Cross-Domain Local Characteristic Enhanced Deepfake Video Detection

Zihan Liu, Hanyi Wang, and Shilin Wang*

School of Electronic Information and Electrical Engineering,
Shanghai Jiao Tong University
{lzh123,why_820,wsl}@sjtu.edu.cn

Abstract. As ultra-realistic face forgery techniques emerge, deepfake detection has attracted increasing attention due to security concerns. Many detectors cannot achieve accurate results when detecting unseen manipulations despite excellent performance on known forgeries. In this paper, we are motivated by the observation that the discrepancies between real and fake videos are extremely subtle and localized, and inconsistencies or irregularities can exist in some critical facial regions across various information domains. To this end, we propose a novel pipeline, Cross-Domain Local Forensics (XDLF), for more general deepfake video detection. In the proposed pipeline, a specialized framework is presented to simultaneously exploit local forgery patterns from space, frequency, and time domains, thus learning cross-domain features to detect forgeries. Moreover, the framework leverages four high-level forgery-sensitive local regions of a human face to guide the model to enhance subtle artifacts and localize potential anomalies. Extensive experiments on several benchmark datasets demonstrate the impressive performance of our method, and we achieve superiority over several state-of-the-art methods on cross-dataset generalization. We also examined the factors that contribute to its performance through ablations, which suggests that exploiting cross-domain local characteristics is a noteworthy direction for developing more general deepfake detectors.

1 Introduction

Recent years have witnessed tremendous progress in face forgery techniques [13, 25, 29, 40], i.e., deepfake, due to the emergence of deep generative models. As such techniques can synthesize highly realistic fake videos without considerable human effort, they can easily be abused by malicious attackers to counterfeit imperceptible identities or behaviors, thereby causing severe political and social threats. To mitigate such threats, numerous automatic deepfake detection methods [6, 9, 20, 31, 34, 46, 53, 54] have been proposed.

Most studies formulated deepfake detection as a binary classification problem with global supervision (i.e., real/fake) for training. They relied on convolutional neural networks (CNN) to extract discriminative features to detect forgeries.

* Corresponding Author

While these methods achieved satisfactory accuracy when the training and test sets have similar distributions, their performance significantly dropped when encountering novel manipulations. Therefore, many works [20, 26, 34, 36] aimed at improving generalization to unseen forgeries with diverse approaches.

With the continuous refinement of face forgery methods, the discrepancies between real and fake videos are increasingly subtle and localized. Inconsistencies or irregularities can exist in some critical local regions across various information domains, e.g., space [1, 43], frequency [6, 31, 34, 42], and time [4, 18, 24, 44] domains. However, these anomalies are so fine-grained that vanilla CNN often fails to capture them. Many detection algorithms exploited local characteristics to enhance generalization performance. However, these algorithms still had some limitations in representing local features. On the one hand, some algorithms [20] solely relied on a specific facial region to distinguish between real and fake videos while ignoring other facial regions, which restricted the detection performance. On the other hand, many algorithms [6, 34, 36] made insufficient use of local representation and cannot aggregate local information from various domains.

In this work, we are motivated by the above observation. It is reasonable to assume that incorporating more local regions and information domains can improve detection performance. We expect to design a specialized model to implement this idea and verify its performance through extensive experiments. We aim to guide the model to capture subtle artifacts around some high-level facial local regions that are sensitive to forgeries due to complicated natural motions. These regions are referred to as the forgery-sensitive local regions (FSLR) in this paper, which are abundant in high-level semantics that can enhance the model's generalization capability. We also consider the feasibility of simultaneously exploiting information from space, frequency, and time domains based on a 3D CNN backbone.

To this end, we propose Cross-Domain Local Forensics (XDLF), a novel pipeline specially designed for feature extraction across multiple domains and local artifacts enhancement. Four forgery-sensitive local regions (i.e., left eye, right eye, nose, and mouth) are extracted to guide the model to capture subtle artifacts around these regions. To simultaneously leverage information from space, frequency, and time domains, we design a two-stream 3D CNN based framework to learn a cross-domain dense representation for forgery detection.

To demonstrate the effectiveness of our framework, extensive experiments were conducted on several benchmark datasets, including FaceForensics++ [43], Celeb-DF [29], and DFDC [13]. Our results show the superiority of the proposed method over many state-of-the-art approaches on cross-dataset generalization.

Our main contributions are as follows:

- We leverage four forgery-sensitive local regions of a human face to guide the model to enhance subtle artifacts and localize potential anomalies around those regions. Using bounding boxes of those regions, we extract regional features as an attention to help the model focus more on those regions. We validated our design through ablations.

- We present a novel deepfake video detection pipeline that simultaneously exploits information from space, frequency, and time domains, thus learning a cross-domain dense representation for better generalization.
- We achieve impressive performance on extensive experiments, and our method outperforms several state-of-the-art methods on cross-dataset generalization.

2 Related Work

2.1 Deepfake Detection

Existing deepfake detection algorithms can fall into two categories, namely image-based methods and video-based methods, depending on whether temporal information is explicitly exploited across frames.

Image-based Methods. Earlier image-based methods employed hand-crafted facial features to detect forgeries, e.g., steganalysis features [55], inconsistent head poses [50], and anomalous visual artifacts [37]. However, these methods underperformed on more realistic forgeries synthesized with more advanced face manipulation technologies recently. With the tremendous progress of deep learning, many works [1, 43] utilized state-of-the-art convolutional neural networks (CNN), e.g., Xception [7], to extract features from facial images and perform binary classification. More recently, an increasing number of CNN-based methods have been proposed from various perspectives. They aimed at exploring the crucial discrepancies between real and fake images, continuously improving the detection performance. These methods included leveraging frequency spectrum [16, 31, 34, 36, 42], attention mechanism [10, 53], extra identity information [3, 9], self-supervised learning [26, 28, 54], etc.

Video-based Methods. Unlike image-based methods, video-based methods distinguish real and fake videos based on a sequence of aligned frames. Most works managed to model the temporal consistency across frames, since current face manipulation techniques struggled to generate temporally coherent fake videos. These methods [19, 30, 36, 44] utilized recurrent neural networks (RNN) or 3D CNN to extract spatio-temporal features of facial movements. They can focus on unnatural eye blinking [27], irregular mouth motion [20], inconsistent visual-auditory modalities [2, 8, 38, 56]. In contrast, our method designs a two-stream 3D CNN based framework to mine forgery patterns from space, frequency, and time domains. We also leverage four facial forgery-sensitive local regions to enhance imperceptible artifacts for forgery defect localization.

2.2 Generalization to Unseen Forgeries

While current methods achieved excellent accuracy in the scenario where the training and test sets have similar distributions, they cannot generalize very well to unseen forgeries and tend to overfit to manipulation-specific artifacts. It is of paramount importance for deployed detectors to learn generalized representation regardless of forgery types. To this end, many works focused on improving

generalization to unseen forgeries with diverse approaches. Several works [34,36] used a two-branch architecture to exploit information from the RGB domain and the frequency domain, exploiting generalized frequency patterns to expose the discrepancies. Our method has a similar idea but far different designs. Moreover, a series of self-supervised methods [26,28,54] demonstrated superior generalization. These methods relied on self-generated fake data targeted at specific patterns without the need for conventional forgery training data. The patterns can be face warping artifacts [28], blending boundary [26], source feature inconsistency [54]. LipForensics [20] exhibited remarkable performance on cross-dataset generalization by pre-training a spatio-temporal network to perform lipreading and fine-tuning on a deepfake dataset. We followed its experimental settings due to similar goals.

3 Proposed Method

3.1 Overview

In this section, we first explain the motivation of our work, and then briefly introduce the pipeline of our proposed method.

Motivation. Recent studies [20, 37, 42, 53] have shown that the discrepancies between real and fake videos contain implicitly in local subtle regions, where manipulation artifacts may exist across various information domains. Unfortunately, most deepfake datasets have no manipulation masks as local supervision. Without external location guidance of facial semantic regions that are sensitive to forgeries, it is often difficult for detectors to localize those subtle artifacts. We observe that current detection algorithms had two limitations in leveraging local information:

- Some algorithms [20, 27] relied on a single facial region as the criterion to detect forgeries, while ignoring the effect of other critical local regions, which may restrict the performance. Our framework exploits four forgery-sensitive local regions (FSLR) of a human face, which are used to guide the model to enhance subtle artifacts and localize more potential anomalies based on our newly proposed FSLR-Guided Feature Enhancement.
- Many algorithms made insufficient use of local regions to detect anomalies, which can be embodied in multiple information domains, e.g., space, frequency, and time domains. To the best of our knowledge, few studies have been done to simultaneously capture local features across these three domains. We note that the Two-branch [36] method extracted spatial/frequency and temporal features at two stages with CNN and RNN, respectively, without cross reference among these features. To this end, we propose a two-stream framework, Cross-Domain Local Forensics, to simultaneously exploit local information from those three domains.

Pipeline. Motivated by the above observations, we propose a novel feature extraction framework **Cross-Domain Local Forensics** (**XDLF**) for more general deepfake video detection. Fig. 1 illustrates the overall pipeline of XDLF.

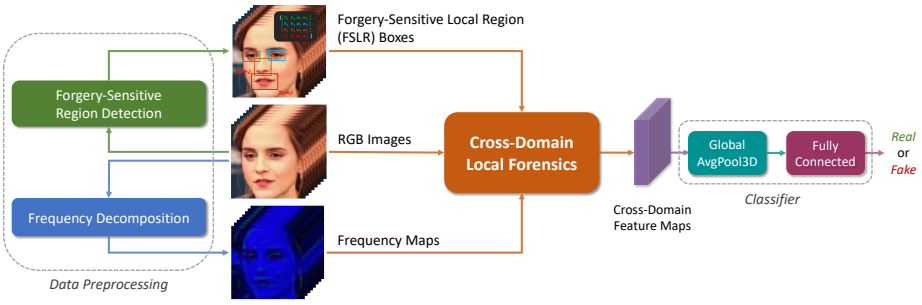


Fig. 1. Pipeline of our proposed framework XDLF. The end-to-end training consists of three stages: Data Preprocessing, Cross-Domain Local Forensics, and Classifier.

The pipeline takes as input a sequence of aligned RGB frames. First, the data preprocessing consists of two procedures. On the one hand, **Frequency Decomposition** takes as input RGB images to generate frequency maps where manipulation traces in the frequency domain are amplified, especially for those videos with high compression. On the other hand, **Forgery-Sensitive Region Detection** takes as input RGB images to extract bounding boxes of four **forgery-sensitive local regions** (**FSLR**) that are abundant in high-level defects. The four FSLRs are left eye, right eye, nose, and mouth. Then, sequences of RGB images, frequency maps, and FSLR boxes are input into **Cross-Domain Local Forensics** (**XDLF**) to learn a comprehensive and generalized cross-domain features. Finally, a classifier comprising a 3D global average pooling layer and a fully-connected layer is used to make predictions.

3.2 Data Preprocessing

Frequency Decomposition. Recent studies [31,52] observed that up-sampling is a necessary step of most existing face manipulation methods, and cumula-

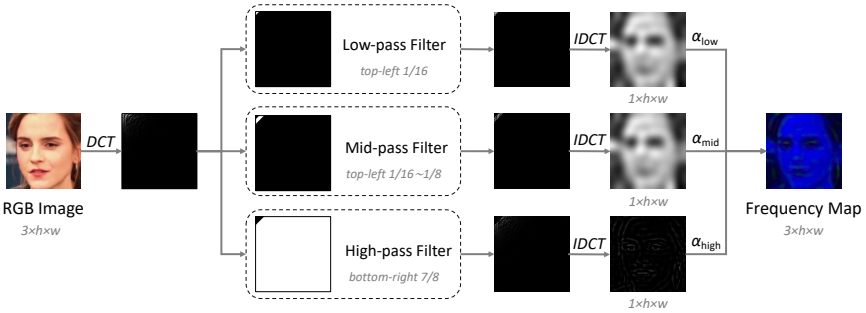


Fig. 2. Pipeline of Frequency Decomposition. This module generates frequency maps where manipulation traces in the frequency domain are amplified adaptively.

tive up-sampling can leave apparent anomalies in the frequency domain, which provides clues for detecting manipulated videos. Inspired by F^3-Net [42], we design Frequency Decomposition to obtain multi-band frequency maps adaptively. Fig. 2 shows the pipeline of this module.

For each RGB image $\boldsymbol{X}$ in a frame sequence, we first calculate the frequency response with Discrete Cosine Transform (DCT) $\mathcal{D}$. Then, filters of low, middle, and high frequency bands $\boldsymbol{f_i},\ i \in \{\text{low}, \text{mid}, \text{high}\}$ are used to obtain three frequency components. We follow the settings in [42] to construct filters. Next, Inversed Discrete Cosine Transform (IDCT) $\mathcal{D}^{-1}$ is applied to three frequency components to obtain the corresponding spatial components $\boldsymbol{Y_i},\ i \in \{\text{low}, \text{mid}, \text{high}\}$. Finally, the three spatial components are concatenated to attain the frequency map $\boldsymbol{Y}$. Before concatenation, each component is multiplied by a learnable weight $\alpha_i \in (0, 1),\ i \in \{\text{low}, \text{mid}, \text{high}\}$ to enable the model to adaptively concentrate on the interested frequency band for a flexible representation of frequency features. The above can be summarized as Eq. 1, 2, where $\odot$ is the element-wise product.

$$\boldsymbol{Y_i} = \mathcal{D}^{-1}\{\mathcal{D}(\boldsymbol{X}) \odot \boldsymbol{f_i}\},\ i \in \{\text{low}, \text{mid}, \text{high}\} \tag{1}$$

$$\boldsymbol{Y} = \text{Concat}(\alpha_{\text{low}}\boldsymbol{Y}_{\text{low}},\ \alpha_{\text{mid}}\boldsymbol{Y}_{\text{mid}},\ \alpha_{\text{high}}\boldsymbol{Y}_{\text{high}}) \tag{2}$$

Forgery-Sensitive Region Detection. Current face manipulation techniques struggled to generate temporally coherent fake faces, especially in high-level semantic regions that have continual motions and thereby sensitive to forgeries. We hope to guide the model to pay more attention to these regions. Therefore, we extract bounding boxes of four forgery-sensitive local regions (FSLR): left eye, right eye, nose, and mouth. These four manually selected regions are further leveraged by **FSLR-Guided Feature Enhancement** (**FGFE**) as an external guidance. For each RGB image, we first compute 68 facial landmarks based on a face detector. Then, the landmarks are used to crop bounding boxes of those four regions based on preset box sizes. Each box can be expressed as a quadruple (h_1, h_2, w_1, w_2) where (h_1, w_1) is the top-left vertex and (h_2, w_2) is the bottom-right vertex. The four boxes are stacked to generate the 4×4 FSLR box matrix.

3.3 Cross-Domain Local Forensics

We propose a novel two-stream collaborative learning framework for cross-domain feature extraction, Cross-Domain Local Forensics (XDLF), which is based on a spatio-temporal convolutional backbone. As is illustrated in Fig. 3, the framework consists of two symmetric 3D CNN backbones: 3D-CNN(A) extracts spatio-temporal features of RGB images, and 3D-CNN(B) extracts frequency-temporal features of frequency maps. The features of the two modalities are cross-referenced and merged at low, middle, and high levels of the backbone, with **Cross Attention** and **Feature Fusion**, respectively. Moreover, we apply **FSLR-Guided Feature Enhancement** to the low-level features of both streams, thus enhancing the local subtle artifacts of shallow features under the guidance of forgery-sensitive regions. The ultimate cross-domain features are obtained with **Feature Ensemble** to integrate features of three different levels of abstraction.

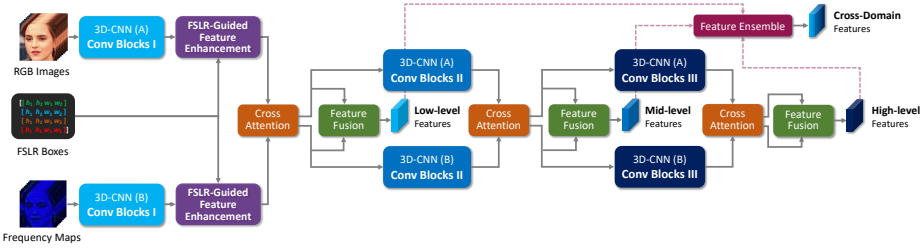


Fig. 3. Framework of Cross-Domain Local Forensics. We adopt a two-stream architecture for cross-domain feature extraction based on two symmetric spatio-temporal convolutional backbones, e.g., 3D ResNet-50 [21, 22].

FSLR-Guided Feature Enhancement. Many studies [31, 53] showed that local textural artifacts represent the high frequency component of shallow features, which is essential for the face forgery detection task. These artifacts are especially salient nearby critical facial regions that are sensitive to forgeries. As aforementioned, we exploit four forgery-sensitive local regions to enhance subtle artifacts and guide the model to localize more possible anomalies in these regions. The module structure is shown in Fig. 4.

The module takes as input low-level RGB (or frequency) features $\boldsymbol{X} \in \mathbb{R}^{c\times d\times h\times w}$ (of c channels, depth d, height h, width w) and FSLR boxes $\boldsymbol{r} \in \mathbb{Z}^{d\times 4\times 4}$ and outputs the enhanced features of the same shape. First, the region coordinates are scaled down (i.e., region projection) according to the size difference between the RGB image (or frequency map) and low-level features. Then, FSLR features $\boldsymbol{R} \in \mathbb{R}^{4\times c\times d\times H\times W}$ are obtained with region pooling, which refers to ROI pooling [17] in object detection. Specifically, we crop four sub-features with region coordinates and generate four FSLR features of fixed size $(H \times W)$ using adaptive max-pooling (Eq. 3). FSLR features condense the irregular semantics of local textural patterns in these four regions, which serve as an attention for global features. Next, transformed features $\boldsymbol{X'} \in \mathbb{R}^{c\times hw}$ are calculated

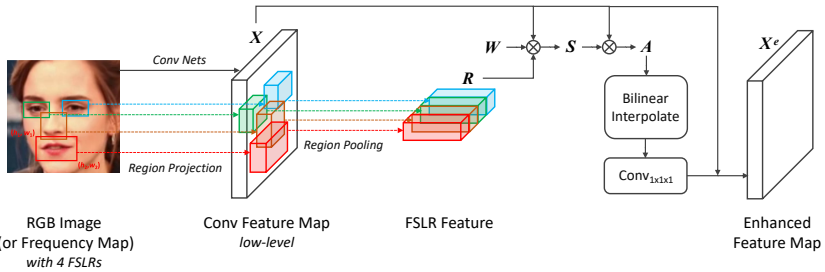


Fig. 4. Structure of FSLR-Guided Feature Enhancement. This module is designed to guide the model to enhance subtle artifacts of shallow features and localize more anomalous regions.

by temporally averaging the RGB (or frequency) features $\boldsymbol{X}$ and flattening spatial dimensions. And features $\boldsymbol{R'} \in \mathbb{R}^{4c \times dHW}$ are also obtained by flattening the FSLR features $\boldsymbol{R}$. Later, the similarity matrix $\boldsymbol{S} \in \mathbb{R}^{hw \times dHW}$ between $\boldsymbol{X'}$ and $\boldsymbol{R'}$ (Eq. 4) is computed, where $\boldsymbol{W} \in \mathbb{R}^{c \times 4c}$ is a learnable weight matrix. Each value in $\boldsymbol{S}$ represents the similarity between each row in $\boldsymbol{X'}^T$ and each column in $\boldsymbol{R'}$. By the similarity matrix, we model the internal relevance between those local regions for cross-region forgery mining. And then the attention matrix $\boldsymbol{A} \in \mathbb{R}^{c \times dHW}$ is calculated to enhance the original features (Eq. 5). Moreover, the upsampled $\boldsymbol{A'} \in \mathbb{R}^{c \times d \times h \times w}$ is obtained by reshaping, bilinear interpolation, and $1 \times 1 \times 1$ convolution (Eq. 6, 7). Finally, the enhanced features $\boldsymbol{X^e} \in \mathbb{R}^{c \times d \times h \times w}$ are attained by element-wise product and residual addition (Eq. 8). We apply this module to the low-level features of both streams, which enables the model to pay more attention to the regularity and consistency of local semantic regions.

$$\boldsymbol{R} = \text{AdaMaxPool}(\boldsymbol{X}, \text{Proj}(\boldsymbol{r})) \quad (3)$$

$$\boldsymbol{S} = \boldsymbol{X'}^T \boldsymbol{W} \boldsymbol{R'} \quad (4)$$

$$\boldsymbol{A} = \boldsymbol{X'} \boldsymbol{S} \quad (5)$$

$$\boldsymbol{A'} = \text{BilinearInterpolate}(\boldsymbol{A}) \quad (6)$$

$$\boldsymbol{A'} = \text{ReLU}(\text{BN}(\text{Conv}_1(\boldsymbol{A'}))) \quad (7)$$

$$\boldsymbol{X^e} = \boldsymbol{X} + \boldsymbol{X} \odot \boldsymbol{A'} \quad (8)$$

Cross Attention. In this module, RGB and frequency features are cross-referenced at low, middle, and high levels of the backbone, which enables the model to learn a more comprehensive cross-domain representation. The module takes as input RGB features $\boldsymbol{X}$ and frequency features $\boldsymbol{X_f}$. First, the two features are concatenated on the channel axis and then applied $1 \times 1 \times 1$ convolution (Eq. 9, 10). Next, $3 \times 3 \times 3$ convolution with output channel 2 and sigmoid activation is used to obtain two attention maps (Eq. 11). Finally, the original features are enhanced with attention maps by element-wise product (Eq. 12).

$$\boldsymbol{U} = \text{Concat}(\boldsymbol{X}, \boldsymbol{X_f}) \quad (9)$$

$$\boldsymbol{U'} = \text{ReLU}(\text{BN}(\text{Conv}_1(\boldsymbol{U}))) \quad (10)$$

$$\boldsymbol{A}, \boldsymbol{A_f} = \text{Sigmoid}(\text{Conv}_3(\boldsymbol{U'})) \quad (11)$$

$$\boldsymbol{X^c} = \boldsymbol{X} \odot \boldsymbol{A}, \ \boldsymbol{X_f^c} = \boldsymbol{X_f} \odot \boldsymbol{A_f} \quad (12)$$

Feature Fusion. In this module, RGB and frequency features are fused in a complementary way based on Squeeze-and-Excitation (SE) [23]. SE block improves the quality of cross-domain features by explicitly modeling the interdependence between the channels of RGB and frequency features. The module structure is shown in Fig. 5.

This module also takes as input RGB features $\boldsymbol{X} \in \mathbb{R}^{C \times D \times H \times W}$ and frequency features $\boldsymbol{X_f} \in \mathbb{R}^{C \times D \times H \times W}$. The two features are first concatenated to obtain $\boldsymbol{U} \in \mathbb{R}^{2C \times D \times H \times W}$ (Eq. 13). Then, the spatial information is squeezed into a value by global pooling to get channel descriptor $\boldsymbol{V} \in \mathbb{R}^{2C}$ (Eq. 14, 15). Next, we enable channel descriptor $\boldsymbol{V}$ to capture the interdependency between channels and obtain channel attention $\boldsymbol{A_c} \in \mathbb{R}^{2C}$ (Eq. 16, 17, 18), where $\boldsymbol{W_1} \in \mathbb{R}^{2C \times \frac{2C}{r}}$ and $\boldsymbol{W_2} \in \mathbb{R}^{\frac{2C}{r} \times 2C}$ are learnable weight matrices, r is the reduction ratio. Finally, the fused features $\boldsymbol{X^v}$ are computed as Eq. 19.

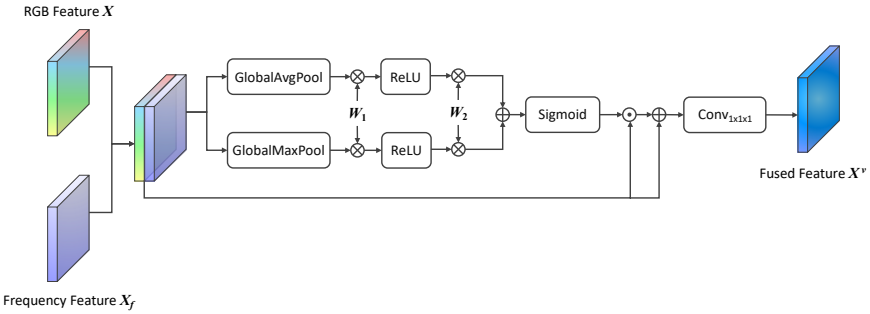


Fig. 5. Structure of Feature Fusion. This module is designed to model the interdependence between RGB and frequency features for improved cross-domain fusion.

$$\boldsymbol{U} = \text{Concat}(\boldsymbol{X}, \boldsymbol{X_f}) \quad (13)$$

$$\boldsymbol{V}_{\text{avg}} = \text{GlobalAvgPool}(\boldsymbol{U}) \quad (14)$$

$$\boldsymbol{V}_{\text{max}} = \text{GlobalMaxPool}(\boldsymbol{U}) \quad (15)$$

$$\boldsymbol{V}'_{\text{avg}} = \boldsymbol{W_2}\text{ReLU}(\boldsymbol{W_1}\boldsymbol{V}_{\text{avg}}) \quad (16)$$

$$\boldsymbol{V}'_{\text{max}} = \boldsymbol{W_2}\text{ReLU}(\boldsymbol{W_1}\boldsymbol{V}_{\text{max}}) \quad (17)$$

$$\boldsymbol{A_c} = \text{Sigmoid}(\boldsymbol{V}'_{\text{avg}} + \boldsymbol{V}'_{\text{max}}) \quad (18)$$

$$\boldsymbol{X^v} = \text{ReLU}(\text{BN}(\text{Conv}_{1\times1\times1}(\boldsymbol{U} + \boldsymbol{U} \odot \boldsymbol{A_c}))) \quad (19)$$

Feature Ensemble. This module aggregates low, middle, and high-level features through adaptive average pooling and concatenation (Eq. 20, 21, 22, 23).

$$\widetilde{\boldsymbol{X}}^{\text{low}} = \lambda_{\text{low}}\text{AdaAvgPool}(\boldsymbol{X}^{\text{low}}, (d_{\text{high}}, h_{\text{high}}, w_{\text{high}})) \quad (20)$$

$$\widetilde{\boldsymbol{X}}^{\text{mid}} = \lambda_{\text{mid}}\text{AdaAvgPool}(\boldsymbol{X}^{\text{mid}}, (d_{\text{high}}, h_{\text{high}}, w_{\text{high}})) \quad (21)$$

$$\widetilde{\boldsymbol{X}}^{\text{high}} = \lambda_{\text{high}}\boldsymbol{X}^{\text{high}} \quad (22)$$

$$\widetilde{\boldsymbol{X}} = \text{Concat}(\widetilde{\boldsymbol{X}}^{\text{low}}, \widetilde{\boldsymbol{X}}^{\text{mid}}, \widetilde{\boldsymbol{X}}^{\text{high}}) \quad (23)$$

where $\boldsymbol{X}^i \in \mathbb{R}^{c_i \times d_i \times h_i \times w_i}$, $i \in \{\text{low}, \text{mid}, \text{high}\}$ are fused features of three abstraction levels, and $\lambda_{\text{i}} \in (0, 1)$, $i \in \{\text{low}, \text{mid}, \text{high}\}$ are three learnable parameters for adaptive feature combination.

4 Experiment and Discussion

4.1 Experiment Setup

Datasets. We used **FaceForensics**++ (FF++) [43] for training and validation, and evaluated the cross-dataset generalization on **Celeb-DF** (CDF) [29] and **DeepFake Detection Challenge** (DFDC) [13]. (1) FF++ is the most commonly used benchmark dataset containing 1,000 real videos and 4,000 fake

videos. Each real video is manipulated by four face forgery techniques, i.e., DeepFakes (DF) [11], FaceSwap (FS) [15], Face2Face (F2F) [48], and NeuralTextures (NT) [47]. We adopted the slightly-compressed (HQ/c23) and heavily-compressed (LQ/c40) versions for our experiments. (2) CDF is a challenging dataset that includes 590 real videos and 5,639 fake videos synthesized by an improved algorithm. (3) DFDC is a large-scale dataset with extreme filming conditions and various perturbations, which is also very challenging for current deepfake detectors. We used the preview version [14] that includes 1,131 real videos and 4,113 fake videos for our evaluation.

Evaluation Metrics. Following most existing works [20,26,36], Accuracy (ACC) and Area Under the Receiver Operating Characteristic Curve (AUC) were used as the metrics to evaluate our method. As in [20], we reported video-level metrics for fair comparison with image-based methods. Specifically, all frame/clip predictions were averaged across the video and hence all models predicted based on an equal number of frames.

Implementation Details. For each video, we sampled non-overlapping frame clips with a length of 16, and oversampled the minority class (e.g., real in FF++) to tackle label imbalance. We used the state-of-the-art face detector RetinaFace [12] to crop facial images with a size of 224×224 and FSLR box matrices with a size of 4×4. The preset FSLR size is 40×40 for the mouth and 30×30 for the other three. For data augmentations, we applied several traditional image augmentations such as random horizontal flipping. Moreover, as in [41], we conducted Mixup [51] augmentation on aligned real-fake pairs to reduce overfitting. For XDLF, we adopted 3D ResNet-50 [21, 22] as the backbone which is pre-trained on large-scale action recognition datasets to accelerate the model convergence. For FSLR-Guided feature enhancement, we set FSLR feature size $H = W = 7$. For feature fusion, we set reduction ratio $r = 16$. For training, we used a batch size of 4 and AdamW [33] optimizer with initial learning rate 1×10^{-4} and weight decay 1×10^{-4}. The learning rate decayed with a cosine annealing [32] strategy with $T_{\max} = 32$.

4.2 In-dataset Evaluation

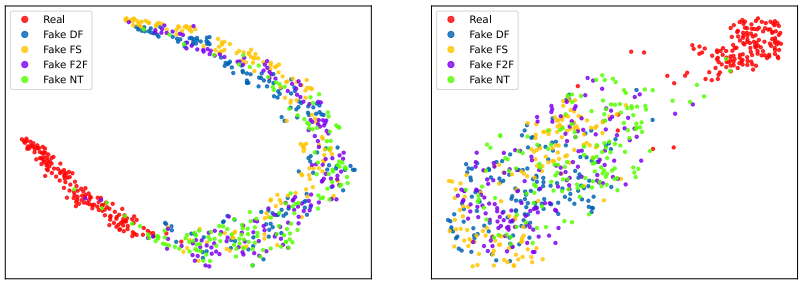
We evaluated our method in the in-dataset scenario where the training and test sets have identical distributions. Following [20], we compared our method with current state-of-the-art approaches in FF++ under different quality settings (HQ/LQ). As shown in Table 1, we achieve great improvements over most current methods, especially under the challenging low-quality (LQ) setting where frequency statistics are partly destroyed. However, our method still maintains good performance when exploiting frequency spectrum, which we attribute to our two-stream architecture that learns to be biased towards RGB features. Note that we gain comparable results with LipForensics [20], which leverages dynamic lip features from pre-trained lipreading models. Unlike LipForensics, our method does not need any external pre-training data and can be more efficiently trained.

Moreover, we show the t-SNE [35] visualization of features extracted from classifiers of LipForensics and our method on FF++ high-quality (HQ) test set

Table 1. In-dataset performance comparisons. We report video-level ACC/AUC (%) when trained and tested on FF++ slightly-compressed (HQ) and heavily-compressed (LQ) videos. The results of other methods are quoted from [20]. The best results are in **bold**, and the second-best results are underlined.

Method	FF++(HQ)		FF++(LQ)	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)
Xception [43]	97.0	99.3	89.0	92.0
CNN-aug [49]	96.9	99.1	81.9	86.9
Patch-based [5]	92.6	97.2	79.1	78.3
Two-branch [36]	–	99.1	–	91.1
Face X-ray [26]	78.4	97.8	34.2	77.3
CNN-GRU [44]	97.0	99.3	90.1	92.2
LipForensics [20]	**98.8**	**99.7**	94.2	**98.1**
XDLF (ours)	98.1	**99.7**	**94.5**	96.7

in Fig. 6. We observe that although both methods can well distinguish real and fake data, they learn different feature distributions. For LipForensics, the separation distances between real and fake data are smaller than our method, which can easily lead to classification ambiguity in those in-between videos, especially for some real and NeuralTextures-based fake samples. On the other hand, our method learns a more mixed and gathered feature representation of FF++ fake data without obviously separating different forgery types. It proves that our method can learn a generalized feature to detect novel forgeries.



(a) LipForensics (b) XDLF

Fig. 6. The t-SNE feature visualization of the baseline LipForensics [20] (a) and our proposed XDLF (b) on FF++(HQ) test set. Each dot represents the feature of a video clip. Red dots are real clips, while the rest are fake ones with different forgery types.

4.3 Cross-dataset Evaluation

In real-world scenarios, a deployed detector is expected to identify videos crafted by unseen manipulations with unknown source videos. Therefore, we conducted

Table 2. Cross-dataset performance comparisons. We report video-level AUC (%) on CDF and DFDC when trained on FF++(HQ). The results of other methods are quoted from [20]. The best results are in **bold**.

Method	CDF AUC (%)	DFDC AUC (%)
Xception [43]	73.7	70.9
CNN-aug [49]	75.6	72.1
Patch-based [5]	69.6	65.6
Face X-ray [26]	79.5	65.5
CNN-GRU [44]	69.8	68.9
Multi-task [39]	75.7	68.1
DSP-FWA [28]	69.5	67.3
Two-branch [36]	76.7	–
LipForensics [20]	82.4	73.5
XDLF (ours)	**82.6**	**73.8**

cross-dataset evaluation as in [20] to verify the generalization capability of our method. Specifically, we trained the models on FF++(HQ) and tested them on CDF and DFDC. As shown in Table 2, our method outperforms all listed methods on both unseen datasets, surpassing the recent state-of-the-art LipForensics [20] by 0.2% and 0.3% in terms of AUC on CDF and DFDC, respectively.

4.4 Ablation Study

Evaluations on Core Modules in XDLF. To understand the components responsible for our method's performance, we ablated three core modules in XDLF and examined its in-dataset and cross-dataset generalization performance. The modules are FSLR-Guided Feature Enhancement (**FGFE**), Cross Attention, and Feature Fusion. For the first two, we removed them directly as their inputs and outputs have the same shapes. For Feature Fusion, we replaced it with a simple channel-axis concatenation of RGB and frequency features. We trained all the models on FF++(HQ) and tested them on FF++(HQ), CDF, and DFDC.

The results are shown in Table 3. We have the following observations: (1) Training our model without FGFE leads to a performance drop on all datasets.

Table 3. Evaluations on core modules in XDLF. We report video-level ACC/AUC (%) on FF++(HQ), CDF, and DFDC when trained on FF++(HQ). The highlighted row is our original setting. We ablated core modules in our feature extraction framework to verify their effects. The best results are in **bold**.

Method	FF++(HQ)		CDF		DFDC	
	ACC	AUC	ACC	AUC	ACC	AUC
XDLF (ours)	**98.1**	**99.7**	**74.2**	**82.6**	**66.2**	**73.8**
w/o FGFE	97.9	99.3	71.7	79.2	65.3	69.8
w/o Cross Attention	97.8	99.4	72.2	79.9	65.9	71.0
w/o Feature Fusion	98.0	99.4	73.8	81.5	66.1	72.3

In cross-dataset evaluation, the model decreases by 3.4% and 4.0% in terms of AUC on CDF and DFDC, respectively. This suggests that the model learns more generalized features by enhancing subtle artifacts in those forgery-sensitive local regions. (2) Both Cross Attention and Feature Fusion play an essential role in performance improvements. Although they have the same goal to complementarily exploit forgery patterns from different domains, they work differently and enhance the model's performance mutually.

To further understand the effect of FGFE, we show the Grad-CAM [45] visualization of the model without/with FGFE in Fig. 7. It visually explains that FGFE serves as external guidance to help the model focus on four forgery-sensitive local regions. As can be seen, these regions are abundant in motions that contain more subtle artifacts. The model can localize more potential anomalies to detect forgeries with the help of FGFE, which is consistent with our motivation.

Evaluations on Different Information Domains. We altered our feature extraction framework XDLF to prove the necessity to mine forgery clues from three different information domains, i.e., space, frequency, and time domains. Specifically, we trained three variants with each dropping one of the three domains: (1) **Freq-Freq-3D**: The inputs of both streams are the same frequency maps, and the network structure is unchanged. (2) **RGB-RGB-3D**: The inputs of both streams are RGB images, and the network structure is unchanged. (3) **RGB-Freq-2D**: The inputs are still RGB images and frequency maps, but temporal dimension is merged into batch dimension. We replaced the 3D ResNet-50 backbone with 2D ResNet-50 backbone, and replaced all 3D convolutional layers and 3D batch normalization layers with 2D counterparts.

The results are shown in Table 4. We have the following observations: (1) By using 3D spatio-temporal CNN instead of 2D CNN, the in-dataset and cross-dataset generalization performance are all considerably improved. It indicates that our method can leverage 3D CNN to effectively capture temporal defects for forgery detection. (2) Compared to RGB-RGB-3D, Freq-Freq-3D achieves better cross-dataset generalization. It suggests that frequency statistics are more generalizable features than color textures. However, RGB-RGB-3D gains better in-dataset results which may benefit from manipulation-specific artifacts. (3) We

Table 4. Evaluations on different information domains. We report video-level ACC/AUC (%) on FF++(HQ), CDF, and DFDC when trained on FF++(HQ). The highlighted row is our original setting. We developed three variants of feature extraction framework with each dropping one of the three domains, i.e., space, frequency, and time domains. The best results are in **bold**.

Method	Information Domains			FF++(HQ)		CDF		DFDC	
	Space	Frequency	Time	ACC	AUC	ACC	AUC	ACC	AUC
RGB-Freq-3D (ours)	✓	✓	✓	**98.1**	**99.7**	**74.2**	**82.6**	**66.2**	**73.8**
Freq-Freq-3D	×	✓	✓	97.6	99.0	73.9	**82.6**	64.5	72.5
RGB-RGB-3D	✓	×	✓	**98.1**	99.5	72.7	81.2	63.6	71.9
RGB-Freq-2D	✓	✓	×	96.4	98.5	68.4	76.1	61.3	69.1

note that forgery clues from these three domains work in a complementary way and contribute to the overall performance.

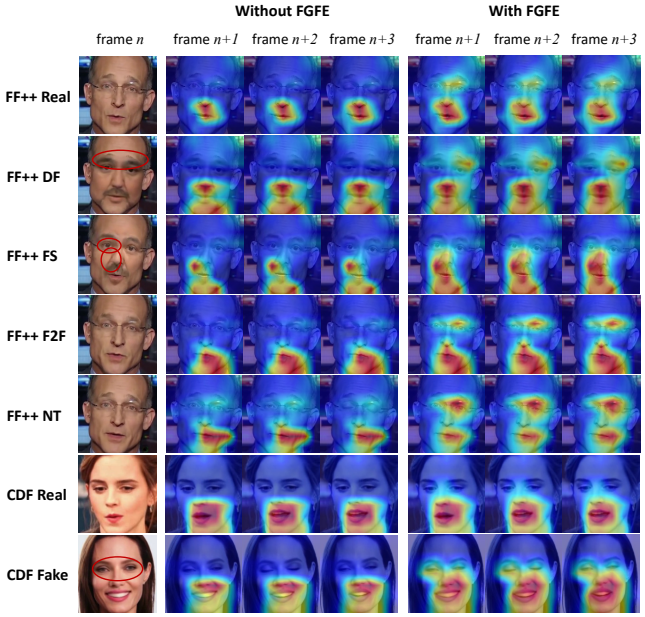


Fig. 7. The Grad-CAM visualization of localized defect regions of the model without/with FSLR-Guided Feature Enhancement (FGFE). We show several examples including four forgery types in FF++ and another dataset CDF. For each example, red circles mark visually noticeable artifacts, and consecutive frames in a video clip are provided to understand temporal defects. The warmer region suggests a higher probability of cross-domain defects the model believes.

5 Conclusion

In this paper, we propose Cross-Domain Local Forensics (XDLF), a specially designed pipeline for general deepfake video detection. Our approach aims at exploiting forgery patterns from space, frequency, and time domains simultaneously to learn a generalized cross-domain features. We also leverage four forgery-sensitive local regions to guide the model to capture subtle forgery defects. Experiments show that our method achieves impressive performance, especially strong cross-dataset generalization. We hope our work encourages future research on cross-domain forensics for more general deepfake detection.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (62271307, 61771310) and Key Lab of Information Network Security of Ministry of Public Security (The Third Research Institute of Ministry of Public Security). Shilin Wang is the corresponding author.

References

1. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: International Workshop on Information Forensics and Security (WIFS). pp. 1–7. IEEE (2018)
2. Agarwal, S., Farid, H., Fried, O., Agrawala, M.: Detecting deep-fake videos from phoneme-viseme mismatches. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2814–2822. IEEE/CVF (2020)
3. Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H.: Protecting world leaders against deep fakes. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 38–45. IEEE/CVF (2019)
4. Amerini, I., Galteri, L., Caldelli, R., Bimbo, A.D.: Deepfake video detection through optical flow based cnn. In: International Conference on Computer Vision Workshops (ICCVW). pp. 1205–1207. IEEE (2019)
5. Chai, L., Bau, D., Lim, S.N., Isola, P.: What makes fake images detectable? understanding properties that generalize. In: European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 12371, pp. 103–120. Springer (2020)
6. Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., Ji, R.: Local relation learning for face forgery detection. In: AAAI Conference on Artificial Intelligence (AAAI). pp. 1081–1088. AAAI Press (2021)
7. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1800–1807. IEEE/CVF (2017)
8. Chugh, K., Gupta, P., Dhall, A., Subramanian, R.: Not made for each other-audio-visual dissonance-based deepfake detection and localization. In: ACM International Conference on Multimedia (ACM MM). pp. 439–447. ACM (2020)
9. Cozzolino, D., Rössler, A., Thies, J., Nießner, M., Verdoliva, L.: Id-reveal: Identity-aware deepfake video detection. In: International Conference on Computer Vision (ICCV). pp. 15088–15097. IEEE/CVF (2021)
10. Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K.: On the detection of digital face manipulation. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5780–5789. IEEE/CVF (2020)
11. DeepFakes: https://github.com/deepfakes/faceswap
12. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5202–5211. IEEE/CVF (2020)
13. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Canton-Ferrer, C.: The deepfake detection challenge dataset. CoRR **abs/2006.07397** (2020)
14. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Canton-Ferrer, C.: The deepfake detection challenge (dfdc) preview dataset. CoRR **abs/1910.08854** (2019)
15. FaceSwap: https://github.com/MarekKowalski/FaceSwap
16. 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 (ICML). Proceedings of Machine Learning Research, vol. 119, pp. 3247–3258. PMLR (2020)
17. Girshick, R.B.: Fast r-cnn. In: International Conference on Computer Vision (ICCV). pp. 1440–1448. IEEE (2015)
18. Gu, Z., Chen, Y., Yao, T., Ding, S., Li, J., Huang, F., Ma, L.: Spatiotemporal inconsistency learning for deepfake video detection. In: ACM International Conference on Multimedia (ACM MM). pp. 3473–3481. ACM (2021)

19. Guera, D., Delp, E.J.: Deepfake video detection using recurrent neural networks. In: International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6. IEEE (2018)
20. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don't lie: A generalisable and robust approach to face forgery detection. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5039–5049. IEEE/CVF (2021)
21. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6546–6555. IEEE/CVF (2018)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE/CVF (2016)
23. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7132–7141. IEEE/CVF (2018)
24. Hu, Z., Xie, H., Wang, Y., Li, J., Wang, Z., Zhang, Y.: Dynamic inconsistency-aware deepfake video detection. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). pp. 736–742. ijcai.org (2021)
25. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Advancing high fidelity identity swapping for forgery detection. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5073–5082. IEEE/CVF (2020)
26. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5000–5009. IEEE/CVF (2020)
27. Li, Y., Chang, M.C., Lyu, S.: In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In: International Workshop on Information Forensics and Security (WIFS). pp. 1–7. IEEE (2018)
28. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 46–52. IEEE/CVF (2019)
29. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3204–3213. IEEE/CVF (2020)
30. de Lima, O., Franklin, S., Basu, S., Karwoski, B., George, A.: Deepfake detection using spatiotemporal convolutional networks. CoRR **abs/2006.14749** (2020)
31. 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: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 772–781. IEEE/CVF (2021)
32. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (ICLR). OpenReview.net (2017)
33. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR). OpenReview.net (2019)
34. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing face forgery detection with high-frequency features. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16317–16326. IEEE/CVF (2021)
35. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(11) (2008)

36. Masi, I., Killekar, A., Mascarenhas, R.M., Gurudatt, S.P., AbdAlmageed, W.: Two-branch recurrent network for isolating deepfakes in videos. In: European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 12352, pp. 667–684. Springer (2020)
37. Matern, F., Riess, C., Stamminger, M.: Exploiting visual artifacts to expose deepfakes and face manipulations. In: Winter Applications of Computer Vision Workshops (WACVW). pp. 83–92. IEEE (2019)
38. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: Emotions don't lie: An audio-visual deepfake detection method using affective cues. In: ACM International Conference on Multimedia (ACM MM). pp. 2823–2832. ACM (2020)
39. Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos. In: International Conference on Biometrics Theory, Applications and Systems (BTAS). pp. 1–8. IEEE (2019)
40. Nirkin, Y., Keller, Y., Hassner, T.: Fsgan: Subject agnostic face swapping and reenactment. In: International Conference on Computer Vision (ICCV). pp. 7183–7192. IEEE/CVF (2019)
41. NTech-Lab: https://github.com/NTech-Lab/deepfake-detection-challenge
42. 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 (ECCV). Lecture Notes in Computer Science, vol. 12357, pp. 86–103. Springer (2020)
43. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: International Conference on Computer Vision (ICCV). pp. 1–11. IEEE/CVF (2019)
44. Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., Natarajan, P.: Recurrent convolutional strategies for face manipulation detection in videos. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 80–87. IEEE/CVF (2019)
45. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: International Conference on Computer Vision (ICCV). pp. 618–626. IEEE (2017)
46. Sun, K., Yao, T., Chen, S., Ding, S., L, J., Ji, R.: Dual contrastive learning for general face forgery detection. CoRR **abs/2112.13522** (2021)
47. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. ACM Trans. Graph. **38**(4), 66:1–66:12 (2019)
48. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2387–2395. IEEE/CVF (2016)
49. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8692–8701. IEEE/CVF (2020)
50. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8261–8265. IEEE (2019)
51. Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (ICLR). OpenReview.net (2018)
52. Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in gan fake images. In: International Workshop on Information Forensics and Security (WIFS). pp. 1–6. IEEE (2019)

53. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N.: Multi-attentional deepfake detection. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2185–2194. IEEE/CVF (2021)
54. Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., Xia, W.: Learning self-consistency for deepfake detection. In: International Conference on Computer Vision (ICCV). pp. 15003–15013. IEEE/CVF (2021)
55. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Two-stream neural networks for tampered face detection. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1831–1839. IEEE/CVF (2017)
56. Zhou, Y., Lim, S.N.: Joint audio-visual deepfake detection. In: International Conference on Computer Vision (ICCV). pp. 14780–14789. IEEE/CVF (2021)