Adaptive Face Forgery Detection in Cross Domain

Luchuan Song[1], Zheng Fang[2*], Xiaodan Li[3], Xiaoyi Dong[1], Zhenchao Jin[1],
Yuefeng Chen[3], and Siwei Lyu[4]

[1] University of Science and Technology of China
{slc0826, dlight blwx96}@mail.ustc.edu.cn
[2] Shopee Inc.
fangzheng0827@gmail.com
[3] Alibaba Group
{fiona.lxd, yuefeng.chenyf}@alibaba-inc.com
[4] University at Buffalo
siweilyu@buffalo.edu

Abstract. It is necessary to develop effective face forgery detection methods with constantly evolving technologies in synthesizing realistic faces which raises serious risks on malicious face tampering. A large and growing body of literature has investigated deep learning-based approaches, especially those taking frequency clues into consideration, have achieved remarkable progress on detecting fake faces. The method based on frequency clues result in the inconsistency across frames and make the final detection result unstable even in the same deepfake video. So, these patterns are still inadequate and unstable. In addition to this, the inconsistency problem in the previous methods is significantly exacerbated due to the diversities among various forgery methods. To address this problem, we propose a novel deep learning framework for face forgery detection in cross domain. The proposed framework explores on mining the potential consistency through the correlated representations across multiple frames as well as the complementary clues from both RGB and frequency domains. We also introduce an instance discrimination module to determine the discriminative results center for each frame across the video, which is a strategy that adaptive adjust with during inference.

Keywords: Face Forgery Detection, Adaptive Discriminative Centers

1 Introduction

With the rapid developments of face forgery techniques [1, 2, 25, 31, 42, 45], manipulated media (the images and videos) with highly realistic forged faces can be easily generated by off-the-shelf softwares. These advanced face forgery technologies may be abused for malicious purposes, such as generating fake statement videos, causing trust issues and security concerns of the general public. Alternatively, recent studies begin to focus on various clues, *e*.g. RGB pattern [11, 15],

[*] Corresponding author.

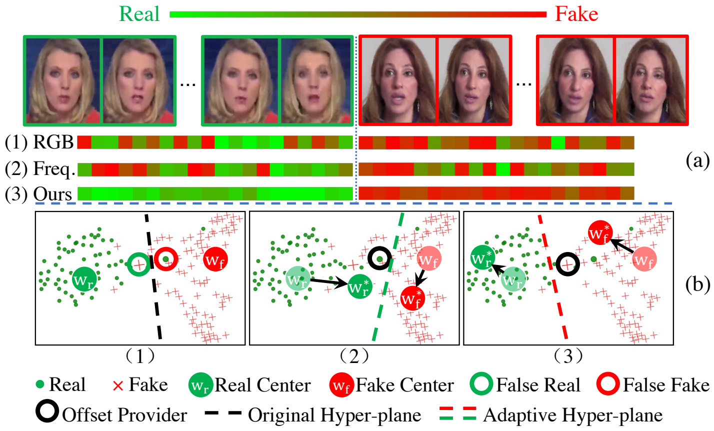

Fig. 1. (a) Probability of the face being fake (red) or real (green). (1)-(3) correspond to the prediction results from Xception, frequency based Xception and our proposed CD-Net, respectively. (b) T-SNE embedding visualization. Instances close to the hyper-plane are easy to be erroneously predicted, since the single classification hyper-plane may be not appropriate for all instances (in (b-1)). Predictions can be corrected by adjusting the hyper-plane adaptively for each instance (in (b-2,3)).

temporal feature [4, 34], optical flow [5], frequency [18, 20, 35, 40, 50], local facial region [10, 47, 55], forgery boundary [27] and biometric signals [12, 13, 37], to capture more robust features for forgery detection. These approaches achieve remarkable performance improvements on several public benchmark datasets.

Albeit the promising performance achieved by previous detection models, they are far from the ultimate solution to the problem. In particular, most existing deep convolutional neural network (CNN) based solutions suffer from several distinct limitations.

As shown in Fig. 1(a-1), though Xception-based detection [11] can make a correct prediction on a given video with voting strategies, it still makes some wrong predictions on several frames, resulting in low frame-level prediction. Moreover, the frame-level detection results are highly inconsistent across faces within multiple frames, even though they share the same identity, accessories, background and *etc.*. This means that for DeepFake videos detection, the Xception [11] cannot give a robust classification performance due to the unstable discriminative center. Apart from the RGB domain feature, several approaches [35, 40] use information in the frequency domain to boost the performance on face forgery detection. Existing methods using features based on frequency patterns are effective in general image classification [27, 49, 55] show that such models are able to find statistical features. There are some artifacts in the existing GAN-based and graphics-based face synthesis methods [24,44], it is exposed with frequency-aware clues and the performance of frequency augmented Xception can be improved to some extent [40], as shown in Fig.1(a-2). Nonetheless, the frequency features may be inadequate and there remains significant cross-frame inconsistency in the predictions. Therefore, we need to improve the cross-frame detection consistency, especially for faces from the same subject, to further improve the performance of detection algorithms.

Another limitation of existing works originates from the large intra-class distance caused by various artifacts on fake faces. We treat the forgery detection problem as a binary classification problem. It optimizes a one-fold discriminative plane $\mathcal{P}$ decided by the class centers, *i*.e.positive center $\mathbf{w}_{\mathrm{f}}$ and the corresponding negative center $\mathbf{w}_{\mathrm{r}}$. During training, these centers will be optimized to achieve an optimal discriminative performance. The final optimized classification hyper-plane based on SoftMax in previous methods is determined by the discriminative centers. This fixed hyper-plane cannot divide the real and fake faces in all frames accurately. In particular, those instances near the discriminative plane tend to be ambiguous for classification, as shown in Fig. 1(b-1). However, to the best of our knowledge, this discrimination inconsistency has not been addressed in previous face forgery detection works.

To solve the inter-frame and inter-instance instability of current forgery detection methods, we propose a novel deep architecture focusing on both *consistent* feature extraction and *discriminative* center adjusting, named as CD-Net. The proposed CD-Net consists of two novel components: *Dual-domain Intra-Consistency Module (DICM)* and *Instance-Discrimination Module (IDM)*. The DICM is designed to enhance intra-consistency by promoting the correlation of features from multiple frames in a video. In contrast with existing methods where either only the temporal consistency is considered in the final embedding (Two-branch RNN [35], STIL [22]), nor temporal information is utilized in frequency-based methods (F^3-Net [40]), the proposed DICM utilizes both temporal and frequency information at the feature level. In this way, it can boost the intra-consistent representations in both the RGB and frequency domains by extracting communal patterns existed in different frames, as shown in Fig. 1(a-3). To give a robust classification performance, we further propose a novel component *Instance-Discrimination Module (IDM)* in CD-Net, which aims to make the predictions adaptive to the features of individual instance and adjust the discriminative centers based on the features of each individual instance. As far as we know, there is no similar approach to explore the intra-class distance of various artifacts and the primary discriminative center of the whole training set is not suitable for those hard fake instances. As shown in Fig. 1(b-2,3), IDM extracts the offsets from the instance feature and adjust centers for real&fake instances to get the instance-adaptive discriminative centers $\mathbf{w}_{\mathrm{r}}^{*}$ and $\mathbf{w}_{\mathrm{f}}^{*}$. Helped with the instance-level offsets, the false predictions can be corrected with the adjusted discriminative hyper-plane determined by $\mathbf{w}_{\mathrm{r}}^{*}$ and $\mathbf{w}_{\mathrm{f}}^{*}$.

To validate the effectiveness of the proposed CD-Net for face forgery detection, we experiment on two different backbones under both the in-domain and out-domain settings. Experimental results show that the proposed method can

achieve state-of-the-art performance on various datasets, showing a promising result for face forgery detection. Our contributions are summarized as follows:
1) We introduce a *Dual-domain Intra-Consistency Module (DICM)* to improve consistency and stability of instance representation, which is extracted based on multiple frames in various domains, *i*.e.RGB and frequency patterns.
2) We introduce an *Instance-Discrimination Module (IDM)* to adjust the discriminative centers. It can dynamically adjust the position of the hyper-plane according to the input instance, which can help to improve the detection performance further.
3) We verify that our approach can achieve state-of-the-art performance on several widely-used datasets under both in-domain and out-domain settings.

2 Related Works

Spatial-Based Forgery Detection. Early approaches mainly focus on examining the appearance features in the spatial domain such as RGB or HSV color spaces. A few studies [27, 33, 55] extract color-space features for classification. GramNet [33] extracts global textures to tackle the distortion perturbations. Face X-ray [27] explores the detection task on locating the boundary of face forgery. PCL [55] employs 1x1 convolution module and extracts the relationship between each pixel in spatial. However, these approaches only utilize the RGB/spatial information, and some important features are difficult to be discovered with CNN models, especially on fake media with fewer artifacts. These clues are generally better revealed in the frequency domain, this is often the case with heavily compressed frames.
Frequency-Based Forgery Detection. Recently there is a growing number of studies [10,35,40,49,50] focus on frequency features. F^3-Net [40] uses frequency-aware image decomposition and local frequency statistics to mine forgery patterns while Two-branch RNN [35] uses the Laplacian of Gaussian operator and merges information extracted from both the RGB domain and the frequency domain. Nevertheless, the frequency information extractors used in these approaches are limited to each face itself, without considering faces of the same person in other frames of the video, and the correlation of features extracted from multiple frames in network backbone is absent.

3 CD-NET

In this section, we describe the proposed CD-Net in detail. As shown in Fig.2, given a sequence of extracted faces from the input video and its corresponding frequency maps extracted via DCT and IDCT transforms, the proposed DICM performs element-wise summation over frequency and RGB features for robust classification. Last, the dual-domain features are merged and fed into the IDM module to dynamically adjust the hyper-plane by the bias and output the final fake score of the input sequence.

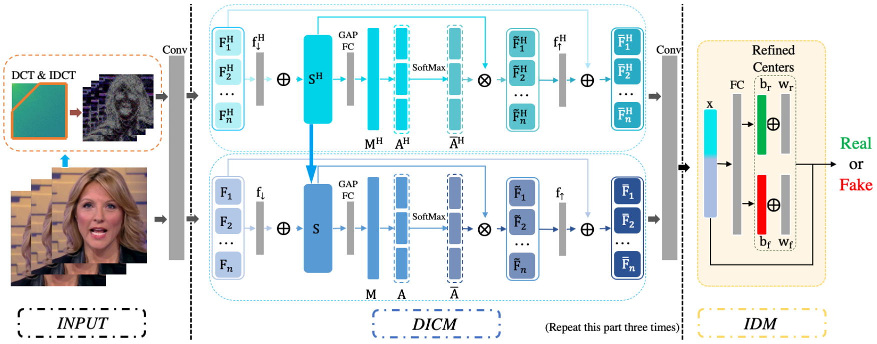

Fig. 2. The overview of the CD-Net. From left to right are the input modules, *Dual-domain Intra-Consistency Module (DICM)* and *Instance-Discrimination Module (IDM)*. The DICM takes RGB features and frequency features as input to obtain stable dual-domain feature for IDM classification. The IDM adaptively adjusting discriminative centers based on the individual instance. The "Repeat this part three times" represents the concatenation of three identical modules. $\otimes$ and $\oplus$ are element-wise multiplication and summation respectively.

3.1 Dual-domain Intra-Consistency Module

Forgery face videos are usually generated in a frame-by-frame manner. In other words, each fake frame is generated individually, which may result in artifacts in temporal dimension. However, most previous studies [33, 35, 40, 55] only adopt features from the current face frame in the backbone, without considering other frames in the video, which leads to inadequate forgery-patterns and introduces inconsistency in predictions. Though some methods [28] propose to model the sequence smoothness along the temporal dimension, their applications are limited in frame-level detection. We argue that for a robust model, the fake faces' features of the same identity in multiple frames should be consistent with each other. For this, instead of focusing on the modeling in temporal dimension, we propose a Dual-domain Intra-Consistency Module (DICM) to extract consistent representations in both the RGB and the frequency domain from the input multiple n frames to inter-act with each other. The architecture of DICM is shown in the Fig.2. Three DICM modules are cascaded in our CD-Net.

In order to improve the robustness of classification through frequency and RGB information, we extract features in a dual-domain way. For the frequency domain, given the features $\{\mathbf{F}_1^{\mathrm{H}}, \mathbf{F}_2^{\mathrm{H}}, \cdots, \mathbf{F}_n^{\mathrm{H}}\}$ of the input multiple faces where n is the number of frames ($n = 3$ in our experiments), it will be fed into the *Intra-Consistency Module (ICM)* to extract generalized communal frequency maps $\mathbf{S}^{\mathrm{H}}$. Specifically, we first perform element-wise summation over the frequency features from the sequence to acquire the common feature $\mathbf{S}^{\mathrm{H}}$, to enhance the features activated by most frames and weaken the noise features. Formally, we have:

$$\mathbf{S}^{\mathrm{H}} = \sum_{i}^{n} f_{\downarrow}^{\mathrm{H}}(\mathbf{F}_i^{\mathrm{H}}), \tag{1}$$

where $f_{\downarrow}^{\mathrm{H}}$ is an 1×1 convolution to reduce the dimension. Note that we adopt face recognition models to ensure that the faces of one input sequence are of the same identity.

After that, we use a channel-wise SoftAttention to extract the attention embedding $\bar{\mathbf{A}}_i^{\mathrm{H}}$ from frequency feature to each instance, inspired by [53]. Followed by the Global Average Pooling (GAP) layer and the fully-connected (FC) layer, the global contextual information with embedded channel-wise statistics of the frequency feature $\mathbf{S}^{\mathrm{H}}$ is gathered into the feature $\mathbf{M}^{\mathrm{H}} \in \mathbb{R}^{n \times C}$, where C is the number of channels in $\mathbf{F}_i^{\mathrm{H}}$. We further transform the $\mathbf{M}^{\mathrm{H}}$ to the feature $\mathbf{A}^{\mathrm{H}} \in \mathbb{R}^{n \times C}$ with each row represents the global contextual information for a single frame. The $\mathbf{M}^{\mathrm{H}}$ is reshaped to feature $\mathbf{A}^{\mathrm{H}} \in \mathbb{R}^{n \times C}$ with each row represents the global contextual information for a single frame. Since the communal feature required by each instance is different, a SoftMax function is performed on $\mathbf{A}_i^{\mathrm{H}}$ to get the channel-wise SoftAttention embedding $\bar{\mathbf{A}}_i^{\mathrm{H}}$. The architecture for extracting the RGB feature is the same as the frequency branch. Formally, the c-th channel of $\bar{\mathbf{A}}_i^{\mathrm{H}}$ is calculated as:

$$\bar{\mathbf{A}}_i^{\mathrm{H}}(c) = \frac{\exp(\mathbf{A}_i^{\mathrm{H}}(c))}{\sum_j^n \exp(\mathbf{A}_j^{\mathrm{H}}(c))}, \tag{2}$$

This channel-wise attention is performed on the dual-domain feature $\mathbf{S}^{\mathrm{H}}$ and $\mathbf{S}$ to extract robust supplementary feature for each instance to enhance the prediction stability. The original feature and the robust supplementary feature are summed together as the refined output $\bar{\mathbf{F}}_i^{\mathrm{H}}$:

$$\bar{\mathbf{F}}_i^{\mathrm{H}} = f_{\uparrow}^{\mathrm{H}}(\mathbf{S}^{\mathrm{H}} \otimes \bar{\mathbf{A}}_i^{\mathrm{H}}) \oplus \mathbf{F}_i^{\mathrm{H}}, \tag{3}$$

where $\otimes$ and $\oplus$ are element-wise multiplication and element-wise summation respectively. $f_{\uparrow}^{\mathrm{H}}$ is the 1×1 convolution for restoring the dimension to the size of the original input.

To make full use of the frequency information together with the RGB information, the frequency communal feature $\mathbf{S}^{\mathrm{H}}$ is merged to the RGB communal feature $\mathbf{S}$. Other parts of the architecture for extracting the RGB feature is the same as the frequency branch. The merge operation concatenates $\mathbf{S}^{\mathrm{H}}$ and $\mathbf{S}$ on channel, and then pass the concatenated features through the fully-connected layer to downsample to origin channel. The merged feature utilizes both RGB maps and frequency maps, which is then used to enhance the stability of the RGB features learned from different instances as mentioned above. Finally, the output of DICM is the concatenate of the refined frequency squence $\{\mathbf{F}_1^{\mathrm{H}}, \mathbf{F}_2^{\mathrm{H}}, \cdots, \mathbf{F}_n^{\mathrm{H}}\}$ and refined RGB squence $\{\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_n\}$.

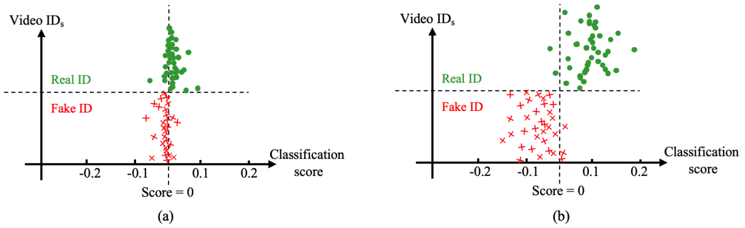

Fig. 3. Toy examples under the normalized SoftMax in (a) and Instance-Discrimination SoftMax in (b). We perform this on 30 real videos and 30 fake videos. The "Video IDs" represents different video markers, and the "Classification score" represents the prediction score get from the classification centers. From (a), the clusters are all around score 0. But from (b), the clusters will be sparser and the classification is more robust.

3.2 Instance-Discrimination Module

Apart from the prediction inconsistency of features, discriminative centers based on the whole training set is another problem for face forgery detection. The discriminative centers need to cover all data, so as to be adaptable to on unseen data. However, the original center based on the whole training set is not suitable for the requirement, especially when there are ambiguous cases near the classification hyper-plane.

We propose a novel Instance-Discrimination Module to adaptively adjust the discriminative center based on the instance itself to make robust and efficient predictions.

The input to IDM $\mathbf{x}$ is the Max-Pooling output of DICM, which is an embedding feature vector. And the corresponding label as y (real or fake), then the conditional probability output (fake score) $P(Y = y|\mathbf{x})$ by a deep neural network can be estimated via the SoftMax operator after FC layer:

$$P(Y = y|\mathbf{x}) = \frac{\exp(\mathbf{w}_y^\top \mathbf{x})}{\sum_j^N \exp(\mathbf{w}_j^\top \mathbf{x})}, \tag{4}$$

where $[\mathbf{w}_1, \cdots, \mathbf{w}_N] \in \mathbb{R}^{d\times N}$ is the weight tensor of the last fully-connected layer. N denotes the number of classes (N is 2 in our task). d is the dimension of embeddings.

The normalized SoftMax is:

$$P(Y = y|\mathbf{x}) = \frac{\exp(\tau \frac{\mathbf{w}_y^\top}{\|\mathbf{w}_y^\top\|_2} \frac{\mathbf{x}}{\|\mathbf{x}\|_2})}{\sum_j^N \exp(\tau \frac{\mathbf{w}_j^\top}{\|\mathbf{w}_j^\top\|_2} \frac{\mathbf{x}}{\|\mathbf{x}\|_2})}, \tag{5}$$

where τ is a scaling factor. Different from Eq. 4, $[\mathbf{w}_1, \cdots, \mathbf{w}_N]$ can be viewed as the discriminative centers, this is because the embedding $\mathbf{x}$ will calculate its cosine distance with $\mathbf{w}_0$ and $\mathbf{w}_1$. The $\mathbf{w}_0$ and $\mathbf{w}_1$ can be considered as positive discriminative center ($\mathbf{w}_\mathrm{r}$) and negative discriminative center ($\mathbf{w}_\mathrm{f}$). Some previous approaches [16, 39] use a fixed positive margin like Equ.5 on all instances to make predictions closer to the correct center and away from other centers. Nonetheless, the fixed positive margin is not optimal for all instances and sometimes a negative margin is better for hard cases [32].

Instance-Discrimination SoftMax. Due to the large intra-class variances in large-scale face datasets, the learned discriminative centers can not appropriately represent some instances distributed differently from the most instances (*i*.e., instances with few visible artifacts). In this paper, we propose the IDM to adaptively adjust the discriminative centers based on the instance itself. The architecture of IDM is illustrated in Fig. 2(a). Two fully-connected layers gathered with Batch Normalization and ReLU are utilized to extract bias embeddings $\mathbf{b}_{\mathrm{r}}$ and $\mathbf{b}_{\mathrm{f}}$ for classification centers $\mathbf{w}_{\mathrm{r}}$ and $\mathbf{w}_{\mathrm{f}}$, respectively. Our Instance-Discrimination SoftMax is formalized as:

$$P(Y=y|\mathbf{x}) = \frac{\exp(\tau \frac{\mathbf{w}_y^\top + \mathbf{b}_y^\top(\mathbf{x})}{\left\|\mathbf{w}_y^\top + \mathbf{b}_y^\top(\mathbf{x})\right\|_2} \frac{\mathbf{x}}{\|\mathbf{x}\|_2})}{\sum_j^N \exp(\tau \frac{\mathbf{w}_j^\top + \mathbf{b}_j^\top(\mathbf{x})}{\left\|\mathbf{w}_j^\top + \mathbf{b}_j^\top(\mathbf{x})\right\|_2} \frac{\mathbf{x}}{\|\mathbf{x}\|_2})}, \tag{6}$$

and the corresponding loss $L(\mathbf{x}, y) = -logP(Y = y|\mathbf{x})$. The IDM adjust discriminative centers based on each individual instance. To give insight into it, we compare the difference of Cosine Similarity between normalized SoftMax (T_{Norm}) and Instance-Discrimination SoftMax (T_{IDM}), specifically,

$$\begin{aligned}
T_{\mathrm{Norm}} &= \frac{\mathbf{w}^\top}{\|\mathbf{w}^\top\|_2} \frac{\mathbf{x}}{\|\mathbf{x}\|_2}, T_{\mathrm{Bias}} = \frac{\mathbf{b}^\top(\mathbf{x})}{\|\mathbf{b}^\top(\mathbf{x})\|_2};\\
T_{\mathrm{IDM}} &= \frac{\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2} \frac{\mathbf{x}}{\|\mathbf{x}\|_2}\\
&= \frac{\mathbf{w}^\top}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2} \frac{\mathbf{x}}{\|\mathbf{x}\|_2} + \frac{\mathbf{b}^\top}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2} \frac{\mathbf{x}}{\|\mathbf{x}\|_2}\\
&= \frac{\left\|\mathbf{w}^\top\right\|_2}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2} \Big(\frac{\mathbf{w}^\top}{\|\mathbf{w}^\top\|_2} \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \Big) + \frac{\left\|\mathbf{b}^\top(\mathbf{x})\right\|_2}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2} \Big(\frac{\mathbf{b}^\top(\mathbf{x})}{\|\mathbf{b}^\top(\mathbf{x})\|_2} \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \Big)\\
&= \frac{\left\|\mathbf{w}^\top\right\|_2}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2} T_{\mathrm{Norm}} + \frac{\left\|\mathbf{b}^\top(\mathbf{x})\right\|_2}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2} T_{\mathrm{Bias}}\\
&= \alpha T_{\mathrm{Norm}} - \epsilon,
\end{aligned} \tag{7}$$

where $\alpha = \frac{\left\|\mathbf{w}^\top\right\|_2}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2}$ serves as temperature on T_{Norm}. T_{Bias} is the Cosine Similarity between bias embedding and discriminative center. $\epsilon = -\frac{\left\|\mathbf{b}^\top(\mathbf{x})\right\|_2}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2} T_{\mathrm{Bias}}$ can be viewed as the adaptive margin performed on each instance.

From Equ.7, we can find that the proposed T_{IDM} consists of both original similarity T_{Norm} and the adaptive margin ϵ. Different from the fixed positive margin in previous work [16], ϵ can be either positive or negative, and is derived from the instance feature. Note that previous studies [16, 26, 39] only apply the margin on the the discriminative center corresponding to ground-truth, however, single-side margin is unreasonable for adaptive margin learning. In our Instance-Discrimination SoftMax, the adaptive margin ϵ is used in both real center and fake center to achieve the balance in the training period and force the network to learn effective bias on each center. Besides, the adaptive margin ϵ is used in both training and testing as a learned offset on discriminative centers in IDM rather than only applied in training to simply improve the intra-class compactness like previous work [16, 26]. ϵ can be viewed as the adaptive margin performed on each instance.

The final discriminative center is the summation of primary center and the predicted center bias. In contrast, when directly extracting discriminative centers from the instance feature, that is essentially the same with baseline which regresses the logits with FC. In IDM, center bias is utilized to adjust primary discriminative centers, which forces the center bias to learn the relationship between instance feature and primary discriminative centers. We give a comparison of toy examples of Instance-Discrimination SoftMax and normalized SoftMax in Fig.3, in which we use the same instance feature. The proposed Instance-Discrimination SoftMax can be treated as adaptive margins (either positive or negative) for instances, promoting the optimal performance on all cases.

4 Experiments

4.1 Setting

Datasets. We conduct experiments on several widely-used datasets, including FaceForensics++ (FF++) [42], DeeperForensics [25], Celeb-DF v2 [31] and Deepfake Detection Challenge (DFDC) dataset [17]. Both GAN-based (such as DeeperForensics [25]) and graphics-based (such as FF++ NeuralTextures [46]) forgery datasets are considered. We follow previous settings used in their corresponding datasets and compare with other methods respectively.

Metrics. We use the Area Under the Receiver Operating Characteristic Curve (AUC) and Accuracy score (Acc) as our evaluation metrics following previous methods [27, 35, 40]. In our experiments, AUC is used as the main metric since it is not affected by class imbalance and threshold. Although Acc is widely-used in face forgery detection, we assume that Acc is improper for this task, mainly caused by the sensitivity on class imbalance and the choice of threshold as mentioned in [35]. For fair comparison, Acc is calculated with the threshold of 0.5 without any threshold adjusting tricks. The video-level results are calculated by averaging all frame-level results by default.

To quantify the stability of the proposed method, two metrics are utilized in experiments, namely *Proportion of Unstable Predictions (PUP)* and *Correction Rate (CR)*. PUP_θ represents the proportion of unstable videos with the max score gap among frames higher than θ. Smaller PUP_θ indicates better stability. CR represents the proportion of same-video frame pairs which are originally get different predictions in baseline and are corrected after applying other methods. Higher CR indicates better ability on improving stability.

Implementation Details. In our experiments, Xception [11] pre-trained on the ImageNet dataset is used as the backbone. When training with IDM, we

Table 1. In-domain quantitative results on FF++ dataset with all quality settings. LQ indicates low quality (heavy compression), HQ indicates high quality (light compression) and RAW indicates videos with raw resolution. The bold results are the best. The reported approaches are spited based on whether utilizing 3D Convolution in backbone. The Acc of F^3-Net [40] with threshold of 0.5. The † implies re-implementation.

Methods	AUC (LQ)	Acc (LQ)	AUC (HQ)	Acc (HQ)	AUC (RAW)	Acc (RAW)
Steg.Features [21]	-	55.98%	-	70.97%	-	97.63%
LD-CNN [14]	-	58.69%	-	78.45%	-	98.57%
Constrained Conv [6]	-	66.84%	-	82.97%	-	98.74%
CustomPooling CNN [41]	-	61.18%	-	79.08%	-	97.03%
MesoNet [3]	-	70.47%	-	83.10%	-	95.23%
Face X-ray [27]	0.616	-	0.874	-	0.987	-
Two-branch RNN [35]	0.911	86.34%	0.991	96.43%	-	-
Xception [11]	0.925	84.11%	0.963	95.04%	0.992	98.77%
STIL† [22]	0.948	86.31%	0.986	98.57%	0.993	99.04%
PCL&I2G† [55]	0.939	87.02%	0.990	98.85%	0.997	99.78%
F^3-Net (Xception) [40]	0.933	86.89%	0.981	97.31%	0.998	99.84%
CD-Net (Xception)	**0.952**	**88.12%**	**0.999**	**98.75%**	**0.999**	**99.91%**
I3D [8]	-	87.43%	-	-	-	-
3D ResNet [23]	-	83.86%	-	-	-	-
3D ResNeXt [51]	-	85.14%	-	-	-	-
3D R50-FTCN [56]	0.966	92.35%	0.995	98.59%	0.997	99.84%
Slowfast [19]	0.936	88.25%	0.982	96.92%	0.994	99.34%
F^3-Net (Slowfast) [40]	0.958	92.37%	0.993	98.64%	0.999	99.91%
CD-Net (Slowfast)	**0.985**	**93.21%**	**0.999**	**98.93%**	**0.999**	**99.91%**

firstly fix the IDM to train the rest parameters with cross-entropy loss [54] until converged. Then we unfreeze the IDM and fine-tune the whole network. To demonstrate the generalization of the proposed methods, we also conduct experiments to validate the effectiveness of DICM and IDM on an exsiting video-based backbone, *i*.e.SlowFast R-101 [19] pre-trained on Kinetics-700 [7]. The DICM directly uses the frames in the slow pathway of SlowFast to extract the communal feature. IDM is attached at the end of SlowFast similar to Xception.

4.2 Comparison with previous methods

Face Forgery Detection. The results on FF++ are listed in Tab.1. Our CD-Net outperforms all the previous methods on all quality settings, *i*.e., LQ (c40, compressed with the quantization of 40), HQ (c23, quantization of 23) and RAW respectively. Benefited from the consistent representations on forgery patterns and instance-adaptive discriminative centers, our Xception-based model performs much better than other image-based approaches, with 0.952 in AUC and 88.12% in Acc respectively, which is even better than most video-based approaches. When utilizing the same backbone (*i*.e., Slowfast [19]), our CD-Net

Table 2. Out-domain Video-level evaluation on DFDC [17] and Celeb-DF v2 [30]. The CD-Net^{1,2,3} represents the backbone of CD-Net are Xception-raw, Xception-c23 and Xception-40 respectively. The best results are bolded. † implies re-implementation.

Methods	DFDC	Celeb-DF v2	Methods	DFDC	Celeb-DF v2
Two-Branch [35]	-	0.767	PCL&I2G [55]	0.675	0.900
CNN-aug [50]	0.721	0.756	3DR50-FTCN [56]	0.740	0.869
CNN-GRU [43]	0.689	0.698	Multi-task [38]	0.681	0.757
FWA [29]	0.695	0.673	PatchForensics [9]	0.656	0.696
Face X-ray [27]	0.655	0.795	STIL† [22]	0.661	0.715
VA-LogReg [36]	0.680	0.651	DSP-FWA [29]	0.630	0.693
Xception-raw [11]	0.709	0.655	**CD-Net**^{1}	**0.783**	0.877
Xception-c23 [11]	0.717	0.635	**CD-Net**^{2}	0.770	0.885
Xception-c40 [11]	0.709	0.655	**CD-Net**^{3}	0.753	**0.921**

gains significant improvement compare with F^3-Net (Slowfast), with 0.984 and 93.13% of AUC and Acc in comparison to 0.958 and 92.37%, in LQ task. We further calculate the confidence intervals of the AUC and Acc in Tab.1 over three repeating runs (the last three checkpoints) to verify that our model is a reliable model rather than randomly obtained. The confidence intervals are pretty small and our CD-Net is stably better than previous approaches on all quality settings even in the case of the lower bound of the score.

The comparison between our method and frequency information based methods [35,40] is shown in Tab.1. Although the frequency-based methods are greatly improved compared to the other, our method still has outstanding performance than them, which achieves the best Acc and AUC at any resolution on FF++ dataset than frequency-based methods (*e*.g., LQ (AUC): 88.12% *v.s.* 86.89%). Meanwhile, our method also has advantages over temporal-based methods [56] (*e*.g., LQ (AUC): 88.12% *v.s.* 86.31%). Our DICM is more robust as it depends on the RGB and frequency feature existed in multiple frames other than the temporal cues, which boosts the consistency of predictions from different frames.

Generalization out domain. To validate the effectiveness of the proposed method, we perform the out-domain experiments on DFDC [17] and Celeb-DF v2 [31] with the models trained on FF++ (c23) datasets. The results are listed in Tab.2. Following the previous studies [22, 55, 56], the video-level AUC scores on DFDC and Celeb-DF v2 are presented in experiments. We copy the result from the papers [25, 48, 55, 56], and we also re-implementate the results without open release code († in the Tab.2) to complete the missing values. For our CD-Net, we conduct experiments on FF++ with different resolutions and the results are listed in the Tab.2. Even when compared with strong state-of-the-art methods, *i*.e.PCL&I2G [55] and 3D R50-FTCN [56], our method still has advantages on Acc score ($0.783 v.s. 0.740$) and AUC score($0.921 v.s. 0.900$) in terms of different resolutions. The CD-Net^1 (Xception-c40) based CD-Net achieves the best performance on DFDC since the heavily compressed data boost the model to acquire more generalized patterns. Our CD-Net outperforms the Xception

Table 3. Ablation study of our method on FF++ c40 (low quality) with AUC metric to evaluate the effects of components. PUP_θ represents the proportion of videos with the frame-level score gap higher than θ, the smaller the better. CR represents the proportion of corrected unstable frame pairs in baseline, the bigger the better.

ID	DICM	IDM	AUC	$\text{PUP}_{0.7}$	$\text{PUP}_{0.5}$	$\text{PUP}_{0.3}$	CR
1	-	-	0.925	66.1%	66.1%	67.4%	-
2	√	-	0.944	29.0%	41.7%	48.9%	78.82%
3	-	√	0.938	35.7%	60.9%	66.6%	55.35%
4	√	√	**0.952**	**26.1%**	**39.4%**	**48.7%**	**80.67%**

baseline on all quality settings with much higher AUC score on Celeb-DF v2 and better Acc score on DFDC, demonstrating the robustness of our CD-Net on the unseen data.

4.3 Ablation Study

Effectiveness of DICM & IDM. To evaluate the effectiveness of the proposed DICM and IDM, we quantitatively evaluate our model and its variants: 1) the naked Xception as the baseline (ID 1), 2) Xception with DICM, 3) Xception with IDM, 4) Xception with both DICM and IDM (CD-Net). Both the AUC score and the prediction stability are reported in Tab.3. Smaller PUP_θ represents better stability with a small prediction score gap among frames. Higher CR corresponds to better ability on improving stability.

As shown in Tab.3, even if only utilizing DICM (ID 2) or IDM (ID 3), significant improvement on detection performance is achieved with AUC score 0.944 and 0.938 respectively. When using both DICM and IDM (ID 4), our method achieves the best performance with 0.952 AUC, much better than the 0.925 of baseline. Furthermore, as shown in the ROC curves in Fig. 4(a), CD-Net achieves the best performance with higher true positive rate, demonstrates the effectiveness in mining consistent representations on forgery patterns. The results of stability are positively related with detection performance where higher AUC corresponds to lower PUP and higher CR. The $\text{PUP}_{0.7}$ of CD-Net is smaller than half of the baseline with 26.1% in comparison to 66.1%, and 80.67% inconsistent frame pairs in baseline are corrected by CD-Net.

DICM. To demonstrate the benefits of utilizing correlation among frames on both frequency and RGB in DICM, we evaluate the proposed DICM and its variants by removing or replacing some components, *i*.e., 1) the proposed DICM without frequency component, denoted as DICM w/o frequency (ID 1), 2) the proposed DICM without using correlation among frames, denoted as DICM w/o correlation (ID 2). All the experiments are under the same hyper-parameters for fair comparisons. The performance of each variants is listed in left part of Tab.4. To demonstrate the improvement in DICM is not introduced by simply using multiple frames, we conduct experiments on multiple frames with other components, *i*.e., 1) Xception using 3D convolution (Conv) to correlate features among

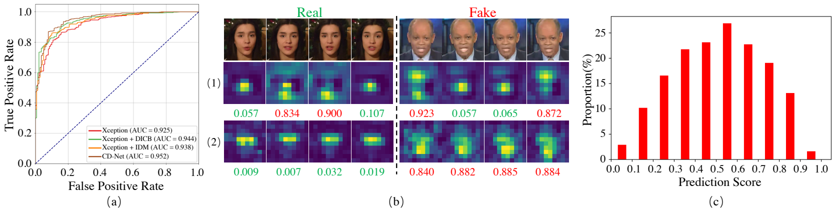

Fig. 4. (a) ROC Curve of the models in ablation studies. (b) The visualization of feature map extracted by Xception (1) and the proposed "Xception + DICM" (2). (c) The proportion of corrected instances after applying IDM relative to the total instances.

Table 4. The left part is the ablation study of DICM on FF++ c40 (low quality) to evaluate the effects of DICM. The right part is ablation study of DICM on FF++ c40 to evaluate the effects of IDM. The "+" in the table indicates different ways of combining Xception, for example, the "+ DICM" means the method of Xception+DICM.

ID	Methods (Xception)	AUC	ID	Methods (Xception)	AUC
1	+DICM w/o frequency	0.938	1*	+Softmax	0.925
2	+DICM w/o correlation	0.932	2*	+Norm Softmax [52]	0.924
3	+3D Conv	0.931	3*	+ArcFace [16]	0.923
4	+DICM w/ Original Image	0.941	4*	+SoftTriple [39]	0.926
5	+DICM w/ Low-frequency	0.939	5*	+IDM w/ Bias on Embed.	0.929
6	+DICM	**0.944**	6*	+IDM	**0.938**

different frames, denoted as "+ 3D Conv" (ID 3). The 3D Conv based model (ID 3) utilizes the similar architecture with singe Intra-consistency Module (ICM) by replacing the summation and SoftAttention with 3D Conv, and the number of parameters in 3D Conv is 27 times of ICM when the count of frames is 3. It shows the efficiency and effectiveness of our DICM. And we further demonstrate the importance of frequency feature as supplementary for RGB domain by quantitatively evaluating the DICM with different kinds of information, *i*.e., low-frequency, frequency and original image. The model with frequency components (ID 6) achieves the best scores, which indicates that frequency is more complementary with others.

To better understand the effectiveness of DICM, the visualization of feature maps extracted by Xception and the "+ DICM" are shown in Fig.4(b). Benefited from the communal patterns existed in various frames, these mispredictions in Fig.4 (b-1) have been corrected. Besides, in Fig.4(b-2), predictions achieve a consistent representation for explicitly utilizing the correlation among frames.

IDM. For better conditioning on the discriminative centers based on instance itself, our IDM is performed gathered with normalized Softmax. To demonstrate the effectiveness of the proposed IDM, we conduct experiments to compare with other classification module, *i*.e., 1) Xception with Softmax (ID 1*), 2) Xception with normalized Softmax (ID 2*), 3) Xception with ArcFace [16] (ID 3*), 4)

Xception with SoftTriple [39] (ID 4*). The result is listed in the right part of Tab.4. There is nearly no performance difference between the Softmax and the normalized Softmax, while our IDM gains a significant improvement with 0.938 in AUC. There are also some previous studies using the metric learning during the training period, such as ArcFace [16] uses a fixed positive margin and SoftTriple [39] uses multiple centers for classification. The proposed IDM achieves excellent performances comparing with these previous metric-based models. The fixed positive margin is not suitable for all instances in ArcFace while the multiple centers used in SoftTriple need the prior knowledge to pre-set the center number and an additional regulation is used to merge centers. The IDM is more versatile than the above-mentioned empirical methods and achieves excellent performances on the FF++. The average ratio between the center bias and the primary discriminative center 1 : 2.51, which demonstrates the importance of center bias on adjusting centers.

To adjust the distance of the instance relative to the discriminative centers, the bias can be either trained to modify the discriminative centers or the instance itself. However, when adjusting the discriminative centers, the model can modify both the distance between the instance relative to the discriminative centers and the distance on the internal of discriminative centers. The latter one could be served as the temperature to pull or push discriminative centers close or away from each other. We perform experiments to utilize the bias on the instance embedding, as shown in right of Tab.4(ID 5*), the result of "+IDM w/ Bias on Embedding" is 0.929 AUC score, slightly better than baseline while worse than using bias on the discriminative centers, demonstrating the advantage of adjusting on discriminative centers.

We further analyze the specific influence of IDM on different predicted score ranges in the test set. As shown in Fig.4(c), horizontal axis represents the predicted (by baseline without IDM) score range, and vertical axis shows the proportion of corrected instances relative to the total instances after applying IDM. The major corrected instances score range from 0.4 to 0.7 and are close to the original discriminative hyper-plane. Benefited from the instance-adaptive adjusting on discriminative centers in IDM, the ambiguous predictions can be corrected to maintain consistent performance on various instances.

5 Conclusions

In this paper, we propose an innovative face forgery detection framework CD-Net that repair defects of the inconsistency of forgery patterns and the suboptimal discriminative centers existed in current approaches. The proposed framework is composed of two components: DICM and IDM. The DICM utilizes the communal feature existed in multiple frames in both frequency domain and RGB domain to promote the stability and consistency. The IDM is capable of adaptively adjusting discriminative centers based on the individual instance feature. Extensive experiments demonstrate the effectiveness and significance of our approaches in in-domain detection, robustness on distortions and the unseen data.

References

1. Deepfakes, https://github.com/deepfakes/faceswap/
2. Faceswap, https://github.com/MarekKowalski/FaceSwap/
3. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–7. IEEE (2018)
4. Agarwal, S., El-Gaaly, T., Farid, H., Lim, S.N.: Detecting deep-fake videos from appearance and behavior. arXiv preprint arXiv:2004.14491 (2020)
5. Amerini, I., Galteri, L., Caldelli, R., Del Bimbo, A.: Deepfake video detection through optical flow based cnn. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
6. Bayar, B., Stamm, M.C.: A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security. pp. 5–10 (2016)
7. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987 (2019)
8. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
9. Chai, L., Bau, D., Lim, S.N., Isola, P.: What makes fake images detectable? understanding properties that generalize. arXiv preprint arXiv:2008.10588 (2020)
10. Chen, Z., Yang, H.: Manipulated face detector: Joint spatial and frequency domain attention network. arXiv preprint arXiv:2005.02958 (2020)
11. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
12. Ciftci, U.A., Demir, I., Yin, L.: Fakecatcher: Detection of synthetic portrait videos using biological signals. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
13. Ciftci, U.A., Demir, I., Yin, L.: How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals. arXiv preprint arXiv:2008.11363 (2020)
14. Cozzolino, D., Poggi, G., Verdoliva, L.: Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. pp. 159–164 (2017)
15. Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K.: On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5781–5790 (2020)
16. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
17. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C.C.: The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854 (2019)
18. Durall, R., Keuper, M., Pfreundt, F.J., Keuper, J.: Unmasking deepfakes with simple features. arXiv preprint arXiv:1911.00686 (2019)
19. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 6202–6211 (2019)

20. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. arXiv preprint arXiv:2003.08685 (2020)
21. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Transactions on Information Forensics and Security **7**(3), 868–882 (2012)
22. Gu, Z., Chen, Y., Yao, T., Ding, S., Li, J., Huang, F., Ma, L.: Spatiotemporal inconsistency learning for deepfake video detection. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3473–3481 (2021)
23. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6546–6555 (2018)
24. He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., Liu, Z.: Forgerynet: A versatile benchmark for comprehensive forgery analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4360–4369 (2021)
25. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2886–2895. IEEE (2020)
26. Kumar, A., Bhavsar, A., Verma, R.: Detecting deepfakes with metric learning. In: 2020 8th International Workshop on Biometrics and Forensics (IWBF). pp. 1–6 (2020). https://doi.org/10.1109/IWBF49977.2020.9107962
27. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5001–5010 (2020)
28. Li, X., Lang, Y., Chen, Y., Mao, X., He, Y., Wang, S., Xue, H., Lu, Q.: Sharp multiple instance learning for deepfake video detection. In: Proceedings of the 28th ACM international conference on multimedia. pp. 1864–1872 (2020)
29. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656 (2018)
30. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df (v2): a new dataset for deepfake forensics. arXiv preprint arXiv:1909.12962 (2019)
31. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3207–3216 (2020)
32. Liu, B., Cao, Y., Lin, Y., Li, Q., Zhang, Z., Long, M., Hu, H.: Negative margin matters: Understanding margin in few-shot classification. arXiv preprint arXiv:2003.12060 (2020)
33. Liu, Z., Qi, X., Torr, P.H.: Global texture enhancement for fake face detection in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8060–8069 (2020)
34. Mas Montserrat, D., Hao, H., Yarlagadda, S.K., Baireddy, S., Shao, R., Horvath, J., Bartusiak, E., Yang, J., Guera, D., Zhu, F., et al.: Deepfakes detection with automatic face weighting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 668–669 (2020)
35. Masi, I., Killekar, A., Mascarenhas, R.M., Gurudatt, S.P., AbdAlmageed, W.: Two-branch recurrent network for isolating deepfakes in videos. arXiv preprint arXiv:2008.03412 (2020)
36. Matern, F., Riess, C., Stamminger, M.: Exploiting visual artifacts to expose deepfakes and face manipulations. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). pp. 83–92. IEEE (2019)

37. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: Emotions don't lie: A deepfake detection method using audio-visual affective cues. arXiv preprint arXiv:2003.06711 (2020)
38. Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos. arXiv preprint arXiv:1906.06876 (2019)
39. Qian, Q., Shang, L., Sun, B., Hu, J., Li, H., Jin, R.: Softtriple loss: Deep metric learning without triplet sampling. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6450–6458 (2019)
40. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. arXiv preprint arXiv:2007.09355 (2020)
41. Rahmouni, N., Nozick, V., Yamagishi, J., Echizen, I.: Distinguishing computer graphics from natural images using convolution neural networks. In: 2017 IEEE Workshop on Information Forensics and Security (WIFS). pp. 1–6. IEEE (2017)
42. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1–11 (2019)
43. Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., Natarajan, P.: Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) **3**(1), 80–87 (2019)
44. Song, L., Liu, B., Yin, G., Dong, X., Zhang, Y., Bai, J.X.: Tacr-net: Editing on deep video and voice portraits. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 478–486 (2021)
45. Song, L., Yin, G., Liu, B., Zhang, Y., Yu, N.: Fsft-net: Face transfer video generation with few-shot views. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 3582–3586. IEEE (2021)
46. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics (TOG) **38**(4), 1–12 (2019)
47. Tolosana, R., Romero-Tapiador, S., Fierrez, J., Vera-Rodriguez, R.: Deepfakes evolution: Analysis of facial regions and fake detection performance. arXiv preprint arXiv:2004.07532 (2020)
48. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
49. Wang, H., Wu, X., Huang, Z., Xing, E.P.: High-frequency component helps explain the generalization of convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8684–8694 (2020)
50. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 7 (2020)
51. Wang, Y., Dantcheva, A.: A video is worth more than 1000 lies. comparing 3dcnn approaches for detecting deepfakes. In: FG'20, 15th IEEE International Conference on Automatic Face and Gesture Recognition, May 18-22, 2020, Buenos Aires, Argentina. (2020)
52. Zhai, A., Wu, H.Y.: Classification is a strong baseline for deep metric learning (2019)
53. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Mueller, J., Manmatha, R., et al.: Resnest: Split-attention networks. arXiv preprint arXiv:2004.08955 (2020)

54. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems **31** (2018)
55. Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., Xia, W.: Learning self-consistency for deepfake detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15023–15033 (2021)
56. Zheng, Y., Bao, J., Chen, D., Zeng, M., Wen, F.: Exploring temporal coherence for more general video face forgery detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15044–15054 (2021)