Exposing Deepfake Face Forgeries With Guided Residuals

Zhiqing Guo®, Gaobo Yang®, Jiyou Chen®, Graduate Student Member, IEEE, and Xingming Sun, Senior Member, IEEE

Abstract—For Deepfake detection, residual-based features can preserve tampering traces and suppress irrelevant image content. However, inappropriate residual prediction brings side effects on detection accuracy. Meanwhile, residual-domain features are easily affected by some image operations such as lossy compression. Most existing works exploit either spatial-domain or residual-domain features, which are fed into the backbone network for feature learning. Actually, both types of features are mutually correlated. In this work, we propose an adaptive fusion based guided residual network (AdapGRnet), which fuses spatial-domain and residual-domain features in a mutually reinforcing way, for Deepfake detection. Specifically, we present a fine-grained manipulation trace extractor (MTE), which is a key module of AdapGRnet. Compared with the prediction-based residuals, MTE can avoid the potential bias caused by inappropriate prediction. Moreover, an attention fusion mechanism (AFM) is designed to selectively emphasize feature channel maps and adaptively allocate the weights for two streams. Experimental results show that AdapGRnet achieves better detection accuracies than the state-of-the-art works on four public fake face datasets including HFF, FaceForensics++, DFDC and CelebDF. Especially, AdapGRnet achieves an accuracy up to 96.52% on the HFF-JP60 dataset, which improves about 5.50%. That is, AdapGRnet achieves better robustness than the existing works.

Index Terms—Deepfake detection, image forensics, guided residuals, attention fusion mechanism.

I. INTRODUCTION

FACE images, which contain rich personal identity information such as gender, race, age and emotion, have been widely-used as the biological modality for access control such as the entrance to restricted areas. However, the rapid development of computer graphics (CG) and image processing, we should not take the credibility of visual media for granted any more. Especially, the recent progress of Deepfake has enabled fake face images generated by Deepfake to create fake profiles on LinkedIn for malicious purposes.

In recent years, many works have been presented to expose face images generated by Deepfake. Among them, some works exploited hand-crafted features to represent biological signal inconsistencies, which include eye blinking [2], head pose [3] and imprecise geometry shapes such as weird ears and asymmetrical faces [4]. The other works introduced deep learning into face image forensics, which learn discriminative features from manipulation traces or biological inconsistencies. However, most existing deep learning based works exposed fake face images by directly learning spatial-domain features from RGB images [5], [6], and the other works improved detection accuracy by learning residual-domain features from prediction residuals [7], [8]. In essence, learning residual-domain features from prediction residuals is an effective way to highlight subtle manipulation traces and suppress irrelevant image content [9], [10]. Note that no matter the residuals are obtained by either fixed predictors [11], [12], [13] or the learning-based predictor [8], there usually exist some prediction biases for the existing predictors. This easily leads to coarse-grained residual-domain features. Apparently, if the residuals can be directly extracted instead of prediction, it will be an intuitive yet fine-grained way.

Though Deepfake can create scarily real face images from scratch, there are still some intractable challenges for Deepfake to keep global illumination and geometry consistency when generating fake face images. This can be alleviated by either explicitly modelling or implicitly learning from the data, yet an imprecise estimation or an improper regularization to the training cost might still lead to some subtle manipulation traces in facial regions. Fig. 1 compares the face image formation processes by camera capturing and Deepfakes. For fake face images generated by Deepfakes, it has been claimed that there exist unique pattern fingerprints or manipulation traces that are specific to different Deepfake models [14], [15]. That is, subtle manipulation traces are possible clues to expose Deepfake, and the key issue is to preserve subtle traces while suppressing image content. Fig. 2 shows the ideal process of extracting subtle manipulation traces

1[Online]. Available: https://www.datingadvice.com/online-dating/generated-photos-can-make-dating-sites-more-welcoming

2[Online]. Available: https://www.entrepreneur.com/article/335293
we present a fine-grained manipulation trace extractor (MTE). Compared with predicting residuals, MTE directly extracts the guided residuals, which can avoid the potential bias caused by inappropriate prediction. Both RGB images and guided residuals are fed into the backbone network to learn spatial features and residual features, respectively. Moreover, an attention fusion mechanism (AFM) is proposed for feature fusion, which fuses the learned features in a mutually reinforcing way by adaptively allocating the weights of two streams in terms of their cross entropy losses. The main contributions of this work are threefold.

1) Instead of predicting residuals, we propose a novel MTE by introducing the guided filter to directly extract the guided residuals, which preserves well subtle manipulation traces while suppressing image content. This expands the applications of the guided filter, and overcomes the potential bias in the prediction-based residuals.

2) An effective AFM is designed for feature fusion, which adaptively allocates the weights to spatial and residual features in terms of the cross entropy loss values of two streams. It also exploits the channel attention module to establish the dependency relationship between manipulation traces.

3) We propose a dual-stream model, namely AdapGRnet, by combining MTE and AFM with a backbone network for Deepfake detection under complex Internet scenarios. AdapGRnet learns both spatial-domain and residual-domain features, and exposes fake face images with either high-qualities or low-qualities. The experimental results on four public fake face datasets prove that the proposed AdapGRnet achieves better accuracy and robustness than the state-of-the-art works.

The remainder of this paper is organized as follows. Section II briefly introduces the related work. Section III presents the AdapGRnet model. Section IV reports the experimental results with some analysis. Conclusion is made in Section V.

II. RELATED WORK

A. Face Image Forgeries

In recent years, many works were presented for face image forgeries [18]. Among them, some works such as Glow [19] and GANimation [20] were proposed for facial expression manipulation. Glow is a generative flow using the invertible $1 \times 1$ convolution, which proves that the generative model optimized towards the plain log-likelihood objective can efficiently synthesize face images with realistic-looking facial expressions. GANimation is a GAN conditioning scheme based on action annotations, and its attention mechanism makes it be robust to background and lighting conditions. The other works, which include BEGAN [21], PGGAN [22] and StyleGAN [23], can synthesize hyper-realistic fake face images directly, or change facial attributes and styles such as gender, age, and hair color [24]. PGGAN, which grows the generator and discriminator progressively, can synthesize fake face images with the spatial resolution up to $1024 \times 1024$. By borrowing the idea of style transfer, StyleGAN proposes an alternative generator architecture for GAN,
which leads to an automatically learned, unsupervised separation of high-level attributes and stochastic variation in generated images. For facial video forgeries, there also exist some approaches such as FaceSwap, DeepFakes, Face2Face [25] and NeuralTextures [26], which replace faces, and even animate facial expressions of the target video from the source actor.

Note that the above-mentioned works achieve photo-realistic face image forgeries, but they might still leave some manipulation traces. Unlike modeling biological signal inconsistencies (such as eye blinking [2] and head poses [3]), implicitly learning features from inherent manipulation traces is a more general way, which can detect various face image forgeries.

B. Face Forensics Works

In recent years, the expose of the AI-enabled fake images has become an emerging topic in the field of image forensics [27], [28], [29], [30]. For the early AI-based face image forgeries, there are still some noticeable artifacts between a real face and a fake one, which include asymmetrical faces, bad teeth and weird ears. Matern et al. [4] exploited the visual artifacts such as imprecise geometry and reflection details to expose fake face images. To expose fake videos generated by Deepfake, Li et al. [2] exploited the inconsistency of eye blinking to design hand-crafted features. Later, deep learning was introduced into face image forgery detection [31]. Dang et al. [32] designed a customized convolutional neural network (CNN) to extract features from a tampered region, which can detect face images manipulated by BEGAN and PGGAN. Jeong et al. [33] proposed a bilateral high-pass filter, which amplifies frequency artifacts in the GAN-synthesized images, to improve generalization capability. Due to the lack of global constraints, face images by Deepfake have different facial part configurations from real faces. Yang et al. [34] presented to expose Deepfake by using landmark locations. Li et al. [35] exploited the disparities in color components to expose Deepfake-generated face images. To detect facial video forgeries, Afchar et al. [5] presented a compact MesoNet, in which a low number of layers are focusing on the mesoscopic properties of images. Zhao et al. [36] presented a dual-stream CNN to simultaneously capture semantic and noise information in face images to expose Deepfake.

There are several works that exploit residual-domain features. Zhou et al. [37] proposed a two-stream Faster R-CNN for image tampering region detection, in which the Spatial Rich Model (SRM) for image steganalysis is exploited to learn residual-domain features. Inspired by the concept of residuals, Bayar et al. [38] developed a constrained convolutional layer, which jointly suppresses image content and highlights manipulation traces, to learn residual-domain features for general-purpose image forensics. To expose fake face images generated by PGGAN, Mo et al. [7] proposed to transform the original images into the residual images via high pass filter, which are fed into a carefully-designed CNN for binary classification. We also proposed an Adaptive Manipulation Traces Extraction Network (AMTEN), which exploits an adaptive convolution layer to predict residuals [8]. The residuals are reused in subsequent layers to maximize manipulation traces by updating weights during back-propagation pass.

In essence, these existing works extract manipulation traces from residuals via a predictor, which may bring potential bias due to the prediction method. In this work, we introduce the guided filter to protect image content. Then, the manipulation traces are highlighted by subtracting the content features from the original image, which avoids the potential prediction bias issue. Nevertheless, learning features from residuals also has its own limitations and drawbacks. That is, the manipulation traces in low-quality face images might have been destroyed, exploiting only residual-domain features easily leads to poor detection accuracy. To address this issue, we fuse the features learned from spatial stream and residual stream to improve detection accuracy and robustness when dealing with high-quality and low-quality face images.

C. Feature Fusion

For image forensics, fusing two or more types of features can improve detection accuracy and robustness [39]. The existing feature fusion strategies mainly include sum, max, min and concatenation [40], [41]. In recent years, Bahdanau et al. [42] proposed an attention-based fusion for machine translation. Later, the attention based feature fusion has attracted extensive attentions in various computer vision tasks such as image classification [43], scene segmentation [44] and fine-grained image recognition [45]. Wang et al. [46] proposed a non-local attention mechanism to capture long-range dependencies for video classification tasks. Huang et al. [47] proposed a criss-cross attention module to capture contextual information for semantic segmentation. Fu et al. [44] proposed a dual attention network to adaptively integrate local features with their global dependencies based on the self-attention mechanism. Zhu et al. [48] also presented an adaptive attention and residual refinement network, namely AR-Net, for copy-move forgery detection. Bonet-tini et al. [49] designed an attention-based module by using $1 \times 1$ convolution and the sigmoid function to promote the network to pay attention to the most relevant portions of the feature maps. Yang et al. [50] proposed a self-texture attention mechanism as the skip connection between encoder and decoder to mine the texture trace.

However, these attention mechanisms are only used in single stream networks for better feature representation, without considering the fusion between multi-streams. For face forensics, spatial-domain and residual-domain features have their own advantages and disadvantages, respectively. In fact, they are complementary to each other. In this work, an AFM scheme is proposed to enable the CNN model to selectively learn more residual or spatial features when dealing with high-quality and low-quality face images, respectively. In addition, we exploit the channel attention module to enhance feature representation capability. Specifically, AFM treats different types of features unequally. Thus, it provides additional flexibility for feature fusion when dealing with different types of features, which enables

---

3[Online]. Available: https://github.com/MarekKowalski/FaceSwap

4[Online]. Available: https://github.com/deepfakes/faceswap
final features be more discriminative for face forensics tasks under complex scenarios.

III. METHODOLOGY

A. Overview

1) Complementarity Analysis: For image forensics tasks, spatial-domain features and residual-domain features are either hand-crafted or implicitly learned from image data in an end-to-end manner. In general, spatial-domain features contain rich manipulation traces for image forensics, but they might also contain image content information that are suitable for image classification and recognition. However, image content information does not do benefit to and even has side impact on image forensics. As we know, image forensics is to detect image forgeries by exposing their subtle manipulation traces that are independent of image contents. Luckily, implicitly learning features from residuals can suppress the side effects of irrelevant image content. Nevertheless, the manipulation traces left by image forgeries might have been partially lost in the residuals when suppressing image content. Especially when candidate images are low-quality, the manipulation traces are partially destroyed, which easily leads to performance loss.

As analyzed above, we exploit the complementarity of spatial and residual features for image forensics. Specifically, for low-quality input images, more spatial-domain features, which contain rich texture information, are exploited to remedy the deficiencies of the residual-domain features. For high-quality images, more residual-domain features, which can alleviate the side effects of irrelevant image content, are exploited, simply because they are more discriminative than the spatial-domain features, especially when they are implicitly learned from manipulation traces via CNNs. Moreover, the guided residuals preserve the manipulation traces much better than the prediction based residuals. Thus, the guided residuals are exploited for learning residual-domain features.

2) Dual-Stream Architecture: In this work, our main goal is to identify fake face images under complex Internet scenarios. To enhance the detectors' robustness when dealing with face images with distinct visual qualities, a dual-stream AdapGRnet is proposed, which learns spatial-domain features and residual-domain features in a mutually reinforcing way. Fig. 3 is the framework of the AdapGRnet.

Given an input RGB face image, we first convert it into a guided residual image by MTE. Then, both RGB images and guided residual images are fed into the backbone network for feature learning, respectively. Since the CNN based approaches are data hungry and small training data might lead to over-fitting, pre-training the CNN model is an effective way to address this issue. In this work, ResNet-18 [51], which has been pre-trained on the ImageNet dataset [52], is exploited as the backbone network for two streams. Since ResNet-18 is only used for feature learning in our framework, we remove full connection layers from it. The output of each stream is a 512-dimension feature. To improve the detectors’ robustness under different scenarios, the spatial and residual features learned from two streams are fused via AFM to emphasize the relationship between feature channel maps and adaptively adjust the weights for them.

To formulate the proposed AdapGRnet, a quadruple variable $M$ is introduced. That is,

$$M = (S_{rgb}, S_{gr}, F, D)$$

where $S_{rgb}$ and $S_{gr}$ are the spatial stream and the residual stream, respectively. $F$ is the AFM and $D$ is the detector. The learned features $f_{rgb}$ and $f_{gr}$ from each stream should have the same dimension, which are input into AFM for feature fusion. The fused feature $F_{attention}$ is expressed as

$$F_{attention} = F(f_{rgb}, f_{gr})$$

Thus, AdapGRnet can be formulated as an optimization problem. That is,

$$\min \frac{1}{N} \sum_{i=1}^{N} L_i[D(F(f_{rgb}, f_{gr})), y_i]$$

where $N$ is the number of samples, $L_i(\cdot, \cdot)$ and $y_i$ represent the cross-entropy loss function and the one-hot encoding label vector, respectively.

---

*Fig. 3. Overview of the proposed framework for detecting fake faces.*
B. Manipulation Trace Extractor

Up to now, existing works use either a fixed predictor [11] or a learning-based predictor [38] to obtain prediction residuals. That is, the residuals are obtained by subtracting the predicted pixel values from the original pixel values. Let $p$ be the input image and $P(\cdot)$ be the predictor. The prediction residuals $R$ can be expressed as

$$ R = P(p) - p \quad (4) $$

For image forensics, many works learn statistical features from the prediction-based residuals [53, 54, 55]. Thus, accurate residuals are critical for image forensics. To avoid the potential bias issue for the prediction-based residuals, we suppress image content and highlight manipulation traces by introducing the guided filter.

The guided filter is an edge-preserving smoothing operator, which computes the output by the content of a guidance image $I$ [16]. The guidance image should be the input image itself or another arbitrary image. In this work, the guidance image uses the input face image to preserve image content and filter out the manipulation traces in flat regions. The key assumption behind the guided filter is a local linear model between the guidance image $I$ and the output image $q$. We assume that $q$ is a linear transform of $I$ in a window $\omega_k$ centered at the pixel $k$. That is,

$$ q_i = a_k I_i + b_k, \quad \forall i \in \omega_k \quad (5) $$

where $a_k$ and $b_k$ are some linear coefficients that are assumed to be constants in $\omega_k$. In reference [16], their values are defined as

$$ a_k = \frac{1}{|\omega|} \sum_{i \in \omega_k} I_i p_i - \mu_k \bar{p}_k $$

$$ b_k = \bar{p}_k - a_k \mu_k \quad (7) $$

where $|\omega|$ is the number of pixels in $\omega_k$, $\bar{p}_k$ is the mean of $p$ in $\omega_k$, $\sigma_k^2$ and $\mu_k$ are the variance and mean of $I$ in $\omega_k$, respectively. $\epsilon$ is a regularization parameter. After obtaining the linear coefficients $\{a_k, b_k\}$, we can compute the filtering output image $q_i$ by Equation (5).

However, for a pixel $i$ that is involved in all the overlapping windows $\omega_k$, its output $q_i$ in Equation (5) is not identical when it is computed in different windows. A simple strategy is to average all the possible values of $q_i$. That is, after computing $\{a_k, b_k\}$ for all the windows $\omega_k$ in the image, the filtering output $q_i$ is obtained as follows.

$$ q_i = \frac{1}{|\omega|} \sum_{k \in \omega_k} (a_k I_i + b_k) \quad (8) $$

Let $\bar{a}_i$ and $\bar{b}_i$ be the average coefficients of all windows overlapping $i$. Equation (8) can be rewritten as

$$ q_i = \bar{a}_i I_i + \bar{b}_i \quad (9) $$

As claimed earlier, a fake face image can be viewed as the combination of the content features $C$ and the manipulation traces $T$ (refer to Fig. 1). Thus, the input face image $p$ can be simply expressed as

$$ p = C + T \quad (10) $$

The local linear model behind the guided filter ensures that the filtering output $q$ preserves well the image content features $C$ and removes the manipulation traces $T$ in flat regions. Thus, the filtering output $q$ of the input image $p$ is equal to $q = C$. Then, the guided residuals $R_{gr}$ is obtained by subtracting the filtering output $q$ from the input image $p$. Fig. 4 shows the extraction of the guided residuals. Note that we take the absolute value to avoid possible negative value. That is,

$$ R_{gr} = |p - q| = T \quad (11) $$

Equation (4) is a general pipeline for predicting the residuals. The guided filter can be treated as a fixed predictor, but it does not exploit the correlations between local pixels for prediction. We exploit the guided filter to suppress image content features and highlight manipulation traces. This is an intuitive yet more suitable way to obtain the guided residuals. Fig. 5 compares the residuals obtained by high-pass filter [7] and MTE for both high-quality and low-quality face images. The first row shows the color images, including real face images and tampered images. The second row shows the prediction residuals obtained by high-pass filter, in which there are still some details about image content. The third and fourth rows show the guided residuals, in which most image contents are removed. In addition, we further enlarge the guided residuals by noise analysis, and the resultant guided residuals are shown in the fifth row. From Fig. 5, we can observe that for high-quality input images, the residuals have rich micro-texture details, and there are obvious differences among the residuals of different forged images. This is very beneficial for the forensics model to learn discriminative features. For the low-quality input image, we can see from the noise analysis in the fifth row that there are many white blocky textures in the residual images. These similar textures make it difficult for the forensics model to discriminate the differences among different forged images. This also proves that the manipulation traces in low-quality images are laundered.

---

5[Online]. Available: https://29a.ch/photo-forensics/#noise-analysis
Fig. 5. Comparison of prediction residuals and guided residuals. To highlight the differences of the guided residuals among different facial forgery techniques, the texture details in the green boxes are enlarged by 5 times, and noisy analysis is made for the guided residual images in the red boxes.

C. Attention Fusion Mechanism

Feature fusion refers to combining two or more feature vectors to form a single feature vector, which should be more discriminative than any of the input feature vectors. For image forensics, feature fusion improves both detection accuracy and robustness. In this work, both spatial-domain features and residual-domain features are learned by the dual-stream AdapGRnet. For each stream, a 512D feature channel map, which can be treated as the response to the manipulation traces left by face image forgeries, is learned. Since feature fusion is to improve feature representation capability, how to exploit the dependency between these responses is a key issue to expose the manipulation traces. AdapGRnet is dedicated to expose Deepfake spreading over Internet, which might have high or low qualities. As claimed earlier, the spatial stream is preferable for learning features from input images with low-qualities, whereas the residual stream is suitable for learning features from high-quality images. How to adaptively allocate the weights to two streams is the key to improve detection accuracy and robustness when dealing with face images with different visual qualities. Thus, we design an AFM, whose structure is shown in Fig. 6.

Firstly, the dependence between feature channel maps is exploited to improve feature representation capability. Let $f_{rgb} \in \mathbb{R}^{C \times H \times W}$ represents the spatial-domain features, where $C$, $H$ and $W$ are the channel, height and width, respectively. $f_{rgb}$ is reshaped into $\mathbb{R}^{C \times N_p}$, where $N_p$ is the number of pixels. Then, the reshaped features are matrix multiplied with the transpose of $f_{rgb}$, and the result is further processed with the softmax function to obtain the spatial attention matrix $M_{rgb} \in \mathbb{R}^{C \times C}$.

$$m_{rgb}^{ij} = \frac{\exp(f_{rgb}^i \cdot f_{rgb}^j)}{\sum_{i=1}^{C} \exp(f_{rgb}^i \cdot f_{rgb}^j)}$$  \hspace{1cm} (12)

where $m_{rgb}^{ij}$ represents the influence of the $i^{th}$ channel on the $j^{th}$ channel. Similarly, the residuals attention matrix $M_{gr} \in \mathbb{R}^{C \times C}$ can be obtained as follows.

$$m_{gr}^{ij} = \frac{\exp(f_{gr}^i \cdot f_{gr}^j)}{\sum_{i=1}^{C} \exp(f_{gr}^i \cdot f_{gr}^j)}$$  \hspace{1cm} (13)
The reshaped \( \{ f_{rgb}, f_{gr} \} \) is also matrix multiplied with the transpose of \( \{ M_{rgb}, M_{gr} \} \) to obtain a new feature representation \( \{ f'_{rgb}, f'_{gr} \} \).

Secondly, we adopt cross-entropy loss \( L \) to generate the spatial stream loss \( L_1 \) and the residual stream loss \( L_2 \), as follows.

\[
L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} y_i^{(j)} \log(D_i^{(j)})
\]

(14)

where \( y_i^{(j)} \) represents the label of the \( i^{th} \) image in the \( j^{th} \) class, \( D_i^{(j)} \) denotes the output of the detector, \( N \) is the number of samples, and \( K \) is the number of neurons in the output layer. \( L_1 \) and \( L_2 \) are used to adaptively allocate the weights of two streams, respectively. Actually, the loss value of each stream should be inversely proportional to its weight. Thus, \( L_1 \) and \( L_2 \) are fed into the softmax function to define the weights \( \alpha_i \), which are subject to \( \sum_i \alpha_i = 1 \).

\[
\alpha_i = 1 - \frac{\exp(L_i)}{\sum_j \exp(L_j)}
\]

(15)

During the back-propagation pass, the weights are iteratively updated through the loss value of each epoch.

Thirdly, we aggregate the two streams via the element-wise sum operation as follows

\[
F_{\text{attention}} = \alpha_1 f'_{rgb} + \alpha_2 f'_{gr}
\]

(16)

In the training stage, AFM obtains the weights of two streams via the loss values. Thus, we only need to fix the corresponding weights in the testing stage, so as to realize the inference.

IV. EXPERIMENTAL RESULTS

A. Datasets

The Hybrid Fake Face (HFF) [8] dataset contains 155 k high-quality face images, which include both real and fake face images. For the real face images, they come from the CelebA dataset [56], the CelebA-HQ dataset [22] and YouTube video frames [57], respectively. The fake face images are obtained by five typical facial forgery techniques including PGGAN [22], StyleGAN [23], Face2Face [25], Glow [19], and StarGAN [24]. Note that both PGGAN and StyleGAN can synthesize face images that do not exist up to the spatial resolution of 1024 \( \times \) 1024. Face2Face and Glow were presented for facial expression transferring, whereas StarGAN was proposed for changing face styles such as hair color and gender by multi-domain image-to-image translation. Since the HFF dataset is an image dataset containing only faces (excluding background information), all images are resized to 128 \( \times \) 128 for experiments. Some samples are randomly selected from the HFF dataset, which are shown in Fig. 7. To verify the detectors’ robustness under different scenarios, the same post-processing operations are applied to the HFF dataset following the work [8]. Specifically, the quality factor of JPEG compression is set to 60 (JP60), which can launder manipulation traces by weakening image quality. The kernel size of the mean filtering is set to 5 \( \times \) 5 (ME5), which smoothes image details including the manipulation traces due to image blur. Thus, three types of data from the HFF dataset are used in the experiment, namely HFF-Raw, HFF-JP60 and HFF-ME5.

The FaceForensics++ (FF) [17] dataset is made up of 1,000 real video sequences that have been manipulated with four facial forgeries including Deepfake, Face2Face, FaceSwap and NeuralTextures, respectively. The FF dataset provides visually lossless high-quality (HQ) videos (the quantization parameter is set with a constant of 23) and visually lossy low-quality (LQ) videos (the quantization parameter is set with a constant of 40 for H.264/AVC compression), which are widely-used in social networks. In the experiment, both FF-HQ and FF-LQ datasets are used for experimental evaluation. To reduce the similarity between frames, one frame is saved every two frames in each video sequence. We extract 50 frames from each real video sequence. Since there are totally 4,000 fake video sequences, 15 frames are extracted from each fake video sequence to balance the real/fake face data. Fig. 8 compares different types of face images in the FF-HQ and FF-LQ datasets.

The DeepFake Detection Challenge (DFDC) [58] dataset contains more than 100 k real and fake video sequences (about 470 GB). There are various actors with diverse attributes such as age, gender and skin-tone. They are recorded with arbitrary background, thus bringing visual diversities. Fig. 9 illustrates some face images from the DFDC dataset. Note that the fake videos are manipulated by two unknown facial forgery algorithms. Thus, only a binary classification task is made for the trustworthiness of the face images. Specifically, we randomly select 6,698 video sequences. For each video sequence, one frame is saved every thirty frames. Then, we select a maximum of ten frames for each video. Thus, there are totally 31,873 real video frames and 30,945 fake video frames in our experiments.

The CelebDF dataset is a challenging dataset, which contains 890 real face videos and 5,639 hyper-realistic Deepfake videos. For real videos, one frame is saved every three frames, and 34
frames are extracted from each video. For fake videos, one frame is saved every twenty frames, and 6 frames are extracted from each video. We randomly select 30,000 real and 30,000 fake video frames from these extracted video frames for cross-dataset evaluation.

In the above three video datasets (namely FF, DFDC and CelebDF), we adopt the open Face Recognition Library\footnote{[Online]. Available: https://github.com/ageitgey/face_recognition} to capture facial regions in video frames to remove useless background information. For the FF dataset, all extracted facial images are resized to $128 \times 128$. For the DFDC and CelebDF datasets, all facial images are resized to $224 \times 224$ for experiments. Table I summarizes the details of the datasets used for model training and testing.

In the experiment, we consider both multi-classification and binary classification tasks to evaluate the detection models. We can observe from Table I that the numbers of real and fake face images in four datasets are roughly balanced, which can avoid the issue of the biased classifiers. However, for multi-classification tasks, it is still difficult to balance the number of images in each category due to the great differences for different categories. Thus, when performing multi-classification tasks on the HFF dataset, we set the real images into three categories (namely CelebA, CelebA-HQ and video frames) to balance the data. The reason behind this is that real face images come from the low-quality CelebA dataset, the high-quality CelebA-HQ dataset and compressed video frames. These three categories of real face images have completely different texture features, while fake face images also have manipulation trace features specific to different forgery techniques. Thus, the number of images in each category is effectively balanced to avoid the bias for model training. For the FF dataset, we randomly select 12,500 and 2,500 real face images from the training and testing sets for experiments, so as to keep the balance for multi-classification tasks on the FF dataset.

### B. Experimental Settings

1) **Evaluation Criterion:** For image forensics tasks, there are two commonly-used metrics, namely accuracy rate (ACC) and area under the ROC curve (AUC), for performance evaluation. In our experiments, ACC is used to evaluate the detectors’ accuracy. AUC is used to measure the performance of binary classification tasks.

2) **Baselines:** Several detection models with open source codes are selected as the baseline models. For fair comparison, they are trained from scratch on our dataset and evaluated by using the same testing set.

- **Meso-4** [5]: It mainly exploits the mesoscopic properties of face images to detect fake videos generated by Deepfake.
- **MesoInception-4** [5]: It is another deep learning based work that has much better performance than Meso-4.
- **HighPass** [7]: It extracts residuals via high pass filters. We exploit the high pass filter that achieves the best detection accuracy for comparisons in our tasks.

| Datasets | Classification | Data type     | Face Frames/Images |
|----------|----------------|---------------|--------------------|
|          | Real           | CelebA        | 20,000            |
|          | Real           | CelebA-HQ     | 8,000             |
|          |                 | Video frames  | 20,000            |
|          | Fake           | Face2Face     | 20,000            |
|          | Fake           | PGGAN         | 8,000             |
|          | Fake           | StarGAN       | 20,000            |
|          |                | SynGAN        | 8,000             |
| HFF      | Real           | Video frame   | 50,000            |
|          | Fake           | Face2Face     | 12,500            |
|          | Fake           | FaceSwap      | 12,500            |
|          |                | Deepfake      | 12,500            |
|          |                | NeuralTextures| 12,500            |
| DFDC     | Real           | Video frame   | 24,873            |
|          | Fake           | Deepfake      | 23,945            |
| CelebDF  | Real           | Video frame   | 30,000            |
|          | Fake           | Deepfake      | 30,000            |

Fig. 8. Face video frames from the FF dataset. The red box indicates that the face image belongs to the HQ dataset and the green box indicates that the face image belongs to the LQ dataset.

Fig. 9. Face video frames from the DFDC dataset.

---

64[Online]. Available: https://github.com/ageitgey/face_recognition
4) MISLnet [38]: A constrained convolution layer, which is equivalent to an adaptive residual predictor, is designed. MISLnet is a pioneer work towards universal image forgery detection. It also serves as a universal detector here for experimental comparisons.
5) XceptionNet [59]: It is originally presented for image classification. However, it is exploited for facial forgery detection and achieves desirable performance in [17].
6) Capsule [60]: It extends the capsule networks to expose various kinds of spoofs such as replay attack and computer generated images/videos.
7) AMTENet [8]: It presents an adaptive manipulation traces extraction network (AMTEN) for facial forgery detection, which achieves the state-of-the-art performance.
8) RFMnet [61]: It designs a representative forgery mining scheme, which realizes state-of-the-art face image forgery detection.

3) Implementation Details: The proposed AdapGRnet is implemented under the PyTorch framework. We use one NVIDIA GeForce GTX 1080 Ti GPU to train the model and employ the ADAM with the decay parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$) for network optimization. An exponential learning rate ($\gamma = 0.5$) is used, and the initial learning rate is 0.0005. For each group of experiments, all the models are trained for 20 epochs, and then a fair comparison is made on the testing set. To improve model generalization capability and robustness, some data augmentation methods, which include horizontal flip, rotation and normalization, are adopted to produce more training data from the original data.

C. Ablation Study
MTE and AFM are two main innovations for the proposed AdapGRnet. Three datasets including HFF, FF and DFDC are used for ablation study. The HFF and FF datasets contain fake face images generated by various face forgeries, whereas there are only two types of real and fake images for the DFDC dataset. Thus, we used the HFF and FF datasets for multi-classification and DFDC dataset for binary classification tasks to better evaluate AdapGRnet. Note that multiple classification is to identify which technique is used to manipulate face images, which is a fine-grained recognition problem. For binary classification, it is used to judge the authenticity of face images, yet it has its own advantage for testing generalization capability.

From the experimental results reported in Table II, we can observe that when only MTE is equipped, the network can achieve the best detection accuracy on high-quality data (HFF-Raw), which also proves that residual features are very good at capturing manipulation traces in high-quality images. For the remaining datasets, they are all low-quality data obtained from datasets with post-processing operations or compressed videos. We design an AFM method to make up for the defect that residual features are not suitable for low-quality data. It can be observed that when MTE and AFM are equipped to AdapGRnet, the best detection accuracy can be obtained on low-quality datasets. It also shows that the adaptive fusion of residual-domain and spatial-domain features effectively improves the robustness of the network in complex scenarios.

D. Key Module Comparison

1) Comparison for Residuals: In existing works, there are several residual extraction methods, which include the fixed predictors (high pass filter [7] and the SRM filter [37]) and the learning-based predictors (AMTEN [8] and constrained convolution layer [38]). They are selected as the baselines to make comparisons with MTE. For fair comparison, we compare the detection accuracies by using the same backbone network following the experimental benchmark in [8]. Since the residual features are not suitable for low-quality images, because they are easily affected by lossy compression operation. Thus, we only use the high-quality HFF-Raw dataset to evaluate the existing residual extraction methods.

Table III reports the detection accuracies when using different residual extraction methods. High pass filter and SRM filter, as the fixed predictors, achieve the detection accuracies of 97.50% and 97.49%, respectively. For the learning based predictors, the constrained convolution layer achieves an accuracy of 95.24%, and AMTEN achieves an accuracy up to 98.52%. The constrained convolution layer works by resetting some specific coefficients in the convolutional kernel after each iteration, whereas AMTEN adaptively updates the coefficients in the convolutional kernel after each iteration. Thus, AMTEN obtains better residuals than the constrained convolution layer, resulting an accuracy improvement of 3.28%. Unlike the fixed predictors and the learning-based predictors, MTE achieves an accuracy rate up to 99.04%. Thus, we can observe that the proposed MTE achieves better accuracies than the fixed predictors and the learning-based predictors. The inherent reason behind this is that the guided residuals have not the potential bias brought by the prediction-based works, which characterize well the manipulation traces left by various face forgeries.

To prove the advantages of the guided residuals obtained by MTE, we also visualize the features learned by different pre-processing methods [62], as shown in Fig. 10. Though the constrained convolution layer greatly suppresses image content, the manipulation traces left in the residuals are fragile. For high pass filter and AMTEN, there still preserve some face contour
information. However, compared with the constrained convolution layer, the residuals obtained by high pass filter and AMTEN reflects better the manipulation traces. Thus, high pass filter and AMTEN achieves better detection accuracies, which are 97.50% and 98.52%, respectively. From the fifth column of Fig. 10, we can observe that the guided residuals almost completely suppress image content, and thus the CNN model learns more discriminative manipulation trace features from the guided residuals.

2) Comparison of Feature Fusion Works: AFM is designed for feature fusion to further improve the detection performance. To show the effectiveness of the AFM, we conduct some experiments by comparing it and four common feature fusion strategies, namely max, min, sum and concatenation. Among them, both max fusion and min fusion compare spatial-domain features and residual-domain features in an element-wise way to return a larger value or a smaller value, respectively. The sum fusion aggregates spatial-domain features and residual-domain features in an element-wise sum way. The concatenation fusion is to splice spatial-domain features and residual-domain features in columns. These fusion methods simply fuse the features of the two streams, without considering adaptive feature fusion according to different data types. The proposed AFM is presented to make up for the defect that residual features are not suitable for low-quality images. Thus, three low-quality datasets (such as HFF-JP60, FF-LQ and DFDC) are selected for performance evaluation. As with Section IV-C, we still use both multi-classification and binary classification tasks for experiments. Table IV reports the detection results when using different feature fusion methods. From it, we can observe that AFM effectively improves the defects of residual features and realizes robust face forgery detection.

E. Comparison With Baseline Approaches

The proposed AdapGRnet is compared with the state-of-the-art works, and the HFF, FF and DFDC datasets are selected for experiments. For the HFF and FF datasets, we provide both multiple classification and binary classification. To show the performance difference between two streams in different types of data, we also provide the detection results of spatial stream and residual stream of AdapGRnet, which are named as “AdapGRnet-Spa” and “AdapGRnet-Res,” respectively.

1) Results on the HFF Dataset: Table V reports the multiple classification accuracies when using different networks. From it, we can observe that the shallow networks including Meso-4 and MesoInception-4 can not achieve desirable results under three scenarios (Raw, JP60, and ME5), and their average accuracies are 70.31% and 74.25%, respectively. Among these baseline models, both XceptionNet and Capsule are deep CNNs with complex structures, and they achieve the average accuracies of 88.89% and 93.74% under three scenarios, respectively. HighPass exploits shallow CNN to learn features from the residuals obtained by the fixed predictor. It achieves an average accuracy of 79.78%, which is about 5.53% higher than that of MesoInception-4. MISLnet and AMTENnet are also shallow convolutional networks, yet their inputs are the learning based residuals. The residuals are updated iteratively during the back-propagation pass, from which more discriminative features can be learned. As a result, MISLnet and AMTENnet achieve the average accuracies of 86.38% and 93.99%, respectively. Note that the accuracies are close to slighter or better than that of XceptionNet. From this, we can conclude that for the deep learning based face forensics tasks, extracting the residuals as...
TABLE VI

| Methods          | HPF-Rev | HPF-FP60 | HPF-MI5S | ACC   | AUC   | ACC   | AUC   |
|------------------|---------|----------|----------|-------|-------|-------|-------|
| Meso-4 [5]       | 76.83%  | 75.54%   | 62.40%   | 62.45%| 63.52%| 64.22%|
| MesoInception-4 [5] | 94.33%  | 94.88%   | 73.63%   | 75.29%| 84.43%| 85.56%|
| HighPass [7]     | 88.06%  | 87.23%   | 72.97%   | 72.55%| 76.88%| 78.62%|
| MISNet [38]      | 93.71%  | 94.89%   | 87.37%   | 89.91%| 84.76%| 86.38%|
| XceptionNet [59] | 92.82%  | 92.55%   | 74.13%   | 74.83%| 76.62%| 77.10%|
| Capsule [60]     | 95.66%  | 98.01%   | 98.12%   | 98.22%| 92.63%| 96.14%|
| AMTENet [8]      | 97.66%  | 97.83%   | 98.31%   | 97.31%| 81.36%| 82.11%|
| RFMnet [61]      | 98.52%  | 98.75%   | 98.30%   | 98.32%| 98.16%| 97.96%|
| AdapGRNet-Spa.   | 66.20%  | 73.34%   | 94.16%   | 97.66%| 67.40%| 72.62%|
| AdapGRNet-Res.   | 97.66%  | 99.51%   | 94.39%   | 98.34%| 95.20%| 99.71%|
| AdapGRNet        | 99.44%  | 99.99%   | 99.63%   | 97.94%| 99.25%| 99.98%|

TABLE VII

| Methods          | FF-HQ | FF-LQ | Average |
|------------------|-------|-------|---------|
| Meso-4 [5]       | 23.30%| 22.50%| 22.90%  |
| MesoInception-4 [5] | 61.31%| 44.09%| 52.70%  |
| HighPass [7]     | 68.05%| 39.50%| 53.78%  |
| MISNet [38]      | 84.09%| 68.25%| 76.17%  |
| XceptionNet [59] | 67.72%| 59.74%| 63.73%  |
| Capsule [60]     | 97.01%| 95.98%| 96.50%  |
| AMTENet [8]      | 89.26%| 69.38%| 79.32%  |
| RFMnet [61]      | 94.58%| 85.64%| 90.11%  |
| AdapGRNet-Spa.   | 91.93%| 91.55%| 91.74%  |
| AdapGRNet-Res.   | 74.02%| 60.67%| 67.27%  |
| AdapGRNet        | 97.88%| 97.30%| 97.57%  |

appropriate as possible can improve the feature representation capability of the CNN model. For RFMnet, it combines data augmentation with XceptionNet, which achieves better detection results. Furthermore, the proposed AdapGRnet combines residual-domain and spatial-domain features in a mutually reinforced way, achieving an average accuracy of 98.63%, which is 2.5% higher than that of the state-of-the-art work with the best performance.

Table VI reports the results for the binary classification task, which are similar to the results in Table V. In the HFF dataset, most face images are generated by GANs that leave unique manipulation traces in the global images, instead of just manipulated local regions, which can be easily extracted by MTE. As shown in Tables V and VI, the proposed AdapGRnet has achieved much better accuracies than the baseline methods.

2) Results on the FF Dataset: The face images in the FF dataset have relatively poor visual qualities, simply because they are obtained from compressed video frames. Thus, the FF dataset is much more challenging to detect face image forgeries than the HFF dataset. Table VII reports the experimental results, which are similar to the multiple classification results reported in Table VI. It is well-known that compressed video frames launder the manipulation traces in the face image, which brings great challenges. By using capsule mechanism, Capsule network still maintains good modeling ability when dealing with low-quality data. Although the guided residuals are not applicable to low-quality data, we avoid the defects of the guided residuals as much as possible via the AFM, thus achieving better performance than the Capsule.

3) Results on the DFDC Dataset: The DFDC dataset is made up of face images that captured from video frames. There are only two types of face images, namely real and fake. For face images in the DFDC dataset, it is a relatively easy task to make binary classification. Table IX reports the experimental results, which compares the ACC and AUC values among the proposed approach and the existing baseline works. From the results, the proposed AdapGRnet still achieves the best accuracy of 99.08% and the highest AUC value of 99.88%.

F. Generalization Ability

To validate the generalization capability of the model, we introduce the latest benchmark, which evaluates the generalization ability of existing methods on the CelebDF dataset [63]. We use the FF dataset with two quality levels for binary classification training to obtain the pre-trained classifiers, which are referred to as AdapGRnet-HQ and AdapGRnet-LQ for cross-dataset evaluation on the CelebDF dataset. Table X reports the experimental results. We can observe that AdapGRNet achieves better cross-dataset evaluation results than the existing works, especially AdapGRNet-LQ achieves the highest AUC value up to 71.5%. For AdapGRNet-HQ trained on the FF-HQ dataset, its cross-dataset evaluation results are not competitive results. We 63.73%, which is quite far away from satisfaction. The residuals based works including HighPass, MISNet and AMTENet achieve the average accuracies of 53.78%, 76.17% and 79.32%, respectively. For Capsule and RFMnet, they both achieve competitive results, with an average accuracy of 96.50% and 90.11%, respectively. For binary classification, Table VIII reports the experimental results, which are similar to the multiple classification results reported in Table VII.
suspect that the slight compression operation may not completely remove the subtle manipulation traces in the image, which makes the classifier trained on the FF-HQ dataset learn more specific manipulation features specific to the FF dataset, but not the common artifacts left by various face forgeries. On the contrary, the FF-LQ dataset lacks the specific manipulation features attributed to the FF dataset. Thus, the AdapGRNet-LQ classifier trained on the FF-LQ dataset learns more common artifact features, thus achieving better cross-dataset evaluation results. We also notice that both Xception-HQ and Xception-LQ classifiers have achieved desirable results. As we know, Xception has a more complex network structure than AdapGRnet, which preserves well subtle manipulation traces. Compared with the existing prediction based residuals, the guided residuals address well the defects of the existing prediction based residuals, the guided residuals, which promotes the detection model to keep a desirable accuracy for low-quality data. However, due to matrix multiplication, our fusion mechanism is costly. If the size of the input image is 224, AdapGRNet has 23.21 M model parameters and 3.64GMac FLOPs. Compared with RFMnet (parameters: 20.81 M, FLOPs: 4.59GMac), AdapGRNet has similar parameters and computational complexity. However, they all belong to large models, which are difficult to deploy on some platforms with limited computing resources (such as mobile devices and embedded devices).

V. CONCLUSION

Deepfake detection under complex Internet scenarios is a challenging problem in the community of image forensics due to diverse face forgeries and low-quality visual content. In this work, we proposed an AdapGRNet for Deepfake detection under complex scenarios. Specifically, MTE obtains the guided residuals, which preserves well subtle manipulation traces. Compared with the existing prediction based residuals, the guided residuals avoid the potential bias. We also designed an AFM that selectively emphasizes the relationship among feature channel maps and adaptively allocates the weights for two streams, further improving the detection performance. The extensive experimental results on four open datasets show that the proposed AdapGRNet achieves much better accuracies and robustness than the state-of-the-art works, which is promising for face image forensics under complex scenarios.

REFERENCES

[1] F. Peng, L.-P. Yin, L.-B. Zhang, and M. Long, “CGR-GAN: CG facial image regeneration for Antiforensics based on generative adversarial network,” IEEE Trans. Multimedia, vol. 22, no. 10, pp. 2511–2525, Oct. 2020.
[2] Y. Li, M.-C. Chang, and S. Lyu, “In ictu oculi: Exposing AI created fake videos by detecting eye blinking,” in Proc. IEEE Int. Workshop Inf. Forensics Secur., 2018, pp. 1–7.
[3] X. Yang, Y. Li, and S. Lyu, “Exposing deep fakes using inconsistent head poses,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2019, pp. 8261–8265.
[4] F. Matern, C. Riess, and M. Stamminger, “Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations,” in Proc. IEEE Winter Appl. Comput. Vis. Workshops, 2019, pp. 83–92.
[5] D. Afchar, V. Nozik, J. Yamagishi, and I. Ishizaki, “MesoNet: A compact facial video forgery detection network,” in Proc. IEEE Int. Workshop Inf. Forensics Secur., 2018, pp. 1–7.
[6] L. M. Dang et al., “Deep learning based computer generated face identification using convolutional neural network,” Appl. Sci., vol. 8, no. 12, 2018, Art. no. 2610.
[7] H. Mo, B. Chen, and W. Luo, “Fake faces identification via convolutional neural network,” in Proc. 6th ACM Workshop Inf. Hiding Multimedia Secur., 2018, pp. 43–47.
[8] Z. Guo, G. Yang, J. Chen, and X. Sun, “Fake face detection via adaptive manipulation traces extraction network,” Comput. Vis. Image Understanding, vol. 204, 2021, Art. no. 103170.
[9] X. Zhao, Y. Lin, and J. Heikinl, “Dynamic texture recognition using volume local binary count patterns with an application to 2D face spoofing detection,” IEEE Trans. Multimedia, vol. 20, no. 3, pp. 552–566, Mar. 2018.
[10] X. Peng, I. J. Cox, and G. Doerr, “Normalized energy density-based forensics detection of resampled images,” IEEE Trans. Multimedia, vol. 14, no. 3, pp. 536–545, Jun. 2012.
[11] A. C. Popescu and H. Farid, “Exposing digital forgeries by detecting traces of resampling,” IEEE Trans. Signal Process., vol. 53, no. 2, pp. 758–767, Feb. 2005.
[12] X. Qiu, H. Li, W. Luo, and J. Huang, “A universal image forensic strategy based on steganalytic model,” in Proc. 2nd ACM Workshop Inf. Hiding Multimedia Secur., 2014, pp. 165–170.

[13] M. Kirchner, “Fast and reliable resampling detection by spectral analysis of padded linear prediction residual,” in Proc. 10th ACM Workshop Multimedia Secur., 2008, pp. 11–20.

[14] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, “Do GANs leave artificial fingerprints?,” in Proc. IEEE Conf. Multimedia Inf. Process. Retrieval, 2019, pp. 506–511.

[15] N. Yu, L. S. Davis, and M. Fritz, “Attributing fake images to GANs: Learning and analyzing GAN fingerprints,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 7556–7566.

[16] K. He, J. Sun, and X. Tang, “Guided Image Filtering,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[17] A. Rossler et al., “FaceForensics++: Learning to detect manipulated facial images,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 1–11.

[18] L. Liang and X. Zhang, “Adaptive label propagation for facial appearance transfer,” IEEE Trans. Multimedia, vol. 21, no. 12, pp. 3068–3082, Dec. 2019.

[19] D. P. Kingma and P. Dhariwal, “Glow: Generative flow with invertible 1x1 convolutions,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 10215–10224.

[20] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, “Gamification: Analogically-aware facial animation from a single image,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 818–833.

[21] D. Berthelot, T. Schumm, and L. Metz, “Began: Boundary equilibrium generative adversarial networks,” 2017, arXiv:1703.10177.

[22] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” in Proc. Int. Conf. Learn. Representations, 2018, pp. 1–26.

[23] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4401–4410.

[24] Y. Choi et al., “StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 7879–7897.

[25] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2Face: Real-time face capture and reenactment of RGB videos,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2387–2395.

[26] J. Thies, M. Zollhofer, and M. Nießner, “Deferred neural rendering: Image synthesis using neural textures,” ACM Trans. Graph., vol. 38, no. 4, pp. 1–12, 2019.

[27] I. Masi, A. Killekar, R. M. Sacreanal, S. P. Gurudatt, and W. AbdAlmageed, “Two-branch recurrent network for isolating deepfakes in videos,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 667–684.

[28] H. Zhao et al., “Multi-attentional deepfake detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 2185–2194.

[29] S. Das et al., “Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 3776–3785.

[30] T. Zhao et al., “Learning self-consistency for deepfake detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 15023–15033.

[31] Y. Li and S. Lyu, “Exposing deepfake videos by detecting face warping artifacts,” in Proc. CVPR Workshops, 2019, pp. 46–52.

[32] L. M. Dang, S. I. Hassan, S. Im, and H. Moon, “Face image manipulation detection based on a convolutional neural network,” Expert Syst. Appl., vol. 129, pp. 156–168, 2019.

[33] Y. Jeong et al., “BiHPF: Bilateral high-pass filters for robust deepfake detection,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis., 2022, pp. 1–7.

[34] X. Yang, Y. Li, H. Qi, and S. Lyu, “Exposing GAN-synthesized faces using landmark markers,” in Proc. ACM Workshop Inf. Hiding Multimedia Secur., 2019, pp. 113–118.

[35] H. Li, B. Li, S. Tan, and J. Huang, “Identification of deep network generated images using disparities in color components,” Signal Process., vol. 174, 2020, Art. no. 107616.

[36] Z. Zhao, P. Sun, and W. Lu, “Detecting deepfake video by learning two-level features with two-stream convolutional neural network,” in Proc. 6th Int. Conf. Comput. Artif. Intell., 2020, pp. 291–297.

[37] P. Zhou, X. Han, V. I. Morarui, and L. S. Davis, “Learning rich features for image manipulation detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1053–1061.

[38] B. Bayar and M. C. Stamm, “Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection,” IEEE Trans. Inf. Forensics Secur., vol. 13, no. 11, pp. 2691–2706, Nov. 2018.