**Discover** Computing

Research

# Robust manipulated media localization and detection based on high frequency and texture features

**Qian Jiang[1] · Shuai Liu[1] · Shengfa Miao[1] · Huasong Yi[2] · Xin Jin[2] · Yuru Kou[2] · Hanxian Duan[2]**

Received: 14 May 2024 / Accepted: 3 January 2025
Published online: 10 February 2025


**Abstract**
Advances in facial manipulation techniques have resulted in the increasing trend of realistic and indistinguishable identity swap media, which mislead the viewers and accompanied by severe security concerns. While current deepfake detectors demonstrate strong performance under high-quality conditions, they still face notable limitations. This article proposes a novel framework mining high frequency and degraded texture features for locating manipulated traces and improving the generalization ability. To improve the universality of the proposed detector, we design the Multi-feature Mining Stream for capturing the global and subtle discrepancies of undegraded images. Moreover, the Encoder-Decoder Structure is introduced for gaining high localization accuracy and full resolution manipulated regions. This work attempts to solve the tampered region localization issue and achieve face forgery image detection at the meantime. This contributes to help the model perform a more effective differentiation between real and fake content when confronted with high- or low-quality compressed images. Comprehensive experiments on the popular FaceForensics++, Celeb-DF, and DFDC datasets demonstrate the superior performance and robustness of our proposed framework, in particular, achieving performance improvements ranging from 1% to 10% in comparison with the most recent related work.

**Keywords** Face forgery detection · Manipulated media localization · Generalization improving · Media properties

## 1 Introduction

Multiple manipulated media including images and videos with facial information are currently rampant on social media channels and the Internet [1], considering people obtain various kinds of information through networks everyday, these tampered media could mislead viewers and even cause adverse effects in the real society [2]. As shown in the Fig. 1, the examples facial part of several celebrities are swapped with another face to generate the manipulation media, which are too realistic to distinguish from the truth even by human eyes. Meanwhile, thanks to the open access to extensive public databases and rapid advancements in deep learning technique, in particular Variational Auto-encoders [3] and Generative Adversarial Networks(GAN) [4], anyone could generate high-quality and realistic manipulated media. Utilizing accessible apps [5, 6] and open-source tools [7, 8], even non-specialists can easily manipulate face forgery media. Although these manipulation techniques were originally created for amusement, they pose significant risks, as they could be used by malicious actors for unscrupulous purposes, such as reputation

✉ Shengfa Miao, miaosf@ynu.edu.cn; Qian Jiang, jiangqian_1221@163.com; Shuai Liu, sdlyls@mail.ynu.edu.cn; Huasong Yi, yihuasong@stu.ynu.edu.cn; Xin Jin, xinxin_jin@163.com; Yuru Kou, kouyvru@163.com; Hanxian Duan, hanxian_duan@163.com | [1]Engineering Research Center of Cyberspace, Yunnan University, Kunming 650091, China. [2]School of Software, Yunnan University, Kunming 650091, Yunnan, China.

**Fig. 1** Some vivid examples of identity swap images in the social network. The real images are in the top row, and the fake images, generated by deep learning, are in the bottom row

damage, spreading fake news, or even political manipulation. Misinformation often spreads rapidly through social networks [9], leading to significant concerns about public safety and a crisis of trust. To mitigate these risks and counter various face forgery techniques, it is paramount to develop more effective and robust detection and localization methods for practical applications.

Due to the privacy issues and potential security threats posed by facial manipulation media, both academia and industry have devoted attention to detecting forged face images. With the rapid development of deep neural network(DNN) in the image classification field [10, 11], many facial manipulation detection methods [12–21] have been proposed based on DNN. Many of these techniques [12, 13] treat deepfake detection as a basic classification problem, focusing only on extracting global features for the deep neural network classification layer, while a few of them explore in other ways. In our investigation, some advanced manipulation techniques only generate small-scale and subtle discrepancies [22, 23], which are difficult for the global feature layer to capture. The recent works observe this phenomenon and explore the correlation between local region and global feature for extracting the discrepancies and artifacts. Reference [14] designed a bilinear attention map to indicate local manipulated regions and shallow layer artifacts, [15] introduced various subtle discrepancies by multiple spatial attention heads, [21] extracted different facial attributes and characteristics of the forgery images. While the discussed methods focus on identifying local artifacts within the spatial domain, their effectiveness diminishes when faced with color-space compression, thus affecting the robustness and accuracy of detection across varied databases and compression rates. This limitation prompts the exploration of forgery patterns through the lens of frequency features, introduced due to the checkerboard artifacts produced by up-sampling operations in generators [24, 25]. This approach offers a fresh perspective for enhancing detection precision. Recent research [17, 19] has explored combining spatial and frequency domain features for more holistic feature extraction, with [16] specifically leveraging local frequency statistics and frequency-aware decomposition in the context of face forgery detection. Furthermore, techniques [26, 27] employ Fast Fourier Transform (FFT) or Discrete Fourier Transform (DFT) to derive the frequency spectrum, subsequently utilizing these frequency-based features for classification, thereby enhancing detection capabilities. In addition, tampering can be recognized by embedding watermark information into the image and through watermark extraction and detection. For example, [28] and [29] proposed a blind steganography analysis method based on SPAM features and integrated classifiers, and Sahu et al. [30] and [31] proposed a tamper detection and localization method based on dual images and reversible fragile watermarking.

While several approaches focus on detecting forgeries, only a select few delve into pinpointing the specific regions within fake images where facial manipulations are most prevalent. The capability to accurately localize manipulated areas significantly enriches the domain of facial tampering forensics, offering enhanced clarity and insights. Achieving detailed localization maps that precisely highlight manipulated zones at high resolution is crucial for augmenting the interpretability of deep learning techniques in authentic forensic contexts. Although [32] leverages an attention mechanism to identifying manipulated regions, its output, a binary mask, lacks comprehensive attribute generality. Li et al. [33] introduced the Face X-ray technique for identifying fake regions, and the method [34] addressed localization through a gray-scale fakeness map, achieving commendable precision. However, these methods tend to falter with heavily compressed forged images. An ideal localization approach should maintain its robustness across varying levels of image compression quality and across different databases, ensuring reliable detection in diverse forensic situations.

Given the landscape of current methods for localizing and detecting manipulated media, our work is primarily driven by the following motivations: (a) It is imperative not only to harness global spatial features but also to meticulously extract local discrepancies. The rationale stems from the observation that forgery signals often undergo compression and blurring effects from JPEG or H.264 compression algorithms [25]. Incorporating frequency components into our analysis emerges as a potent strategy to enhance the robustness and generalization capabilities of our proposed model, making it more adept at identifying subtle forgery cues amidst compression artifacts. (b) Most DNN-based models focus on detection tasks, but manipulated region localization is more valuable in the manipulation forensic field. The localization method should be implemented with high localization accuracy and full resolution. (c) Simultaneously solving the tampered localization issue and achieving manipulation media detection is useful in the practical forensic scenario. The framework has tremendous potential in the industry and the social network, which will greatly reduce online fraud and malicious rumors, improve public judgment, and enhance the credibility of facial media disseminated on the Internet.

Inspired by the above thoughts, we propose a novel Robust Manipulated Media Localization and Detection Based on High Frequency and Texture Features (RMLD-HFTF), which is aimed at face forgery detection and manipulated region localization. Corresponding to the aforementioned motivations, our method is composed of three sophisticated modules. Firstly, Multi-features Mining Stream (MMS) is designed for extracting multiple local and subtle discrepancies and improving the generalization capabilities. Specifically, the high-frequency band discrepancy between the original and manipulation image is pretty apparent as illustrated in Fig. 2. We adopt high-pass filter function from SRM [35] to capture high frequency feature in MMS, the differences of facial attributes are obvious after frequency filters' processing. To exploit inconsistency across different compression rate images and improve robustness, we utilize data augmentation (e.g., JPEG compression, Gaussian Blurring, etc.) to obtain quality degradation images as another input. Local features from shallow layers are extracted using Steganalysis Feature Extracting Filter (SFEF) based on SRM and learnable filters, which serve as attention weights multiplied with features from the bottom stream. Secondly, the Artifact Guided Attention Module (AGAM) is proposed to fuse dual encoder features and prediction artifact region. AGAM extract robust co-occurrence features and the artifacts could guide the classifier for improving the detection accuracy. Thirdly, we design the Encoder-Decoder Structure based on Atrous Spatial Pyramid Pooling (ASPP) [36] and DeepLabv3 [37]. The RMLD-HFTF framework takes both raw images and quality-degraded images as inputs and outputs the classification result and predicted manipulation region.

FaceForensic++ (FF++) [13], Celeb-DF [23], and DFDC [38] datasets are widely used in the manipulation media forensic research. To demonstrate the efficiency and robustness of the proposed RMLD-HFTF, we conduct comprehensive experiments compared with other state-of-the-art manipulation detection and localization methods on the above datasets. We compare with several localization methods, all of the methods do not satisfy high-resolution and universality simultaneously. Moreover, the proposed framework keeping superior generalization on the cross-dataset evaluation of challenging Celeb-DF [23] and cross compression rate experiment of FF++ [13].

To provide a more comprehensive understanding of our approach, the following sections detail the related work, methodology, and experimental validation of the proposed framework. In Section 2, we review the generation processes of facial manipulation media, examine existing detection methods for identifying forged facial images, and discuss

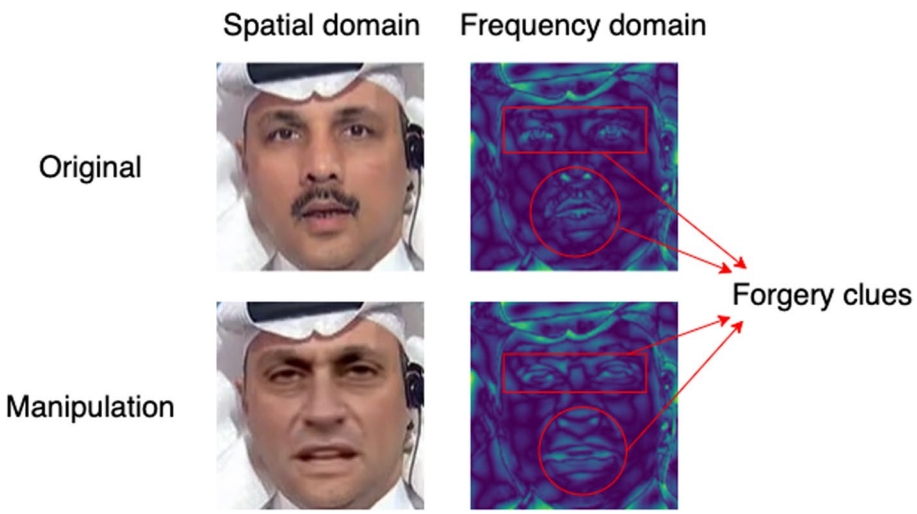


**Fig. 2** Inconsistency in frequency domain could serve as an important forgery clues. The left column is the original and manipulation images in spatial domain, and the frequency information based on high-pass frequency filter is in the right column. The visualization of forgery clues are obviously marked with red label

approaches for localizing manipulated regions. Section 3 presents the proposed Robust Manipulated Media Localization and Detection framework based on High Frequency and Texture Features (RMLD-HFTF), outlining its components and processes. In Section 4, we conduct extensive experiments to evaluate the performance of our method compared to state-of-the-art techniques, particularly in challenging cross-dataset and cross-compression rate scenarios. Finally, Section 5 covers ablation studies where we assess the contributions of key components such as the Multi-features Mining Stream (MMS) and Artifact Guided Attention Module (AGAM) to the overall effectiveness of our framework. The main contributions are summarized as follows.

- Our research introduces an innovative framework, termed RMLD-HFTF, which marks a pioneering effort to concurrently addressing the challenge of localizing tampered regions and detecting manipulated media. This dual-purpose approach significantly advances the field of face forgery detection and localization, setting a new benchmark for future studies.
- We integrate two novel components, MMS and AGAM within our framework. These elements are specifically designed to detect delicate forgery indicators across both frequency and spatial domains, thereby substantially enhancing the framework's resilience against manipulation.
- Through extensive experimentation and rigorous evaluation, our methodology has proven to outperform existing techniques in the realm of face forgery detection and localization. This is particularly evident in scenarios involving varied compression rates and cross-dataset environments. Our ablation studies, complemented by detailed visualizations, offer deep insights into the mechanisms driving our framework's state-of-the-art efficacy.

## 2 Related work

In this section, we provide a concise overview of the generation processes for facial manipulation media, examine detection methodologies for identifying forged facial images, and discuss approaches for localizing manipulated regions.

### 2.1 Generation of facial manipulation media

Manipulated media have spread in the social network due to their high-quality and easy-to-use ability of applications [5–8] that are developed based on deep learning techniques. The types of facial manipulation media can be classified into four categories: entire face synthesis, attribute manipulation, expression swap and identity swap.

Entire face synthesis generates non-existent facial forgery images in the real world. DCGAN [43], WGAN [44], PGGAN [45], and StyleGAN [46] are several classical examples of entire face synthesis. In [43], CNN and GAN were combined for the first time, DCGAN focused on unsupervised learning with pre-trained discriminator. WGAN [44] addressed the mode-dropping phenomenon between generator and discriminator, minimizing a reasonable approximation of the Earth-Mover (EM) distance. Karras et al. [45] proposed an effective method for generating high-resolution images of size $1024 \times 1024$ pixels. StyleGAN [46] has proposed a novel design to automatically learn the unsupervised separation of high-level attributes.

Attribute manipulation is aimed at editing face attributes, such as skin color, hair color, and adding glasses. StarGAN [47] and STGAN [48] are two classic methods of this manipulation technique. Choi et al. [47] proposed a single model to achieve image-to-image translations and synthesize images with diverse attributes. STGAN [48] took the difference between target and source attribute vectors as the input, and modified the encoder matrix to the encoder-decoder.

Expression swap replaces the facial expression of the target image with the facial expression of source image. Face2Face [49] and A2V [50] are two classical manipulation techniques. Face2Face [49] exploited the best-matching facial parts to generate a realistic forgery face based on the target and source images. Suwajanakorn et al. [50] used a recurrent neural network to synthesize high- quality mouth textures and map raw audio features.

Identity swap replaces the face in the target image with the face in the source image, which is similar to expression swap. The classical works include FaceSwap [51] and CycleGAN [39]. FaceSwap [51] used face alignment, Gauss-Newton optimization, and image blending to generate identity-changed media. In [39], Zhu et al. built a mapping and similarly constructed an inverse mapping. Although CycleGAN does not mention identity swap, it can be used for identity swap easily. Through the above research, we consider expression swap and identity swap to be the most perilous manipulation techniques, which could generate realistic and compelling forgery media. This work mainly pays attention to the deepfakes detection and manipulated regions localization of expression swap and identity swap.

### 2.2 Face forgery images detection

In the realm of deepfake detection, significant advancements have been realized concerning efficiency, adaptability, and resilience. Initial detection strategies identified unusual artifacts, such as irregularities in eye blinking [39], anomalies in head pose [52], and inconsistencies in facial warping [53]. However, advancements in forgery algorithms [49, 51] have minimized these detectable artifacts, diminishing the effectiveness of early detection approaches. Rossler et al. [13] introduced the FF++ dataset as a standard for evaluating deepfake detection techniques, highlighting the competence of the Xception [10] model. MesoNet [12] developed two CNN architectures, Meso4 and Meso-Inception4, which foucs on the mesoscopic attributes of doctored images to detect forgery. These methods, however, demonstrate vulnerability when faced with highly compressed datasets as the compression process masks the forgery artifacts.

Notably, disparities in manipulated content become more apparent within the frequency domain, as shown in Fig. 2. The up-convolution stages involved in the manipulation process induce notable irregularities [24, 25] in this domain. Acknowledging this, recent investigations have leveraged high-frequency details to enhance detection precision and robustness. Frank et al. [25] employed frequency representation for the automated identification of deepfake content. F3-Net [16] improved upon this by incorporating local frequency statistics and a frequency-aware decomposition approach, achieving notable detection performance yet facing limitations in generalization due to a focus on subtle discrepancies. Durall et al. [26] proposed an analysis of the amplitude spectrum via DFT to distinguish fake images by observing distribution shifts.

Recent studies [17, 19, 40] have progressively adopted a dual-domain approach, extracting features from both the spatial and frequency domains for a more holistic and complementary feature analysis in deepfake detection. The Two-branch [17] methodology enriches multiple frequency bands using Laplacian of Gaussian (LOG) filters, integrating dual-domain information for enhanced detection. Wang et al. [19] introduced a multi-scale transformer that captures subtle inconsistencies across spatial and frequency dimensions. Meanwhile, SSTNet [40] explored a combination of spatial, steganalysis, and temporal information, utilizing an augmented Xception [10] model alongside LSTM [54] to detect deepfake videos effectively, illustrating the evolving landscape of deepfake detection methodologies.

### 2.3 Manipulated regions localization

Despite numerous initiatives aimed at detecting deepfakes, only a handful of studies focus on pinpointing the exact regions of manipulation within images. Some prior works [41, 42] have ventured into the realm of localizing altered segments for image forgery detection, proposing methods such as a two-stream Faster R-CNN network [41] designed to identify tampered areas through noise discrepancies. Bappy et al. [42] integrated frequency and spatial data to isolate altered zones. However, these methodologies weren't specifically tailored for facial image manipulation localization. In addressing the challenge of localizing deepfakes, emphasis is placed on pixel-level segmentation to detect falsified traces accurately. Dang et al. [32] put forward a model that leverages attention mechanisms for forensic localization, although it did not achieve the desired granularity. Face X-ray [33] attempts to define manipulated regions using a binary mask approach, falling short of capturing the detail at full resolution. Another approach [15] utilizes a patch-based classifier focused on local distortions, consequently overlooking broader manipulation patterns and resulting in a loss of comprehensive information about the forged regions. We summarize the papers mentioned in related works and their merits and their optimizable points in a Table 1.

## 3 The proposed method

In this section, we introduce the Robust Manipulated Media Localization and Detection Based on High Frequency and Texture Features (RMLD-HFTF) for face forgery detection and localization in Fig. 3. We present the Encoder-Decoder features capturing structure in Sub-Section 3.1, introduce the multi-features mining stream (MMS) in Sub-Section 3.2, and report the artifact guided attention module(AGAM) in Sub-Section 3.3.

**Table 1** Summary of papers and their advantages and disadvantages

| Paper | Advantages | Disadvantages |
|---|---|---|
| [39] | Quickly identifies apparent forgeries like irregular eye blinking and head pose anomalies. | Detectable artifacts diminish as forgery techniques improve. |
| [25] | Employs frequency domain representation to automatically detect deepfakes; emphasizes high-frequency details for accuracy. | Limited generalization ability; struggles with diverse forgery techniques. |
| [26] | Analyzes amplitude spectrum using DFT to detect forgery via distribution shifts. | Effective for specific forgery types; less adaptable to complex scenarios. |
| [17] | Enhances multiple frequency bands using Laplacian of Gaussian (LOG) filters; integrates dual-domain features for better detection. | Computational overhead due to model complexity. |
| [19] | Multi-scale transformer captures subtle spatial and frequency inconsistencies. | Limited generalization; relies on dataset distribution. |
| [40] | Combines spatial, steganalytic, and temporal information; uses enhanced Xception model and LSTM for video forgery detection. | May face efficiency bottlenecks when handling complex temporal sequences. |
| [41] | Identifies tampered regions via noise discrepancies; provides accurate localization. | Not specifically designed for facial manipulations; lacks fine-grained localization. |
| [42] | Integrates frequency and spatial data for tampered region localization. | Limited adaptability to sophisticated forgery techniques. |
| [32] | Leverages attention mechanisms for forensic localization of manipulated regions. | Lacks granularity for pixel-level accuracy. |
| [33] | Defines manipulated regions using a binary mask approach. | Fails to capture high-resolution details. |
| [15] | Focuses on local distortions to identify forged regions. | Ignores broader manipulation patterns; lacks comprehensiveness. |

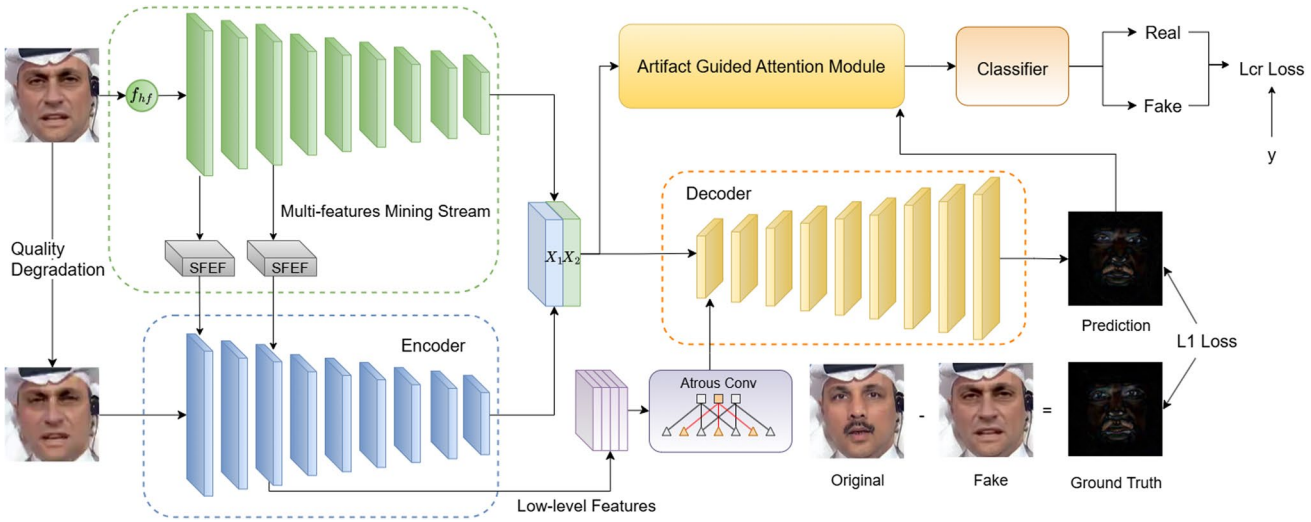


**Fig. 3** Our methodology is underpinned by two critical components within its architectural framework: **a** The Multi-Feature Mining Branch, which is tasked with the extraction of diverse shallow style features through the application of the Shallow Feature Extraction Filter (SFEF); **b** The Artifact-Guided Attention Module, designed to refine and process the amalgamation of features and manipulation traces, employing the cross-attention mechanism for enhanced accuracy and precision

### 3.1 Encoder-decoder features capturing structure

To achieve the goal of local discrepancies capturing and low-level features mining task at the same time, this paper designs an Encoder-Decoder feature-capturing structure. Considering images usually compressed with compression algorithms, such as JPEG and H.264, the highly compression detection accuracy of previous methods has declined. To improve the generalization and robust ability of our proposed model, the input images are degraded with Gaussian blur as the Encoder input:

$$G(x,y,z) = \frac{1}{\sqrt{(2\pi)^3\sigma^6}} \exp\left(-\frac{x^2+y^2+z^2}{2\sigma^2}\right), \tag{1}$$

where $G(x, y, z)$ represents the value after applying Gaussian blur at the point $(x, y, z)$. In image processing, $x$ and $y$ typically represent the horizontal and vertical coordinates of the image, while $z$ can denote the depth in the context of three-dimensional images. $\sigma$ is the standard deviation of the Gaussian function, controlling the extent of the blur. $(2\pi)^3$ and $\sigma^6$ represent the parts of the normalization factor, ensuring that the integral of the entire Gaussian function equals 1. And exp is the exponential function.

As shown in Fig. 3, the low-level features represent the texture forgery clues of the input image, which are captured by shallow convolution layers. Specifically, low-level convolutional layers have smaller receptive fields and focus on local texture and color features. The output feature map $x_1$ from the Encoder corresponds to a component of the fused features $F_{fs}$, which is one part of Decoder input. Then we feed low-level features into Atrous Spatial Pyramid Pooling block (ASPP), providing an additional input to the Decoder. ASPP plays a crucial role in enhancing the model's ability to capture multi-scale texture forgery information and expand the receptive field. And ASPP allows the network to efficiently process contextual information at different scales and facial regions, thereby improving the localization accuracy to segment and localize manipulation regions of varying sizes.

It is noteworthy to highlight that the loss function consists of two parts. The first part is the cross-entropy loss $L_{cr}$ used in the spatial domain of the training phase classifier:

$$L_{cr} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)], \tag{2}$$

where $N$ is the number of samples in the dataset, $y_i$ is the true label of the $i$-th sample, which can be either 0 or 1. And $\hat{y}_i$ is the predicted probability that the $i$-th sample belongs to the real images class without forgery. The summation $\sum_{i=1}^{N}$ iterates over all samples in the training dataset.

The second part of loss function is $L_1$ loss (Least Absolute Deviations, LAD) calculated by forged area prediction image, and the ground truth image, which is generated by corresponding real label image and fake forgery image:

$$L_1 = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{3}$$

where $N$ is the number of samples in the dataset. $y_i$ is the true value of the $i$-th sample, and $\hat{y}_i$ is the predicted result for the $i$-th sample. $|y_i - \hat{y}_i|$ presents the absolute difference between the ground truth pixel value and the predicted region value for each sample. As the above equation, the summation $\sum_{i=1}^{N}$ iterates over all samples.

The designed loss function is used to balance the tasks of forgery image detection and forgery area prediction image generation, so that the model can generate high-resolution and 3-channel forgery area prediction images while improving detection accuracy and generalization:

$$L_t = \lambda_1 L_{cr} + \lambda_2 L_1, \tag{4}$$

where the two hyper-parameters $\lambda_1$ and $\lambda_2$ are set to 0.6 and 0.4, respectively, based on empirical validation. Specifically, we determined this parameter by the trial experiments on a small-scale dataset, testing various combinations of $\lambda_1$ and $\lambda_2$ values ranging from 0.1 to 0.9. Thus, we got the optimal configuration that balances the detection and localization performance.

The generated fake area prediction image is used as an attention feature weight in the fake area guided attention enhancement module (AGAM) part to guide the classifier to learn high-frequency shallow texture features, thereby further improving the detection effect of the classifier. AGAM is introduced in the below section C.

### 3.2 Multi-features mining stream

Multi-Feature Mining Stream (MMS), aims to extract multiple local and subtle differences and improve generalization performance. Specifically, high-frequency band differences between the original and tampered images are clearly visible. Steganalysis Feature Extracting Filter (SFEF) based on SRM [35] is used to capture high-frequency features in MMS. After processing by the frequency filter $f_h f$, the difference in facial attributes is obvious. Discrete Cosine Transform (DCT) noted $D$ is applied for extracting different level frequency band with masks:

$$Y_{fd} = f_{hf}[D(x) \odot (f^i_{base} + f^i_{learn})], \quad i = 1, 2, 3, \tag{5}$$

where $Y_{fd}$ is frequency domain features, and $x$ is the input image. We manually design binary filter $f^i_{base}$, and $f^i_{learn}$ presents learnable filter.

Shallow local features will be computed using steg-analytic feature extraction filters (SFEF) based on SRM and learnable filters, which are treated as attention weights element-wise adding with features of the underlying Encoder stream. The key idea behind the SFEF filter is extracting local pixel dependencies independently of frequency features by calculating the residuals between adjacent pixels. The process of calculating the residual $R_{ij}$ is:

$$R_{ij} = \hat{X}_{ij}(N_{ij}) - X_{ij}, \tag{6}$$

where $N_{ij}$ denotes the neighborhood pixels surrounding pixel $X_{ij}$, and $\hat{X}$ is defined as the predictor of $X_{ij}$. Employing filters with distinct weights enables the capture of intricate dependencies among local pixel values. Furthermore, the SFEF filter can be viewed as a high-pass filter to learn local high-frequency features in the noise domain to mine hidden traces of tampering.

During the data processing stage, each learnable SFEF filter is characterized by a convolution kernel size of 5$x$5$x$3. Three distinct SFEF filters are employed to extract high-frequency noise maps from the Red (R), Green (G), and Blue (B) channels, aiming to generate an output high-frequency map of the same dimensions as the input image. Moreover, experimentally chosen filters serve as initial parameters for the learnable filters. Notably, a non-linear layer is deliberately omitted after the learnable filter to avoid potential damage to tampering traces within the noisy feature maps. This design choice ensures that the learnable SFEF filter can autonomously optimize and configure the filter weights to address the specific requirements of the face forgery detection task in a trainable manner.

### 3.3 Artifact guided attention module

The Artifact Guided Attention Module (AGAM) is designed to extract co-occurrence features and artifacts, which guide the classifier to improve detection accuracy. Attention maps could highlight areas of tampered images and thereby guide the network to detect these areas, AGAM is useful for face forgery detection. In effect, each pixel in the attention map calculates the probability that its receptive field corresponds to the tampered region in the input image. Digital forensics has shown that forgery identification is possible due to the fingerprint of high-frequency information in real images. Therefore, it is feasible to detect forgery clues in high-frequency information due to algorithmic processing. The attention map is inserted before the classifier network, where the receptive fields correspond to local patches of appropriate size, then the features preceding the prediction region encode the high-frequency features of the corresponding patch.

As shown in the Fig. 4, the input consists of two parts. The input of the first part is the fused features, and the second part input is the predicted 3-channel regional image after convolution and activation function, as the attention feature maps guides the classifier. The calculated attention map is multiplied element by element with the feature map, and the obtained features are input into the classifier. Finally, the fake image detection model obtains the result of classification.

We firstly apply convolution layer and fully connection layer to the prediction region image, and set a sigmoid activation function to get the attention map noted as *Matt*. Then the fusion features $F_{fs}$ is element-wise added with mean map $M_i \in R^{1\times H\times W}$, and the middle layer features is noted with $M_{midd} \in R^{n\times H\times W}$. After convolution layer calculating, $M_{midd}$ element-wise multiply with *Matt*, the final features $F_f n$ are partitioned by:

$$F_{fn} = Conv(F_{fs} + M_i) \odot Matt, \quad i = 1, 2, ..., C, \tag{7}$$

where $i$ is the $i$-th channel of fusion features. We put the features $F_{fn}$ into classifier to get the detection results of every batch images, which present the facial image is real or fake.

## 4 Experiments

In this section, we commence by delineating the comprehensive experimental framework. Subsequently, we provide a detailed exposition of the extensive experimental outcomes, which underscore the efficacy and resilience of the proposed framework.

### 4.1 Experimental setup

**Databases.** To thoroughly assess the precision and resilience of our approach in both detection and localization tasks, we utilized datasets including FaceForensic++ (FF++) [13], Celeb-DF [23], and DFDC [38]. FF++ [13] stands out as a benchmark dataset, comprising 1,000 authentic and 4,000 manipulated videos through methods such as DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). It offers videos at three compression levels: raw (C0), lightly compressed (c23/HQ), and heavily compressed (c40/LQ). Adhering to the dataset's standardized procedure [13], we allocated 720 videos for training, 140 for validation, and another 140 for testing. For fake videos, we extracted 128 frames, and for

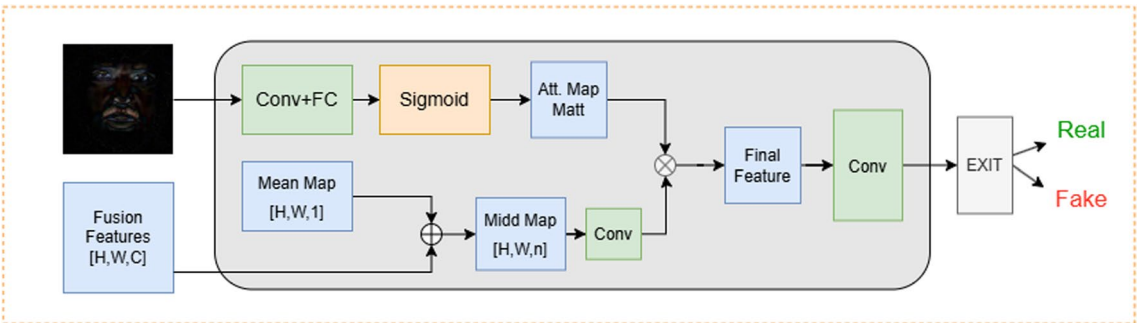


**Fig. 4** Artifact Guided Attention Module.The fusion features are produced by multi-features mining stream and image quality degradation stream, and predicted manipulation region image is produced by Decoder. We apply the prediction map as an attention weight map, which guides classifier having better detection result based on fusion features

genuine videos, we gathered 512 frames using the Dlib algorithm [55], ensuring the real frames outnumber the fake ones by a ratio of four to one.

We trained FF++ with more positive and negative samples because this dataset is our key training set. In real application scenarios, there are often far more real samples than fake samples, so we followed a sampling ratio of 4:1, which helps to better simulate the actual detection scenarios and improve the performance of the model in real situations. FF++, as a standard benchmark dataset, provides different compression levels and multiple forgery methods that can help models learn richer features.

The Celeb-DF [23] dataset, featuring both genuine and DeepFake synthesized footage, comprises 590 original clips sourced from YouTube, showcasing a diverse array of ages, ethnicities, and genders, alongside 5,639 DeepFake counterparts. Following the official guidelines of Celeb-DF [23], we extracted 128 frames from the designated video list.

Lastly, the DFDC [38] dataset, the most extensive public collection of face-swapping video clips, includes over 100,000 videos from 3426 remunerated actors, created using an assortment of DeepFake, GAN, and traditional methods. By accessing the example dataset on Kaggle, we adhered to the provided testing list [38] to extract 128 frames from both real and manipulated videos.

Celeb-DF and DFDC are mainly used for performance evaluation across datasets. We extract a fixed number of samples from these datasets in order to maintain the consistency of the evaluation and ensure the generalization ability and robustness of the models in multiple scenarios.

**Evaluation metrics.** The detection metrics include the frame-level accuracy (ACC) and Area Under the Curve of ROC (AUC) following the previous studies [12, 14, 16] setting. ACC is used to evaluate the detection accuracy of our method in the practical scenario. AUC is a performance indicator that measures the advantages and disadvantages of the classifier, and we apply AUC as another detection metric to give a comprehensive perspective.

**Implementation details.** For every frame retrieved from the FF++, Celeb-DF(v2), and DFDC datasets, we employ Dlib for facial alignment and cropping, resulting in images of dimensions 256 x 256. Within our framework, DeepLabv3 [37], leveraging an Xception [10] backbone pretrained on ImageNet [56], serves as the foundational architecture for the RMLD-HFTF model. This model undergoes optimization via the Adam optimizer, configured with hyperparameters ($\beta_1 = 0.9$, $\beta_2 = 0.99, \epsilon = 10^{-8}$). We initialize the learning rate at 2e-3, establish the random seed at 42, and conduct training over 20 epochs and a batch size of 32 across 40,000 iterations, with all other parameters maintained at their default settings. The computational work is performed on a GeForce RTX 4090 GPU, equipped with 24 GB of video memory, and the model's development is facilitated by the PyTorch library.

### 4.2 Manipulation detection effectiveness evaluation

**Different manipulation methods detection.** We firstly compared RMLD-HFTF with previous detection methods against different manipulation methods on FF++ low quality dataset to demonstrate the effectiveness of compressed forgery image detection. Although our primary objective is to enhance the robustness and generalization of facial forgery image detection, our approach also achieved excellent detection capability in the intra-dataset scenario as shown in Table 2. In situations where official implementations are unavailable, we mark the respective results as '-' in the tables to denote that the testing outcome is not defined. The bolded results highlight the top performance in our evaluation metrics. Addressing the challenge of local manipulations by Face2Face (F2F) and NeuralTextures (NT) techniques, the RMLD-HFTF framework goes beyond merely leveraging global phase differences, as seen in SPSL [18]. It intricately captures essential high-frequency artifacts within local patches, significantly enhancing detection capabilities. Particularly against the complex NT manipulation technique, our approach registers a substantial increase in accuracy (ACC) by 3.7% and Area Under the Curve (AUC) by 1.86%, surpassing the benchmark set by Xception [10]. This advancement owes to the meticulous integration of the Multi-scale Mining Module (MMS) and the Encoder branch, adeptly tracing forgery signatures across spatial and frequency realms while focusing on both low-level features and prominent high-frequency artifacts.

**Different video compression detection.** In assessing the performance of RMLD-HFTF across different compression levels within the FF++ dataset, our methodology distinctly surpasses prior detection techniques, as delineated in Table 3. Notably, our framework exhibits superior efficacy over earlier methods such as Steg.Features [35], LD-CNN[57], CP-CNN [58], and MesoNet [12]. Benchmark models like Xception [10] and EfficientNet-B4 [63], often employed as the foundational networks in contemporary approaches, also fall short when compared with RMLD-HFTF. Specifically, our method demonstrates an enhancement of 1.99% and 2.26% in the AUC metric for the LQ and HQ settings, respectively, over Xception, as well as a 3.09% and 0.46% augmentation over EfficientNet-B4. Despite a marginal underperformance in comparison to MADD [14] under the HQ criterion, RMLD-HFTF excels in the more stringent LQ context. While MADD

**Table 2** The table presents quantitative analysis of detection performance, specifically accuracy (ACC%) and area under the curve (AUC%), across the FaceForensics++ (FF++ LQ) dataset under four distinct facial manipulation methodologies: DeepFake (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT)

| | DF | | F2F | | FS | | NT | |
|---|---|---|---|---|---|---|---|---|
| Methods | ACC | AUC | ACC | AUC | ACC | AUC | ACC | AUC |
| Steg.Features [35] | 73.64 | – | 73.72 | – | 68.93 | – | 63.33 | – |
| LD-CNN [57] | 85.45 | – | 67.88 | – | 73.79 | – | 78.00 | – |
| CP-CNN [58] | 84.55 | – | 73.72 | – | 82.52 | – | 70.67 | – |
| MesoNet [12] | 87.27 | – | 56.20 | – | 61.17 | – | 40.67 | – |
| Xception [10] | 95.13 | 99.24 | 87.34 | 93.62 | 92.42 | 97.08 | 77.54 | 84.51 |
| SPSL [18] | 93.48 | 98.50 | 86.02 | 94.62 | 92.26 | 98.10 | 76.78 | 80.49 |
| SurFake[59] | 88.46 | 94.71 | 75.62 | 86.43 | 84.87 | 90.42 | 52.81 | 55.49 |
| FreMask [60] | 92.09 | 96.34 | 85.16 | 93.71 | 87.11 | 94.28 | 58.41 | 64.23 |
| UnseenArti [61] | 94.49 | 98.09 | 87.12 | 93.15 | 88.14 | 95.47 | 64.81 | 69.81 |
| MADNet [62] | 95.17 | 98.73 | 86.87 | 93.84 | 90.19 | 96.47 | 70.37 | 77.36 |
| Ours | **97.37** | **99.46** | **89.54** | **95.13** | **92.67** | **98.32** | **81.24** | **86.37** |

Bold indicates the best results

**Table 3** Presenting the quantitative outcomes for accuracy (ACC in %) and the area under the curve (AUC in %) across the FaceForensics++ (FF++) dataset, evaluated under two distinct compression quality conditions: high-quality (c23, with minimal compression) and low quality (c40, with substantial compression)

| | LQ(c40) | | HQ(c23) | |
|---|---|---|---|---|
| Methods | ACC | AUC | ACC | AUC |
| Steg.Features [35] | 55.98 | – | 70.97 | – |
| LD-CNN [57] | 58.69 | – | 78.47 | – |
| CP-CNN [58] | 61.18 | – | 79.08 | – |
| MesoNet [12] | 70.47 | – | 83.10 | – |
| SPSL [18] | 81.57 | 82.82 | 91.50 | 95.32 |
| Two Branch [17] | 86.34 | 86.59 | 96.43 | 98.70 |
| Xception [10] | 86.86 | 89.30 | 95.73 | 96.30 |
| EfficientNet-B4 [63] | 86.67 | 88.20 | 96.63 | 98.18 |
| GFFD [64] | 86.89 | 88.27 | 96.87 | 98.45 |
| MADD [14] | 87.28 | 89.12 | **97.34** | **99.29** |
| DIFLD [65] | 84.95 | 89.40 | 93.00 | 97.57 |
| Ours | **87.63** | **91.29** | 96.92 | 98.56 |

Bold indicates the best results

adeptly extracts nuanced features for forgery identification in the RGB spectrum, it encounters challenges in discerning tampered regions within LQ scenarios, which are characterized by diminished artifact visibility. The introduction of our MMS module, adept at isolating high-frequency elements and subtle variances, significantly elevates the detection accuracy for images subjected to severe compression. It's worth noting the enhanced generalization capability of our approach over both MADD and GFFD [64], a topic further explored in the subsection on validating generalization prowess.

### 4.3 Generalization ability validation

**Generalization of cross-manipulation detection.** Generalizability is paramount in the realm of DeepFake detection. To evaluate the RMLD-HFTF model's adeptness at generalizing across various manipulation methodologies, we embarked on cross-manipulation testing using the high-quality version of the FaceForensics++ (FF++(HQ)) dataset, which includes four distinct manipulation techniques. The framework is uniquely trained on forged images derived from a singular method and subsequently tested against all four techniques. In our comparative analysis, which includes RMLD-HFTF, the foundational model EfficientNet-B4, and two cutting-edge methods, MADD [14] and GFFD [64], it was observed that EfficientNet-B4's binary classification approach tends to overfit to particular manipulation styles, thereby compromising its generalizability in diverse manipulation contexts. While MADD improves detection accuracy within specific manipulation types through a multi-attention mechanism that zeroes in on forged traces across facial regions, its generalization efficacy remains somewhat limited in comparison to other methods. GFFD, on the other hand, exhibits enhanced

**Table 4** Cross manipulation techniques quantitative detection results on FF++ (HQ) dataset with AUC(%) evaluation metric

| Training Set | Method | Testing Set (AUC) | | | |
|---|---|---|---|---|---|
| | | DF | F2F | FS | NT |
| DF | EfficientNet-B4 [63] | 99.53 | 69.91 | 49.54 | 75.68 |
| | MADD [14] | **99.67** | 68.12 | 49.25 | 71.68 |
| | GFFD [64] | 99.42 | 72.21 | 50.12 | 78.12 |
| | Ours | 99.56 | **73.14** | **52.63** | **79.32** |
| F2F | EfficientNet-B4 [63] | 85.10 | 99.26 | 59.71 | 69.49 |
| | MADD [14] | 85.87 | **99.33** | 58.12 | 70.12 |
| | GFFD [64] | 86.76 | 99.02 | 61.92 | 72.02 |
| | Ours | **86.81** | 99.29 | **63.19** | **73.20** |
| FS | EfficientNet-B4 [63] | 67.86 | 72.10 | 99.69 | 53.75 |
| | MADD [14] | 68.12 | 71.18 | **99.78** | 50.18 |
| | GFFD [64] | 69.87 | 74.31 | 99.21 | 54.42 |
| | Ours | **70.83** | **76.45** | 99.65 | **54.73** |
| NT | EfficientNet-B4 [63] | 87.16 | 70.19 | 48.35 | 99.43 |
| | MADD [14] | 89.12 | 72.28 | 50.16 | **99.68** |
| | GFFD [64] | 89.23 | 71.32 | 51.04 | 99.12 |
| | Ours | **89.75** | **72.64** | **51.53** | 99.57 |

The bold results are the best, while the second-best results are marked with underline

generalization by identifying regional inconsistencies, leveraging features from both the spatial and noise domains. Our RMLD-HFTF method stands out by effectively capturing multi-scale textural artifacts and regional noise inconsistencies, significantly bolstering its generalization capabilities in varied manipulation scenarios, as demonstrated in Table 4.

**Generalization evaluation on cross-dataset.** To more rigorously test the generalizability of our approach, we expanded our evaluation beyond the scope of cross-manipulation tests within the FF++. Specifically, the synthetic media produced by the four manipulation techniques originate from identical real footage, potentially skewing generalization assessments. Therefore, to more accurately gauge our model's real-world applicability, cross-dataset evaluations were performed. The RMLD-HFTF was trained on the FF++(HQ) dataset and subsequently assessed against the Celeb-DF(v2) and DFDC datasets, both renowned for their high-quality fake content. As elucidated in Table 5, our RMLD-HFTF exhibited exceptional generalization prowess, securing an 8.33% and 4.69% AUC metric enhancement over the benchmark Xception model on Celeb-DF(v2) and DFDC, respectively. Moreover, our method outstripped the performance of recent advancements in face forgery detection, establishing new state-of-the-art benchmarks. While MADD demonstrates

**Table 5** Cross-dataset detection results from FF++ high-quality dataset to Celeb-DF(v2) and DFDC with AUC(%) evaluation metric

| Methods | Training set | FF++ (HQ) | Celeb-DF (v2) | DFDC |
|---|---|---|---|---|
| Two-Stream [66] | Private dataset | 70.10 | 53.80 | 61.40 |
| Meso4 [12] | FF++ (HQ) | 84.70 | 54.80 | – |
| MesoInception4 [12] | FF++ (HQ) | 83.00 | 53.60 | – |
| FWA [67] | Private dataset | 80.10 | 56.90 | 72.70 |
| Xception-c23 [10] | FF++ (HQ) | 98.23 | 65.30 | 69.83 |
| Xception-c40 [10] | FF++ (HQ) | 95.50 | 65.50 | – |
| Two-Branch [17] | FF++ (HQ) | 93.18 | 73.41 | – |
| F3-Net [16] | FF++ (HQ) | 98.10 | 65.17 | – |
| EfficientNet-B4 [63] | FF++ (HQ) | 99.18 | 64.29 | 70.53 |
| GFFD [64] | FF++ (HQ) | 99.42 | 70.85 | 73.01 |
| M2TR [19] | FF++ (HQ) | 99.50 | 65.70 | – |
| MADD [14] | FF++ (HQ) | **99.67** | 67.44 | 72.18 |
| 3D-CS [68] | FF++ (HQ) | 98.70 | 72.04 | – |
| TAN-GFD [69] | FF++ (HQ) | 99.21 | 72.33 | 73.46 |
| Ours | FF++ (HQ) | 99.56 | **73.63** | **74.52** |

Bold indicates the best results

a refined detection capability, possibly edging out our method in certain respects, the RMLD-HFTF's generalization capability distinctively exceeds that of MADD, underscoring its robustness and efficacy in broader, real-world contexts.

### 4.4 Predicted manipulation region experiments

As shown in Fig. 5, fake images are generated by four manipulated techniques including DeepFake, Face2Face, FaceSwap, and Neural Texture. Among the result images, Neural Texture forgery method tends to replace the mouth of the real face, and the predicted image also reflects the high-frequency shallow texture features corresponding to the forgery technology. Moreover, DeepFake and FaceSwap forgery technologies replace the facial features of the entire face area. It can be seen that the forged area prediction map in the fourth row is concentrated in the eyes, nose, and mouth, which is in line with our expectations. Face2Face visually manipulates the nose and mouth area as shown in the Fig. 5 third and fourth column, and subtly changes eyes and eye-brown, the color of mouth is changed to pink in the third column which is marked in the prediction row.

## 5 Ablation studies

In this section, we conduct a series of ablation studies to assess the contribution of our newly introduced components, the Multi-features Mining Stream (MMS) and the Artifact Guided Attention Module (AGAM), across both intra-dataset (FF++) and cross-dataset (Celeb-DF and DFDC) contexts. MMS is specifically crafted to capture high-frequency features from the shallow layers, a critical step given that distinct forgery techniques introduce unique manipulation characteristics. AGAM, on the other hand, is deployed to pinpoint the exact locations of forgery within an image by leveraging an attention mechanism, facilitating the interactive fusion of predicted regions with multiple feature sets for enhanced localization precision. To isolate the impact of these components, we systematically remove each one from our framework, maintaining all other conditions constant, and then evaluate the performance of the resulting models: 1) the baseline Xception model without MMS and AGAM, 2) our model without MMS, and 3) our model without AGAM.

### 5.1 Effectiveness to in-dataset detection

Table 6 presents our experimental findings on both the high-quality (HQ) and the more demanding low-quality (LQ) settings of the FF++ dataset, illustrating the contributions of the MMS and AGAM modules to detection efficacy. Relative to the baseline Xception framework, the incorporation of both MMS and AGAM demonstrates a noticeable improvement in detection capabilities. Notably, AGAM contributes more significantly to the enhancement of detection performance than

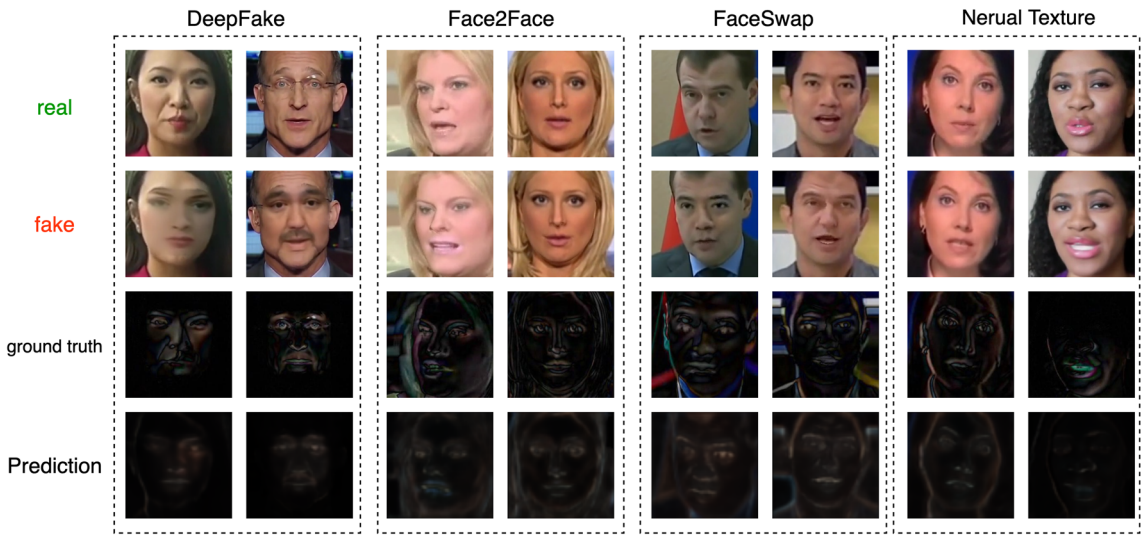


**Fig. 5** Manipulated region localization results. There are the results of four different manipulation techniques, which generate fake images by real facial images. Ground truth images are calculated by fake images and real images

**Table 6** Ablation study outcomes for the FF++ dataset under both low quality (LQ) and high quality (HQ) configurations

| Methods | FF++(LQ) | | FF++(HQ) | |
|---|---|---|---|---|
| | ACC | AUC | ACC | AUC |
| Baseline | 86.86 | 89.30 | 95.73 | 96.30 |
| w/o MMS | 86.93 | 90.39 | 96.42 | 98.24 |
| w/o AGAM | 87.34 | 91.17 | 96.13 | 97.36 |
| Ours(RMLD-HFTF) | **87.63** | **91.29** | **96.92** | **98.56** |

Bold indicates the best results

MMS, highlighting its effectiveness in guiding the receptive field towards manipulated regions and leveraging artifact information from dual streams. The superior performance of RMLD-HFTF, as compared to models without MMS or AGAM, underscores the complementary nature of features sourced from the high-frequency domain and those derived from the degraded RGB domain through MMS. Furthermore, AGAM's strategic utilization of these integrated features further bolsters our framework's detection accuracy.

### 5.2 Effectiveness to cross-dataset detection

In our exploration of RMLD-HFTF's enhanced generalization capabilities, cross-dataset evaluations, as depicted in Table 7, reveal significant improvements in both accuracy and generalization over the baseline Xception model. The AGAM notably bolsters the model's detection accuracy, contributing to an increase of 1.12%. Additionally, the MMS module significantly enhances the model's generalization ability, evidenced by an 8.26% improvement in Celeb-DF (v2) and a 4.54% enhancement in DFDC. This indicates MMS's pivotal role in enriching the model's adaptability across diverse datasets. Collectively, these findings underscore our framework's adeptness at navigating the delicate balance between robust detection performance and broad generalization across varied forgery contexts.

## 6 Conclusion

In this article, we proposed a novel framework for face forgery detection in social networks, which contributes to preserving the integrity and reliability of facial information, enhancing the security of personal data in the era of information technology. The framework addresses the limitations of current deepfake detectors, such as their low performance under different image compression qualities and cross-database scenarios, as well as their lack of focus on manipulated trace localization. The proposed framework includes a Multi-feature Mining Stream for capturing global and subtle discrepancies in undegraded images, as well as an Encoder-Decoder Structure for achieving high localization accuracy and full-resolution manipulated regions. The framework outperforms existing detectors in terms of performance and robustness in cross-database scenarios and manipulation media localization. This is the first attempt to solve the tampered region localization issue and achieve face forgery image detection simultaneously. The results of comprehensive experiments and evaluations support the superior performance of the proposed framework. Overall, the framework contributes to addressing the security concerns associated with realistic and indistinguishable identity swap media on social networks. The framework, however, faces limitations in balancing detection and localization tasks, which may result in modest

**Table 7** Ablation results on cross-dataset setting with AUC(%) evaluation metric

| Methods | Traning set | FF++ (HQ) | Celeb-DF (v2) | DFDC |
|---|---|---|---|---|
| baseline | FF++ (HQ) | 98.23 | 65.30 | 69.83 |
| w/o MMS | FF++ (HQ) | 99.34 | 72.92 | 73.21 |
| w/o AGAM | FF++ (HQ) | 98.57 | 73.56 | 74.37 |
| Ours(RMLD-HFTF) | FF++ (HQ) | **99.56** | **73.63** | **74.52** |

Bold indicates the best results

improvements for each task individually. Future work will aim to better integrate these two objectives, optimizing both detection accuracy and localization precision to enhance overall performance in real-world applications.

**Acknowledgements** This study is supported by the National Natural Science Foundation of China (No. 62261060), Yunnan Fundamental Research Projects (Nos. 202401AT070470, 202301AW070007, 202301AU070210), National Key Research and Development Program of China (No. 2024YFC3014300), Major Scientific and Technological Project of Yunnan Province (No. 202202AD080002 and 202302AD080006 ), Postgraduate Research Innovation Project of Yunnan University (Nos. 2222962), and Yunnan Province Expert Workstations (202305AF150078), and Xingdian Talent Project in Yunnan Province.

**Author contributions** Qian Jiang: Writing-original draft, Methodology, Resources, Funding acquisition, Supervision. Shuai Liu: Conceptualization, Writing-original draft, Methodology. Shengfa Miao: Visualization, Writing-review&editing, Funding acquisition, Supervision. Huasong Yi: Software, Visualization, Writing-review&editing. Xin Jin: Supervision, Resources. Yuru Kou: Software, Visualization. Hanxian Duan: Writing-Review&Editing, Resources.

**Data availability** Data openly available in a public repository. The data that support the findings of this study are openly available in a github repository at github.com/ondyari/FaceForensics.

## Declarations

**Competing interests** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.



## References

1. Thomas D. Deepfakes: a threat to democracy or just a bit of fun. BBC News 2020.
2. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J. Deepfakes and beyond: a survey of face manipulation and fake detection. Inform Fusion. 2020;64:131–48.
3. Pu Y, Gan Z, Henao R, Yuan X, Li C, Stevens A, Carin L. Variational autoencoder for deep learning of images, labels and captions. Adv Neural Inform Process Syst. 2016;29.
4. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. Commun ACM. 2020;63(11):139–44.
5. Faceapp. http://faceapp.com/app.1
6. deepfakesweb. https://deepfakesweb.com/
7. DeepFakes. https://github.com/deepfakes/faceswap
8. DeepFaceLab. https://github.com/iperov/DeepFaceLab
9. Kwok AO, Koh SG. Deepfake: a social construction of technology perspective. Curr Issues Tour. 2021;24(13):1798–802.
10. Chollet F. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017;1251–1258.
11. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR. 2019;6105–6114.
12. Afchar D, Nozick V, Yamagishi J, Echizen I. Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE. 2018;1–7.
13. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. Faceforensics: a large-scale video dataset for forgery detection in human faces. arXiv. 2018. https://doi.org/10.48550/arXiv.1803.09179.
14. Zhao H, Zhou W, Chen D, Wei T, Zhang W, Yu N. Multi-attentional deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021;2185–2194
15. Chai L, Bau D, Lim S-N, Isola P. What makes fake images detectable? understanding properties that generalize. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16. Springer. 2020;103–120.
16. Qian Y, Yin G, Sheng L, Chen Z, Shao J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: European Conference on Computer Vision. Springer, Berlin. 2020;86–103.
17. Masi I, Killekar A, Mascarenhas RM, Gurudatt SP, AbdAlmageed W. Two-branch recurrent network for isolating deepfakes in videos. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII. Springer. 2020;16:667–684.

18. Liu H, Li X, Zhou W, Chen Y, He Y, Xue H, Zhang W, Yu N. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021;772–781.
19. Wang J, Wu Z, Ouyang W, Han X, Chen J, Jiang Y-G, Li S-N. M2tr: Multi-modal multi-scale transformers for deepfake detection. In: Proceedings of the 2022 International Conference on Multimedia Retrieval. 2022;615–623.
20. Wang G, Jiang Q, Jin X, Li W, Cui X. Mc-lcr: multimodal contrastive classification by locally correlated representations for effective face forgery detection. Knowl Based Syst. 2022;250:109114.
21. Xiao S, Lan G, Yang J, Li Y, Wen J. Securing the socio-cyber world: multiorder attribute node association classification for manipulated media. IEEE Trans Comput Soc Syst. 2022. https://doi.org/10.1109/TCSS.2022.3213832.
22. Thies J, Zollhöfer M, Nießner M. Deferred neural rendering: image synthesis using neural textures. Acm Trans Graph (TOG). 2019;38(4):1–12.
23. Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020;3207–3216.
24. Durall R, Keuper M, Keuper J. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020;7890–7899.
25. Frank J, Eisenhofer T, Schönherr L, Fischer A, Kolossa D, Holz T. Leveraging frequency analysis for deep fake image recognition. In: International Conference on Machine Learning. PMLR. 2020;3247–3258.
26. Durall R, Keuper M, Pfreundt F-J, Keuper J. Unmasking deepfakes with simple features. arXiv. 2019. https://doi.org/10.48550/arXiv.1911.00686.
27. Zhang X, Karaman S, Chang S-F. Detecting and simulating artifacts in gan fake images. IEEE Int Workshop Inform Forens Secur (WIFS). 2019. https://doi.org/10.1109/WIFS47025.2019.9035107.
28. Hemalatha J, Sekar M, Kumar C, et al. Towards improving the performance of blind image steganalyzer using third-order spam features and ensemble classifier. J Inform Secur Appl. 2023;76:103541.
29. Sahu AK. A logistic map based blind and fragile watermarking for tamper detection and localization in images. J Ambient Intell Human Comput. 2022;13(8):3869–81.
30. Sahu AK, Sahu M, Patro P, et al. Dual image-based reversible fragile watermarking scheme for tamper detection and localization. Pattern Anal Appl. 2023;26(2):571–90.
31. Sahu AK, Umachandran K, Biradar VD, et al. A study on content tampering in multimedia watermarking. SN Comput Sci. 2023;4(3):222.
32. Dang H, Liu F, Stehouwer J, Liu X, Jain AK. On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020;5781–5790.
33. Li L, Bao J, Zhang T, Yang H, Chen D, Wen F, Guo B. Face x-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020;5001–5010.
34. Huang Y, Juefei-Xu F, Guo Q, Liu Y, Pu G. Fakelocator: robust localization of GAN-based face manipulations. IEEE Trans Inform Forens Secur. 2022;17:2657–72.
35. Fridrich J, Kodovsky J. Rich models for steganalysis of digital images. IEEE Trans inform Forens Secur. 2012;7(3):868–82.
36. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018;801–818.
37. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv. 2017. https://doi.org/10.48550/arXiv.1706.05587.
38. Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer CC. The deepfake detection challenge (DFDC) preview dataset. arXiv. 2019. https://doi.org/10.48550/arXiv.1910.08854.
39. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. 2017;2223–2232.
40. Wu X, Xie Z, Gao Y, Xiao Y. Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2020;2952–2956.
41. Zhou P, Han X, Morariu VI, Davis LS. Learning rich features for image manipulation detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018;1053–1061.
42. Bappy JH, Simons C, Nataraj L, Manjunath B, Roy-Chowdhury AK. Hybrid lstm and encoder-decoder architecture for detection of image forgeries. IEEE Trans Image Process. 2019;28(7):3286–300.
43. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv. 2015. https://doi.org/10.48550/arXiv.1511.06434.
44. Adler J, Lunz S. Banach wasserstein GAN. Adv Neural Inform Process Syst. 2018;31.
45. Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANS for improved quality, stability, and variation. rXiv. 2017. https://doi.org/10.48550/arXiv.1710.10196.
46. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019;4401–4410.
47. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019;4401–4410.
48. He Z, Zuo W, Kan M, Shan S, Chen X. Attgan: facial attribute editing by only changing what you want. IEEE Trans Image Process. 2019;28(11):5464–78.
49. Thies J, Zollhofer M, Stamminger M, Theobalt C, Nießner M. Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016;2387–2395.
50. Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I. Synthesizing obama: learning lip sync from audio. ACM Trans Graph (ToG). 2017;36(4):1–13.
51. FaceSwap. https://github.com/deepfakes/faceswap
52. Yang X, Li Y, Lyu S. Exposing deep fakes using inconsistent head poses. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019;8261–8265. IEEE.
53. Li Y, Lyu S. Exposing deepfake videos by detecting face warping artifacts. arXiv. 2018. https://doi.org/10.48550/arXiv.1811.00656.

54. Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J. LSTM: a search space odyssey. IEEE Trans Neural Networks Learn Syst. 2016;28(10):2222–32.
55. King DE. Dlib-ml: a machine learning toolkit. J Mach Learn Res. 2009;10:1755–8.
56. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inform Process Syst. 2012;25.
57. Cozzolino D, Poggi G, Verdoliva L. 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. 2017;159–164.
58. Rahmouni N, Nozick V, Yamagishi J, Echizen I. Distinguishing computer graphics from natural images using convolution neural networks. In: 2017 IEEE Workshop on Information Forensics and Security (WIFS). IEEE. 2017;1–6.
59. Ciamarra A, Caldelli R, Becattini F, Seidenari L, Del Bimbo A. Deepfake detection by exploiting surface anomalies: The surfake approach. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops. 2024;1024–1033
60. Doloriel CT, Cheung N-M. Frequency masking for universal deepfake detection. arXiv. 2024. https://doi.org/10.48550/arXiv.2401.06506.
61. Chhabra S, Thakral K, Mittal S, Vatsa M, Singh R. Low-quality deepfake detection via unseen artifacts. IEEE Trans Artif Intell. 2023;5(4):1573–85.
62. Wang Y, et al. Multi-domain awareness for compressed deepfake videos detection over social networks guided by common mechanisms between artifacts. Comput Vision Image Understand. 2024;247:104072.
63. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR. 20196105–6114.
64. Luo Y, Zhang Y, Yan J, Liu W. Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021;16317–16326.
65. Zou Y, Luo C, Zhang J. DIFLD: domain invariant feature learning to detect low-quality compressed face forgery images. Complex Intell Syst. 2023;10(1):357–68.
66. Zhou P, Han X, Morariu VI, Davis LS. Two-stream neural networks for tampered face detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE. 2017;1831–1839.
67. Li Y, Lyu S. Exposing deepfake videos by detecting face warping artifacts. arXiv. 2018. https://doi.org/10.48550/arXiv.1811.00656.
68. Zhu X, Fei H, Zhang B, Zhang T, Zhang X, Li SZ, Lei Z. Face forgery detection by 3d decomposition and composition search. IEEE Trans Pattern Anal Mach Intell. 2023.
69. Zhao Y, Jin X, Gao S, Wu L, Yao S, Jiang Q. Tan-gfd: generalizing face forgery detection based on texture information and adaptive noise mining. Appl Intell. 2023;53(16):19007–27.