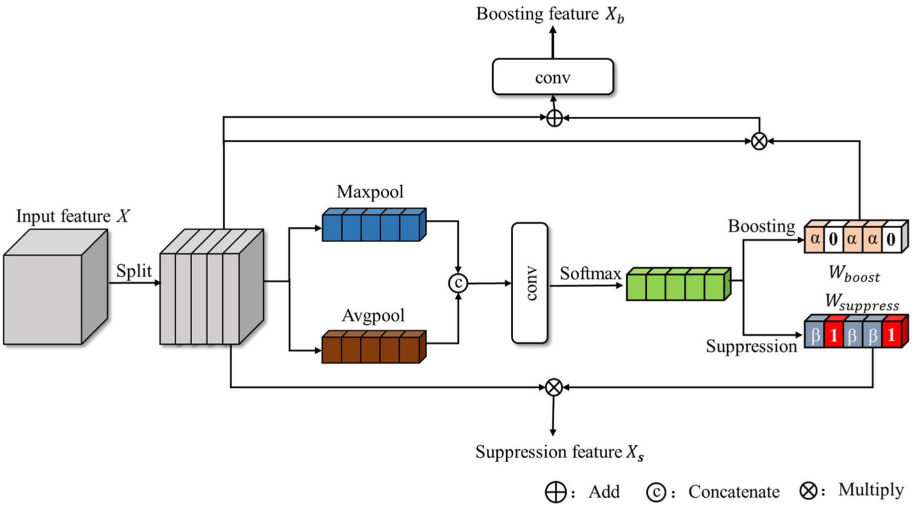


Fig. 3 Pipeline of the Dual Feature Selective Module

of boosting. Otherwise, it is set to 0 and it is considered that no boost is applied to this part. Similarly, the suppression weights $W_{suppress} = \left(w_{s_1}, ..., w_{s_L}\right)$ are calculated as follows:

$$w_{s_i} = \begin{cases} 1, & if\ s_i < T \\ \beta, & else \end{cases}, \tag{2}$$

where β is a hyper-parameter that controls the extent of suppression. The effect of α and β on the performance will be discussed in Section 4.2.

Finally, the boosting features X_B and suppression features X_S are obtained by fusing W_{boost} and $W_{suppress}$ with the feature map X as following:

$$X_B = X + (W_{boost} \otimes X), \tag{3}$$

$$X_S = W_{suppress} \otimes X, \tag{4}$$

where $\otimes$ represents the element dot product. The boosting features X_B will be directly used as the output of DFSM for subsequent feature fusion, and the suppression features X_S will be put into the next boosting-suppression stage to learn potential boosting features in other regions. Learned through several such stages, the boosting features at different scales can be obtained.

3.3 Noise-aware module

3.3.1 Acquisition of phase spectrum and adaptive noise features

After inputting face image I and converting it to grayscale map $I_{gray} \in \mathbb{R}^{H \times W \times 1}$ by an image processing algorithm, the phase and adaptive noise features are obtained as follows:

Phase spectrum. With Discrete Fourier Transform (DFT), the grayscale map I_{gray} is first transformed into the frequency domain to obtain its frequency spectrum. The Inverse Discrete Fourier Transform (IDFT) is then applied with the frequency spectrum without amplitude to obtain the spatial domain representation of the phase spectrum as $I_{phase} \in \mathbb{R}^{H \times W \times 1}$.

Adaptive noise features. A constrained convolution layer [33] is introduced and the kernel parameters of the Conv-SRM are adaptively updated through network training. Specifically, the constraint is applied as follows:

$$\begin{cases} C_k(0,0) = -1 \\ \sum_{m,n\neq 0} c_k(m,n) = 1 \end{cases} \quad k \in \{1,2,3\}, \tag{5}$$

where c_k represents the k^{th} convolution channel updated with the model parameters and $(0,0)$ is its central coordinate. The adaptive noise features $I_{srm} \in \mathbb{R}^{H\times W\times 3}$ of the image can be obtained by inputting I_{gray} into the constraint convolution layer.

3.3.2 Complementary fusion of noise and RGB streams

As shown in the noise stream of Fig. 2, the phase features I_{phase} are concatenated with the adaptive noise features I_{srm} in the channel, resulting in 4-channel noise features I_{noise} which are then input into the noise stream, and the noise features at different scales are obtained after different layers of convolution blocks. To enhance the model's ability to detect artifacts, the fusion of noise features with the corresponding scaled RGB boosting features are added to the output of each convolutional block of the noise stream, which is then fed into the next convolutional block for higher-level feature learning.

Denote the features obtained by the three convolutions of the noise stream as $X_N^{(i)}, i \in \{1,2,3\}$ and the corresponding boosting features obtained by the PSAM as $X_B^{(i)}, i \in \{1,2,3\}$. These feature maps are then flattened into two-dimensional vectors along the spatial dimension as $\tilde{X}_N^{(i)}$ and $\tilde{X}_B^{(i)}$, respectively. Inspired by self-attention [34], the dual-stream complementary fusion is as follows:

$$F_{BN}^{(i)} = X_B^{(i)} + \rho \times att_{BN}, \tag{6}$$

$$F_{NB}^{(i)} = X_N^{(i)} + \rho \times att_{NB}, \tag{7}$$

where ρ is the hyper-parameter controlling the fusion level, att_{BN} and att_{NB} represents the complementary weights which are calculated as follows:

$$att_{BN} = softmax\left(\frac{\tilde{X}_B^{(i)}\tilde{X}_B^{(i)^T}}{\sqrt{H\times W\times C}}\right)\tilde{X}_N^{(i)}, \tag{8}$$

$$att_{NB} = softmax\left(\frac{\tilde{X}_N^{(i)}\tilde{X}_N^{(i)^T}}{\sqrt{H\times W\times C}}\right)\tilde{X}_B^{(i)}, \tag{9}$$

The $F_{NB}^{(i)}$ obtained from (7) will be input to the next convolutional block to learn higher-level noise features $X_N^{(i+1)}$. The $F_{BN}^{(i)}$ obtained from (6) is left as the input feature map for the following integrated prediction module, which incorporates the complementary features to provide more detailed information.

3.4 Integrated prediction module

Given the remarkable performance achieved by consistency learning in forgery detection, it is also introduced into the final prediction module and further fused with global information learned from complementary features for discrimination.

3.4.1 Global feature extraction

To further enhance the feature representation capability of the input integrated prediction module, the different scale complementary features $F_{BN}^{(i)}$ obtained from (6) are fused with each other so that the features at each scale aggregate discriminative information from other scales and fully combine the contextual information of the images. A fusion method similar to that in Section 3.3.2 is used here to produce new multi-scale complementary features $F_{BN}^{*(i)}$ as follows:

$$F_{BN}^{*(i)} = F_{BN}^{(i)} + \sum_{j \neq i} \mu \times att^{(i,j)} \quad i, j \in \{1, 2, 3\} \tag{10}$$

where μ is used to control the fusion level, and $att^{i,j}$ represents the complementary weights of $F_{BN}^{(j)}$ for $F_{BN}^{(i)}$ which are calculated as follows:

$$att^{i,j} = softmax\left(\frac{\tilde{F}_{BN}^{(i)} \tilde{F}_{BN}^{(i)^T}}{\sqrt{H \times W \times C}}\right) \tilde{F}_{BN}^{(j)}, \tag{11}$$

where $\tilde{F}_{BN}^{(i)}$ and $\tilde{F}_{BN}^{(j)}$ are the two-dimensional vectors that flattened $F_{BN}^{(i)}$ and $F_{BN}^{(j)}$ along the spatial dimension, respectively. As the feature map $F_{BN}^{*(i)}$ is complementarily fused in spatial and frequency domains and at different scales, it can highlight the artifact region at the different granularities and supplement the detail information from other scales and noise features. A bilinear interpolation algorithm is used to make the feature maps equal in size. They are then concatenated along the channels and learned by a certain number of convolutional layers to obtain the features $\hat{F}_{BN}$ which can be regarded as a global feature covering different artifacts of concern at different scales.

3.4.2 Local consistency calculation

Local similarity calculation is introduced to measure local consistency. The robust multi-scale complementary feature map $F_{BN}^{*(i)}, i \in \{1, 2, 3\}$ obtained in (10) is applied as input and the bilinear interpolation algorithm is used to resize the different scale features into the same dimensions. They are then concatenated into $Z \in \mathbb{R}^{H \times W \times C}$ in the channel. Z is equally divided into $M \times M$ patches along the spatial dimension and flattened into a one-dimensional vector. Denote the one-dimensional vectors obtained from the i^{th} and j^{th} patches as $p_i, p_j \in \mathbb{R}^{M^2 C}$, respectively. The similarity $sim_{i,j}$ between the patches is calculated based on the dot product as follows:

$$sim_{i,j} = \delta\left(\frac{p_i p_j}{\sqrt{C}}\right), \tag{12}$$

where δ is the Sigmoid activation function, and the value of $sim_{i,j}$ is between 0 and 1; the closer to 1, the higher the similarity between the two patches, otherwise, the lower the similarity. The multi-scale local features F_{sim} can be constructed by iteratively calculating the similarity between each patch and all other patches.

3.4.3 Integrated Prediction

To improve the classifier's decision performance, $\hat{F}_{BN}$ is flattened into a one-dimensional vector together with the local similarity F_{sim}, and the final prediction probability $\hat{y}$ is then obtained through the fully connected layer followed by the Sigmoid activation function. The cost function uses the cross-entropy loss as follows:

$$loss\left(y, \hat{y}\right) = -\left[ylog\hat{y} + (1-y)log(1-\hat{y})\right], \tag{13}$$

where y is set to 1 if the input image is a forged face, otherwise it is set to 0. During the training process, the loss back propagation drives the network to learn the difference between real and forged faces.

4 Experiments

4.1 Settings

Datasets. Following recent works of face forgery detection, the benchmark dataset FaceForensics++(FF++) is used for evaluation. The dataset consists of 740 videos for training, 140 videos for validation, and 140 videos for testing. There are four versions of FF++ in terms of common face manipulation methods, i.e., DeepFakes (DF) [35], Face2Face (F2F) [36], FaceSwap (FS) [37] and NeuralTextures (NT) [38]. Additionally, FF++ consists of three versions of compression level, i.e., raw, lightly compressed (HQ), and heavily compressed (LQ). We adopt the HQ version by default and otherwise specify the version. To evaluate the robustness of our method, we also conduct experiments on the recently proposed large-scale face manipulation datasets, i.e., Celeb-DF [39], DeepfakeDetection (DFD) [9], DeepForensics-1.0 (DF1.0) [40].

Evaluation metrics. In our experiments, the Accuracy rate (ACC) and the Area Under Receiver Operating Characteristic Curve (AUC) are mainly used as evaluation metrics. A higher ACC or AUC value indicates better performance. ACC is used as the major evaluation metric while AUC is adopted to evaluate the performance on cross-dataset. Parameter count (Params) and the number of floating-point operations (FLOPs) are additionally introduced to measure the model's computational workload.

Implementation. All experiments were done in Ubuntu 16.04 operating system equipped with 12 GB of RAM and a GTX 2080 Ti GPU. The proposed framework is implemented in Pytorch 1.3.0 with a configuration environment of CUDA10.2, CUDNN7.6.5. The backbone network used for the RGB streams is the EfficientNet-B4 [41] model pre-trained on ImageNet, and the noise branch is a custom stacked convolutional block. The MTCNN [42] is used for face extraction and alignment, and the aligned faces are resized to 256 × 256. The model has a batch size of 32 and an iteration epoch number of 30. The RAdam optimizer is used to train the model with a learning rate of 0.002 and a weight decay of 0.0005. The learning rate is adjusted using the cosine annealing algorithm. The model hyper-parameters α, β, ρ and μ are set by cross-validation on the training set with regard to the value of ACC and AUC in each procedure.

4.2 Ablation Study

In order to analyze the individual contributions of each module in the proposed model, we conduct experiments using the EfficientNet-B4 model as a backbone. The performance metrics of the algorithm on the low-quality (LQ) version of the FF++ dataset are evaluated when progressively adding the PSAM, the NAM, and the IPM respectively. This allows us

Table 1 Comparison of the performance gains generated by each improvement strategy (%)

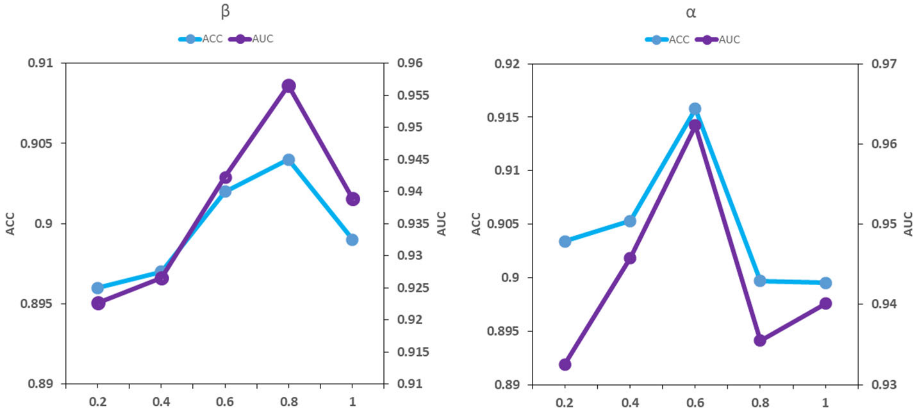
EfficientNet-B4	PSAM	NAM	IPM	ACC	AUC
✓	-	-	-	88.67	91.20
✓	✓	-	-	90.73	94.71
✓	-	✓	✓	89.72	92.68
✓	✓	✓	-	91.09	95.27
✓	✓	✓	✓	**91.65**	**96.34**

Bold fonts highlight the best results of the experiment

to examine in detail the performance gains generated by each module in detail. The results are shown in Table 1.

As can be seen from Table 1, each module contributes to the performance improvement, with the progressive spatial attention module contributing the most, increasing the model performance gain by almost 2%. This is mainly due to its ability to force the model to learn other discriminative regions and boost the features of useful ones (this property will be further discussed in subsequent sections using Grad-CAM [43] visualization). The model's performance in dealing with post-processed videos can be further improved by adding noise-aware and integrated prediction modules. Finally, combining all the modules, the proposed method achieves a better performance with an ACC and AUC of 91.65% and 96.34%, respectively.

In addition, as shown in (1) and (2), different boosting weights α and suppression weights β will have different effects on the model's performance. Specifically, the larger α is, the greater the weight assigned to the boosting spatial part. The smaller β is, the greater the suppression level of the boosting spatial part. To investigate the effects of the two hyper-parameters on the model, an experimental analysis was conducted using the LQ version of the FF++ dataset, as shown in Fig. 4, where Fig. 4a shows the ACC and AUC variation curves with respect to the suppression weight β without boosting, and Fig. 4b shows the curves variations with respect to α when the suppression weight β is fixed to the optimal value. From Fig. 4a, it is known that when β is small, i.e., when the suppression level in the boosting spatial part is large, more



(a)

(b)

Fig. 4 Effects of different boosting-suppression weights on the detection performance of the model. (a) Variation of detection metrics with suppression weight when there is no boosting weight. (b) Effect of boosting weight on the model when the suppression weight is fixed

useful information might be lost, which has a greater impact on the model's performance. As β increases, the useful information is retained, and the optimum is finally reached at about 0.8. Fig. 4b indicates that appropriately assigning larger weights to the features of the boosting spatial part can effectively improve the performance of the model, while excessive boosting might introduce additional noise leading to performance degradation. Therefore, α is uniformly taken as 0.6 and β as 0.8 in the following experiments.

4.3 Visualization of PSAM

To enhance the interpretability of the progressive spatial attention module, the Grad-CAM visualization method is adopted to visualize the feature maps output by both the proposed model and EfficientNet-B4 at the same scale. The decision basis of the model is significantly displayed by representing the regions of interest in the pictures as heat maps, as shown in Fig. 5.

It can be observed that although the EfficientNet-B4 model can also capture artifact regions, such as nose and eye regions, it tends to focus excessively on a certain part in the subsequent stage of the model and ignore the effective information of other regions, thus easily causing overfitting issues. Comparatively, the proposed model can effectively address this issue, as shown in the first row of Fig. 5. The proposed model tends to focus on the eye region in the first stage, followed by the suppression operations in the DFSM so that the network captures the nose region, which is also discriminative in the second stage. Similarly, the network focuses on the fusion boundary region in the last stage (Stage 3). Comparatively, the EfficientNet-B4 model is limited to detecting the nose region after stage 2 of the network and is less robust. Therefore, the proposed model not only captures different artifact regions at different scales but also better fits the characteristics that the artifacts of actual forged images tend to be concentrated in key regions such as eyes, nose, and mouth. These regions are the essential discriminative information for the model to perform subsequent learning.

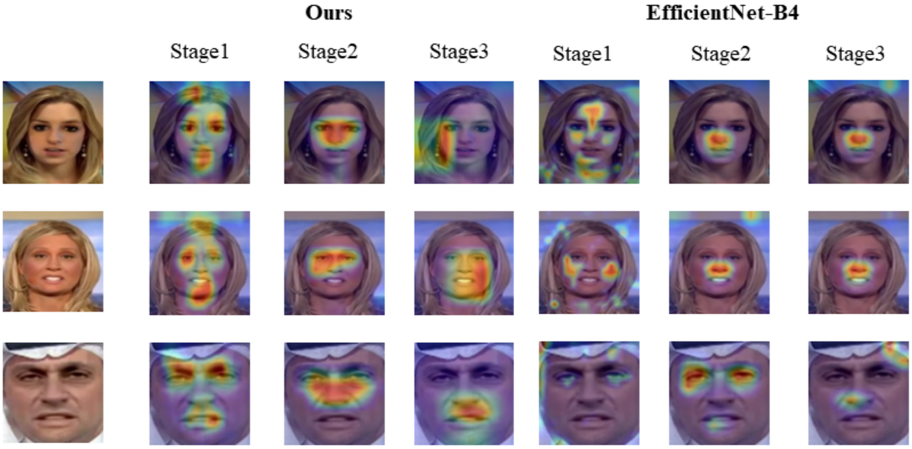


Fig. 5 Visualized heat map at different scales

4.4 Comparison with recent works

To demonstrate the effectiveness of the proposed model, three types of forgery detection methods are selected to compare their performance with the proposed model using HQ and LQ versions of the FF++ dataset. The forgery detection techniques used for comparison are:

(1) Conventional CNN methods, including LD-CNN [10], MesoNet [8] and Xception [9].
(2) Forgery detection methods that use noise features to overcome the effects of post-processing, including Steg.Features [18], Two-branch [16], and F^3-Net [17].
(3) Forgery detection methods based on local correlation, including Face X-ray [44] and LRLF [21].

The comparison results are shown in Table 2. It can be seen that:

Although conventional CNN methods, such as MesoNet and Xception, achieve great results on the HQ version of FF++ dataset, their performance drops significantly on the heavily compressed LQ dataset, while the proposed model achieves remarkable performance on both HQ and LQ datasets and maintains better and stable performance on the LQ version of the dataset. This result is not only due to the progressive spatial attention module that captures the latent artifacts from different boosting regions, but also due to the noise information introduced to overcome the impact caused by post-processing.

Compared with models based on noise features, such as Steg.Features and Two-branch, the proposed model can capture more imperceptible artifacts through adaptive noise features and phase features, thus achieving better performance on datasets of different quality.

Compared with the state-of-the-art LRLF, the proposed model outperforms LRLF in ACC metrics by 0.18% and 0.26% on two different quality FF++ datasets, respectively. Compared with LRLF, which only uses local similarity features for forgery detection, the proposed model not only adds a progressive learning feature extraction method to enrich feature information at different scales, but also captures both local and global information, which has better robustness.

Compared the results of FLOPs and Params, MesoNet achieves the best performance due to its lightweight structure. However its limited network structure hinders it from capturing deeper semantic information, resulting in lower ACC and AUC values. In comparison, the model presented in this paper achieves moderate performance. The computational efficiency,

Table 2 Comparison of ACC, AUC, FLOPs and Params under different quality FaceForensics++ datasets (%)

Methods	LQ		HQ		FLOPs(G)	Params(M)
	ACC	AUC	ACC	AUC		
Steg.Features	55.98	-	70.97	-	20.10	84.05
LD-CNN	58.69	-	78.45	-	-	-
MesoNet	70.47	-	83.10	-	**0.11**	**1.31**
Face X-ray	-	61.60	-	87.40	-	-
Two-branch	86.34	86.59	96.43	88.70	-	-
Xception	86.86	89.30	95.73	-	6.01	22.86
F^3-Net	90.43	93.30	97.52	98.10	16.91	41.62
LRLF	91.47	95.21	97.59	99.46	-	-
Ours	**91.65**	**96.34**	**97.85**	**99.60**	13.81	53.24

Bold fonts highlight the best results of the experiment

Table 3 Comparison of ACC on datasets from different manipulation methods of FF++(LQ) (%)

Methods	DF	FF	FS	NT
Steg.Features	67.00	48.00	49.00	56.00
LD-CNN	75.00	56.00	51.00	62.00
MesoNet	90.00	83.00	83.00	75.00
Xception	96.36	86.86	90.29	80.67
F^3-Net	97.97	**95.32**	96.53	83.32
LRLF	98.12	95.25	97.05	**84.22**
Ours	**98.55**	94.61	**97.32**	82.59

Bold fonts highlight the best results of the experiment

as measured by FLOPs, is fair and the number of parameters is relatively large for end-use applications. This is primarily due to the numerous self-attention computations involved in fusing multi-scale and multi-modal features. Therefore, there is room for significant improvement in this aspect.

To further verify the detection effectiveness of the proposed model for different manipulation methods, comparative experiments were also conducted on four forgery subsets of FF++ (DF, FF, FS, NT), which were generated using different forgery methods. The LQ version of the FF++ dataset is adopted for the experiments, and the results are shown in Table 3.

As shown in Table 3, the performance of the proposed model outperforms most of the advanced methods and demonstrates its robustness in dealing with different forgery methods. However, compared to the best performance on the full face swap datasets DF and FS, the performance of the proposed model is still inferior on the expression swap datasets FF and NT datasets. The reason is that the proposed progressive spatial attention module forces the model to continuously mine hidden artifacts outside the boosting region, and after multiple boosting-suppression stages, the model tends to mine more than one suspected artifact regions. That is beneficial for full-face replacement containing multiple artifacts, whereas for expression replacement, which is often limited to the replacement of a certain attribute of the face, such as eyes or mouth, over-mining of artifacts may introduce noise and ultimately affects the detection results.

4.5 Generalization Ability Evaluation

The generalization ability of the proposed model is verified by training the model on FF++ dataset and testing it on Celeb-DF, DFD, and DF1.0 datasets. The results are shown in Table 4. Three additional methods focusing on generalisability were added to this experiment for comparison, including: Hybrid model [45], Cross-Modality [46] and MA Localization [47].

Celeb-DF, DFD, and DF1.0s are categorized as the second-generation deep forgery dataset. Compared with FF++, they adopt different forgery synthesis methods, which not only significantly improve the synthesis quality of faces, but also make the artifacts less obvious. At the same time, fake faces under different conditions are also taken into account. For example, different capture scenes (indoor and outdoor), different light conditions (day and night), the distance between the video subject and the camera, and head posture changes. Consequently, performing generalization experiments on such datasets is challenging.

As can be seen from the Table 4, our model's generalization results are significantly better than other detection methods. This advantage is mainly due to the following two points: (1) CNN-based detectors tend to overfit to method-specific color textures and thus fail to

Table 4 Cross-dataset evaluation on Celeb-DF, DFD and DF1.0 (AUC (%))

Methods	FF++	Testing AUC		
		Celeb-DF	DFD	DF1.0
MesoNet	84.70	54.80	-	-
Capsule	96.60	57.50	-	-
Xception	99.70	68.29	84.86	71.95
Two-branch	93.18	73.41	-	-
Face X-ray	99.70	74.20	85.60	72.30
F^3-Net	99.30	65.17	87.41	72.59
LRLF	99.80	77.25	89.54	74.58
Hybrid model	99.50	77.40	-	-
Cross-Modality	99.80	76.90	-	-
MA Localization	95.70	67.20	-	-
Ours	**99.90**	**78.47**	**91.55**	**76.87**

Bold fonts highlight the best results of the experiment

generalize. The multi-scale noise information obtained from the noise stream removes color textures and reveals discrepancies between authentic and counterfeit regions. It gives the model a stronger ability to detect post-processing forged faces. As on the DF1.0 dataset with multiple post-processing, the proposed model can still achieve 2-4% higher detection results than the F^3-Net and LRLF. (2) The PSAM prevents the model from focusing excessively on a single region while ignoring potential artifacts in other regions, and improves the model's ability to detect invisible artifacts.

5 Conclusion

To address the issues of poor generalization and low accuracy in existing deepfake detection models, this study proposes a noise-aware multi-scale deepfake detection model that focuses on three key aspects: progressive artefact mining, multi-scale multi-model feature fusion, and joint prediction based on global information and local similarity. The proposed progressive spatial attention module can generate multi-scale features by progressively focusing on different salient regions with a boosting-suppression strategy to enable the network to effectively explore subtle fake features. In addition, the noise-aware dual-stream network integrates adaptive noise features and phase spectrum with RGB multi-scale features to improve the model's ability to handle the effects of post-processing. The robustness of the model is further enhanced by the combined discrimination of local consistency and global features. Experiments on widely used benchmarks show the remarkable improvements achieved by the proposed model in coping with low-quality images and the generalization between datasets. The proposed model achieves an ACC value of 91.65% and an AUC value of 96.34% on the low-quality dataset (LQ) of FF. Furthermore, in the generalization experiments from the FF dataset to Celeb-DF, DFD, and DF-1.0 datasets, the model achieves AUC values of 78.47%, 91.55%, and 76.87% respectively.

However, it was observed that the model exhibits poor performance in detecting single local forgeries, such as expression swaps. This is because the PSAM tends to examine multiple suspicious artefacts across the face through a boosting-suppression mechanism, which may force the model to mine additional noisy information that does not belong to the region of expression-swapping artefacts. Future research will focus on investigating the filtering of false artefact features to improve the accuracy of the model in detecting expression swap.

Acknowledgements This work is partly supported by the National Nature Science Foundation of China (No. 62072286, 61876100, 61572296), Shandong Provincial Science and Technology Support Program of Youth Innovation Team in Colleges (No. 2019KJN041, 2020KJN005) and Shandong Provincial Graduate Education Innovation Plan Project (SDYAL21211).

Availability of data and material The data that support the findings of this study are not openly available due to the sensitivity of face data and are available from the corresponding author upon reasonable request. FF++ and DFD datasets are available through the FaceForensicshttps://github.com/ondyari/FaceForensics, DF1.0 dataset is available through the https://github.com/EndlessSora/DeeperForensics-1.0/tree/master/dataset and DFDC dataset is available through thehttps://www.kaggle.com/competitions/deepfake-detection-challenge/data.

Code availability Our code is available at https://github.com/PPnostalgia/NPMD

Declarations

Conflicts of interest The authors declare that they have no competing interests.

References

1. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion 64:131–148
2. Nguyen TT, Nguyen QVH, Nguyen DT, Nguyen DT, Huynh-The T, Nahavandi S, Nguyen TT, Pham QV, Nguyen CM (2022) Deep learning for deepfakes creation and detection: A survey. Comput Vis Image Underst 223:103525
3. Zhang T (2022) Deepfake generation and detection. Multimedia Tools and Applications 81:6259–6276
4. Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head poses. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8261–8265. IEEE
5. Nguyen HH, Yamagishi J, Echizen I (2019) Capsule-forensics: Using capsule networks to detect forged images and videos. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2307–2311. IEEE
6. Amerini I, Galteri L, Caldelli R, Del Bimbo A (2019) Deepfake video detection through optical flow based cnn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0
7. Dang H, Liu F, Stehouwer J, Liu X, Jain AK (2020) On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5781–5790
8. Afchar D, Nozick V, Yamagishi J, Echizen I (2018) Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7. IEEE
9. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2018) Face forensics: A large-scale video dataset for forgery detection in human faces. arXiv:1803.09179
10. Cozzolino D, Poggi G, Verdoliva L (2017) Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, pp. 159–164
11. Zanardelli M, Guerrini F, Leonardi R, Adami N (2022) Image forgery detection: a survey of recent deep-learning approaches. Multimedia Tools and Applications 82:17521–17566
12. Wang H, Wu X, Huang Z, Xing EP (2020) High-frequency component helps explain the generalization of convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8684–8694
13. Qi H, Guo Q, Juefei-Xu F, Xie X, Ma L, Feng W, Liu Y, Zhao J (2020) Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 4318–4327

14. Nguyen HH, Fang F, Yamagishi J, Echizen I (2019) Multi-task learning for detecting and segmenting manipulated facial images and videos. In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–8. IEEE
15. Zhou P, Han X, Morariu VI, Davis LS (2018) Learning rich features for image manipulation detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1053–1061
16. Masi I, Killekar A, Mascarenhas RM, Gurudatt SP, AbdAlmageed W (2020) Two-branch recurrent network for isolating deepfakes in videos. In: European Conference on Computer Vision, pp. 667–684. Springer
17. Qian Y, Yin G, Sheng L, Chen Z, Shao J (2020) Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: European Conference on Computer Vision, pp. 86–103. Springer
18. Fridrich J, Kodovsky J (2012) Rich models for steganalysis of digital images. IEEE Trans Inf Forensics Secur 7(3):868–882
19. Mayer O, Stamm MC (2019) Forensic similarity for digital images. IEEE Trans Inf Forensics Secur 15:1331–1346
20. Zhao T, Xu X, Xu M, Ding H, Xiong Y, Xia W (2021) Learning self-consistency for deepfake detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15023–15033
21. Chen S, Yao T, Chen Y, Ding S, Li J, Ji R (2021) Local relation learning for face forgery detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1081–1088
22. Kuang L, Wang Y, Hang T, Chen B, Zhao G (2022) A dual-branch neural network for deepfake video detection by detecting spatial and temporal inconsistencies. Multimedia Tools and Applications 81:42591–42606
23. Durall R, Keuper M, Pfreundt FJ, Keuper J (2019) Unmasking deepfakes with simple features. arXiv:1911.00686
24. Liu H, Li X, Zhou W, Chen Y, He Y, Xue H, Zhang W, Yu N (2021) Spatial phase shallow learning: rethinking face forgery detection in frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 772–781
25. Li Y, Chang MC, Lyu S (2018) In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. In: IEEE International Workshop on Information Forensics and Security
26. Hong C, Yu J, Zhang J, Jin X, Lee K (2018) Multimodal face-pose estimation with multitask manifold deep learning. IEEE Trans Industr Inf 15(7):3952–3961
27. Yu J, Tao D, Wang M, Rui Y (2015) Learning to rank using user clicks and visual features for image retrieval. IEEE Transactions on Cybernetics 45(4):767–779
28. Yu J, Tan M, Zhang H, Tao D, Rui Y (2019) Hierarchical deep click feature prediction for fine-grained image recognition. IEEE Trans Pattern Anal Mach Intell 44(2):563–578
29. Hong C, Yu J, Wan J, Tao D, Wang M (2015) Multimodal deep autoen coder for human pose recovery. IEEE Trans Industr Electron 24(12):5659–5670
30. Hong C, Yu J, Tao D, Wang M (2014) Image-based 3d human pose recovery by multi-view locality sensitive sparse retrieval. IEEE Trans Industr Electron 62(6):3742–3751
31. Kohli A, Gupta A (2021) Detecting deepfake, faceswap and face2face facial forgeries using frequency cnn. Multimedia Tools and Applications 80:18461–18478. Springer
32. Woo S, Park J, Lee JY, Kweon IS (2018) Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19
33. Bayar B, Stamm MC (2018) Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. IEEE Trans Inf Forensics Secur 13(11):2691–2706
34. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803
35. Newman MEJ (2013) DeepFakes. http://github.com/deepfakes/faceswap
36. Thies J, Zollhofer M, Stamminger M, Theobalt C, Nießner M (2016) Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2387–2395
37. MarekKowalski: FaceSwap (2013). https://github.com/MarekKowalski/FaceSwap
38. Thies J, Zollhöfer M, Nießner M (2019) Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics (TOG) 38(4):1–12
39. Li Y, Yang X, Sun P, Qi H, Lyu S (2020) Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3207–3216
40. Jiang L, Li R, Wu W, Qian C, Loy CC (2020) Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2889–2898

41. Tan M, Le Q (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR
42. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503
43. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626
44. Li L, Bao J, Zhang T, Yang H, Chen D, Wen F, Guo B (2020) Face x-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5001–5010
45. Nadimpalli A, Rattani A (2023) Facial forgery-based deepfake detection using fine grained features. arXiv:2310.07028v1
46. Zhao L, Zhang M, Ding H, Cui X (2023) Fine-grained deepfake detection based on cross-modality attention. Neural computing & applications 35(15):10861–10874
47. Waseem S, Abu-Bakar SARS, Omar Z, Ahmed BA, Baloch S, Hafeezallah A (2023) Multi-attention-based approach for deepfake face and expression swap detection and localization. EURASIP Journal on Image and Video Processing 2023(1)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.