

# Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization

Shichao Dong<sup>1,\*</sup>, Jin Wang<sup>1,\*</sup>, Renhe Ji<sup>1,†</sup>, Jiajun Liang<sup>1</sup>, Haoqiang Fan<sup>1</sup>, Zheng Ge<sup>1</sup>  
<sup>1</sup>MEGVII Technology

{dongshichao, wangjin, jirenhe, liangjiajun, fhq, gezheng}@megvii.com

## Abstract

In this paper, we analyse the generalization ability of binary classifiers for the task of deepfake detection. We find that the stumbling block to their generalization is caused by the unexpected learned identity representation on images. Termed as the *Implicit Identity Leakage*, this phenomenon has been qualitatively and quantitatively verified among various DNNs. Furthermore, based on such understanding, we propose a simple yet effective method named the *ID-unaware Deepfake Detection Model* to reduce the influence of this phenomenon. Extensive experimental results demonstrate that our method outperforms the state-of-the-art in both in-dataset and cross-dataset evaluation. The code is available at <https://github.com/megvii-research/CADDM>.

## 1. Introduction

Recently, face-swap abusers use different face manipulation methods [17, 29, 29, 31, 65] to generate fake images/videos. Those images/videos are then used to spread fake news, make malicious hoaxes, and forge judicial evidence, which have caused severe consequences. In order to alleviate such situations, an increasing number of deepfake detection methods [12, 14, 42, 59, 60, 63] have been proposed to filter out manipulated images/videos from massive online media resources, ensuring the filtered images/videos are genuine and reliable.

Previous methods usually dealt with the task of deepfake detection with binary classifiers [1, 2, 10, 46, 51]. These methods have achieved great detection accuracy in detecting the seen attacks learned in the training datasets (*i.e.* the in-dataset evaluations). However, when confronted with media generated from newly-proposed deepfake methods (*i.e.* the cross-dataset evaluations), these methods often suffered from significant performance drops. Though plenty of researchers have designed effective methods [32, 74, 75] to

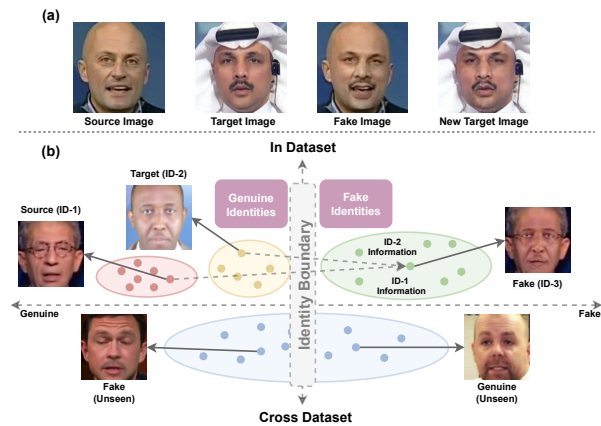


Figure 1. **The Implicit Identity Leakage phenomenon.** Since the fake image retains some features of its source image, its identity should not be completely regarded as its target image. As a consequence, there exists an implicit gap between genuine identities and fake identities in the training set, which is unintentionally captured by binary classifiers. When confronted with images manipulated by unseen face-swap methods, the classifier tends to misuse identity information and make false predictions.

improve the generalization of deepfake detection models, it still lacks a thorough analysis of why binary classifiers fail to perform well on the cross-dataset evaluation.

In this paper, given well-trained binary classifiers of deepfake detection, we find that the stumbling block for their generalization ability is caused by the mistakenly learned identity representation on images. As shown in Fig. 1 (a), a deepfake image is usually generated by replacing the face of the source image with the face of the target image. However, we notice that the procedure of synthesizing the fake image [5, 17, 29] may cause the information loss of ID representations. The identity of the fake image can not be considered as the same as either its target image or its source image. In particular, when the face of the target image is swapped back with the face of the fake image, it is noticeable that the identity of the target image is altered.

In this way, as shown in Fig. 1 (b), when learning a

\*Equal contribution

†Corresponding author

deepfake detection model, there exists an implicit decision boundary between fake images and genuine images based on identities. During the training phase, binary classifiers may accidentally consider certain groups of identities as genuine identities and other groups of identities as fake identities. When tested on the cross-dataset evaluation, such biased representations may be mistakenly used by binary classifiers, causing false judgments based on the facial appearance of images. In this paper, we have qualitatively and quantitatively verified this phenomenon (termed as the Implicit Identity Leakage) in binary classifiers of various backbones. Please see Sec. 3 and Sec. 5.2 for analyses.

Furthermore, based on such understanding, we propose a simple yet effective method named the ID-unaware Deepfake Detection Model to reduce the influence of Implicit Identity Leakage. Intuitively, by forcing models to only focus on local areas of images, less attention will be paid to the global identity information. Therefore, we design an anchor-based detector module termed as the Artifact Detection Module to guide our model to focus on the local artifact areas. Such a module is expected to detect artifact areas on images with multi-scale anchors, each of which is assigned a binary label to indicate whether the artifact exists. By localizing artifact areas and classifying multi-scale anchors, our model learns to distinguish the differences between local artifact areas and local genuine areas at a finer level, thus reducing the misuse of the global identity information.

Extensive experimental results show that our model accurately predicted the position of artifact areas and learned generalized artifact features in face manipulation algorithms, successfully outperforming the state-of-the-art. Contributions of the paper are summarized as follows:

- We discover that deepfake detection models supervised only by binary labels are very sensitive to the identity information of the images, which is termed as the Implicit Identity Leakage in this paper.
- We propose a simple yet effective method termed as the ID-unaware Deepfake Detection Model to reduce the influence of the ID representation, successfully outperforming other state-of-the-art methods.
- We conduct extensive experiments to verify the Implicit Identity Leakage phenomenon and demonstrate the effectiveness of our method.

## 2. Related Work

With the development of Generative Adversarial Network (GAN) [8, 19, 26, 27, 38] techniques, forgery images/videos have become more realistic and indistinguishable. To deal with attacks based on different face manipulation algorithms, researchers tried to improve their deepfake detectors [24, 44, 50] from different perspectives, such

as designing different loss functions [3], extracting richer features [15, 69], and analyzing the continuity between consecutive frames [23, 47]. Most of these deepfake detection methods can be roughly summarized into two categories.

### 2.1. Binary Classifiers

Many researchers [1, 2, 10, 46, 51] treated the deepfake detection task as a binary classification problem. They used a backbone encoder to extract high-level features and a classifier to detect whether the input image has been manipulated. Durall *et al.* [16] first proposed a model analyzing the frequency domain for face forgery detection. Masi *et al.* [43] used a two-branch recurrent network to extract high-level semantic information in original RGB images and their frequency domains at the same time, by which the model achieved good performance on multiple public datasets. Li *et al.* [30] designed a single-center loss to compress the real sample classification space to further improve the detection rate of forged samples. Binary classifiers achieved high detection accuracy on in-dataset evaluation, but they could not maintain good performance when facing unseen forged images.

### 2.2. Hand-crafted Deepfake Detectors

Many works attempted to improve the generalization capability of deepfake detectors by modeling specific hand-crafted artifacts among different face manipulation methods. Li *et al.* [33] believed that some physical characteristics of a real person cannot be manipulated in fake videos. They designed an eye blinking detector to identify the authenticity of the video through the frequency of eye blinking. Since 3D data can not be reversely generated from the fake image, Yang *et al.* [70] did the face forgery detection task from the perspective of non-3D projection generation samples. Sun *et al.* [61] and Li *et al.* [32] focused on precise geometric features (face landmark) and blending artifacts respectively when detecting forged images. Liu *et al.* [37] equipped the model with frequency domain information since the frequency domain is very sensitive to up-sampling operations (which are often used in deepfake detection models), and used a shallow network to extract rich local texture information, enhancing the model's generalization and robustness.

In summary, hand-crafted deepfake detectors guided the model to capture specific artifact features and indicated manipulated images/videos by responding to these features. However, these methods have a common limitation: when forgeries do not contain specific artifacts that are introduced in the training phase, they often fail to work well.

## 3. Implicit Identity Leakage

The Implicit Identity Leakage denotes that the ID representation in the deepfake dataset is captured by binary clas-

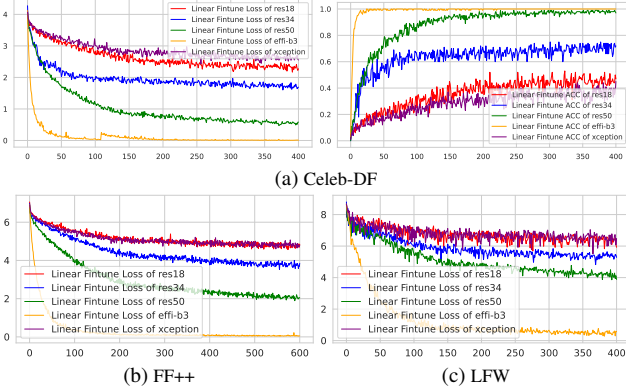


Figure 2. **ID linear classification on frozen features of binary classifiers.** Results show that binary classifiers of different backbones learned ID representation of images, even without explicit supervision of identity labels.

sifiers during the training phase. Although such identity information enhances the differences between real and fake images when testing the model on the in-dataset evaluation, it tends to mislead the model on the cross-dataset evaluation. In this section, we conduct thorough experiments to verify this hypothesis. First, we conduct the ID linear classification experiment to verify that binary classifiers capture identity information during the training phase. Second, we quantified the influence of such ID representation on the in-dataset evaluation and cross-dataset evaluation respectively, to verify its effect on the task of deepfake detection.

### 3.1. Verifying the Existence of ID Representation

**Hypothesis 1:** The ID representation in the deepfake dataset is accidentally captured by binary classifiers during the training phase when without explicit supervision.

In this section, we performed the ID linear classification experiment to verify that binary classifiers accidentally learn the ID representation on images.

Inspired by previous unsupervised pre-training methods [7, 21], we finetuned the frozen features extracted from classifiers to evaluate the generalization of the learned ID representation. Given a binary classifier trained on FF++ [56], we measured the linear classification accuracy of identities on features extracted from the classifier for FF++ [56], Celeb-DF [36] and a face recognition dataset LFW [25]. To be specific, we froze the input feature to the last linear layer of ResNet-18/34/50 [22], Xception [9] and Efficient-b3 [62] to demonstrate the universality of such phenomenon. Fig. 2 shows that linear classification on features of different classifiers converged to varying degrees and achieved varying degrees of accuracy for identity classification. Such results also indicate that although classifiers were never trained on Celeb-DF and LFW before, they still extracted substan-

Datasets	ResNet-18	ResNet-34	ResNet-50	Xception	EfficientNet-b3
FF++	81.53	89.77	99.58	97.32	94.87
Celeb-DF	46.88	47.22	49.47	47.23	44.43

Table 1. **Quantifying the influence of the ID representation on the task of deepfake detection.** Results show that although ID representation could boost the performance of the in-dataset evaluation, *i.e.* FF++, it would hinder the improvements of the cross-dataset evaluation, *i.e.* Celeb-DF.

tial information about identities from images, especially on strong backbones (*e.g.*, Efficient-b3). In other words, deepfake detectors accidentally learned the ID representation of images, without explicit supervision in particular.

### 3.2. Quantifying the Influence of ID Representation

**Hypothesis 2:** Although the accidentally learned ID representation may enhance the performance on the in-dataset evaluation, it tends to mislead the model on the cross-dataset evaluation.

After verifying the existence of ID representation in features of binary classifiers, we performed another experiment to verify its effect on deepfake detection.

The key challenge is how to attribute the output of the binary classifier to the ID representation of the input image quantitatively. Intuitively, the identity of an image is not decided by each image region individually, *e.g.* mouths, eyes and noses. Instead, these regions usually collaborate with each other to form a certain pattern, *e.g.* the identity of the input image. Thus, we used the multivariate interaction metric [71] to quantify the influence of the ID representation. Such a metric can be considered as the attribution score disentangled from the output score of the input image, which is assigned to the interaction of multiple units. Let  $N = \{1, 2, 3, \dots, n\}$  denote all the units of an input image. The multivariate interaction caused by the subset of units  $S \subseteq N$  is calculated as

$$I([S]) = \phi([S]|N_{[S]}) - \sum_{i \in S} \phi(i|N_i) \quad (1)$$

where  $\phi([S]|N_{[S]})$  denotes the Shapley value [57] of the coalition  $[S]$ , which indicates the contribution of  $[S]$  to the output score.  $\phi(i|N_i)$  denotes the Shapley value of the unit  $i$ , which indicates the contribution of the unit  $i$ .  $N_S = N/S \cup \{[S]\}$  and  $N_i = N/i \cup \{i\}$ .

In practice, to reduce the computational cost, we sampled 5 frames from each video and divided the input image into  $16 \times 16$  grids.  $S$  was set as  $S = N$  in experiments, since the input faces are usually cropped and aligned to expand the whole images as a common protocol for the deepfake detection [74, 75]. In this way, we used  $I([N])$  as the score of each image and calculated the frame-level AUC to show the effect of ID representation.

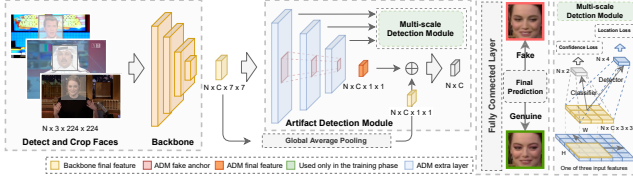


Figure 3. **The overall framework of the ID-unaware Deepfake Detection Model.**  $N$  and  $C$  denote the number of images and channels. With the help of the Artifact Detection Module, our model aims to focus on the local representation of images to indicate face forgeries.

**Understanding**  $I([N])$ .  $I([N])$  can be rewritten as  $I([N]) = \phi([N]|N_{[N]}) - \sum_{i \in N} \phi(i|N_i) = [f(N) - f(\emptyset)] - \sum_{i \in N} [f(\{i\}) - f(\emptyset)]$ . Here  $f$  denotes the binary classifier, which outputs a scalar score for an input image. Then, we have  $f(N) = I([N]) + \sum_{i \in N} [f(\{i\}) - f(\emptyset)] + f(\emptyset)$ .  $f(\emptyset)$  is a constant, representing the output of the classifier when setting the input image as the baseline value. In this way, we can disentangle the output score of the classifier  $f(N)$  into two variables, *i.e.*  $I([N])$  and  $\sum_{i \in N} [f(\{i\}) - f(\emptyset)]$ , which measures the overall interaction among multiple units and the local utility of each unit respectively. Therefore, when setting  $I([N])$  as the output score for each image, if the calculated frame-level  $AUC > 0.5$ ,  $I([N])$  can be considered to contribute positively to the classification task. Meanwhile, if the calculated frame-level  $AUC < 0.5$ ,  $I([N])$  can be considered to contribute negatively.

Results are shown in Table 1. We used ResNet-18/34/50, Xception and Efficient-b3 as the backbones of classifiers to clarify the broad effect of ID representation for deepfake detection models. All classifiers were trained on FF++. In general, when tested during the in-dataset evaluation, all classifiers achieved  $AUC > 0.5$ , indicating the enhancement for differences between real and fake images. In contrast, all classifiers achieved  $AUC < 0.5$  when tested during the cross-dataset evaluation, which indicates the misguidance of ID representation. Such results verify our hypothesis.

## 4. ID-unaware Deepfake Detection Model

Based on the understanding of the Implicit Identity Leakage, we further propose a simple yet effective method termed as the ID-unaware Deepfake Detection Model to improve the generalization ability of binary classifiers.

**Motivation.** It has been widely acknowledged that object detection modules focus on local areas of images, instead of the global representation [34, 39, 41, 73]. Inspired by the fact that local areas usually do not reflect the identity of images, we designed the Artifact Detection Module in our model to focus on local artifact areas on images, so as to pay less attention to the global identity. By preventing our model from learning the global ID representation of images,

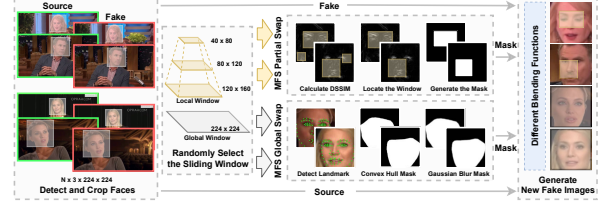


Figure 4. **Overview of the Multi-scale Facial Swap (MFS).** MFS manipulates the paired fake image and source image in two ways, namely global swap, and partial swap, to generate new fake images with the ground truth of artifact areas.

the influence of Implicit Identity Leakage can be reduced.

Moreover, to facilitate the training of the Artifact Detection Module, we propose the Multi-scale Facial Swap method to generate fake images with the ground truth of artifact area positions, which also enriches artifact features in the training phase.

### 4.1. Artifact Detection Module

The overall architecture of the Artifact Detection Module (ADM) is shown in Fig. 3. ADM takes the extracted features from the backbone as the input and detects the position of artifact areas based on multi-scale anchors. Specifically, at the end of the backbone, four extra layers of different scales are added, where the sizes of the feature maps decrease following the tuple  $(7 \times 7, 5 \times 5, 3 \times 3, 1 \times 1)$ . In the training stage, the Multi-scale Detection Module is placed after the first three extra layers, detecting artifact areas on fake images with multi-scale default anchors on images. Similar to [40, 53, 54], each feature map grid is associated with multiple default anchors with different scales on the input images. The Multi-scale Detection Module adds a detector and a classifier after each extra layer to output the position offsets ( $N \times 4$ ) and confidences of categories ( $N \times 2$ , *i.e.* the fake or genuine anchor) for each default anchor on images. A default anchor box is annotated as fake if the Intersection over Union (IoU) between the anchor box and the ground truth of artifact areas is greater than the threshold. Moreover, the final  $1 \times 1$  feature maps of ADM create a short connection with the end of the backbone, which further enriches the artifact features learned by the ADM. Its output is then fed into a fully connected layer to generate the final prediction.

To summarize, ADM determines if there exist artifact areas in multi-scale anchors. Such architecture helps our model to pay less attention to the global identity features on images, reducing the influence of Implicit Identity Leakage.

### 4.2. Multi-scale Facial Swap

The training of the Artifact Detection Module requires fine and local position annotations of artifact areas on im-



ages, which are usually not available in public deepfake datasets [36, 56]. To this end, we propose the Multi-scale Facial Swap (MFS) method, which uses multi-scale sliding windows and different blending functions to create new fake images with the position annotations of artifact areas. Besides, the new fake images also further enrich artifact features in the training set.

The procedure of MFS is shown in Fig. 4. To generate the new fake image with the position annotations of artifact areas, MFS manipulates the paired fake image and source image in two ways, i.e. global swap and partial swap. During the procedure of partial swap, MFS firstly selects a sliding window of a random size to locate the artifact area. In order to find the local area where artifacts most likely exist, the sliding window is selected by the following equation:

$$x_t, y_t = \arg \max_{x,y} \sum_{i=x}^{x+h} \sum_{j=y}^{y+w} \text{DSSIM}(I_F, I_S)_{i,j}. \quad (2)$$

where  $\text{DSSIM}(\cdot)$  indicates the structural dissimilarity [68], a larger value of which usually suggests that the areas are more probable to contain artifacts.  $x, y$  denote the top-left position of the sliding window on images.  $h, w$  denote the height and width of the sliding window.  $I_F, I_S$  denote the fake image and source image. Based on the selected sliding window, we then calculate a mask  $M$  to generate the new fake image. Specifically, we get the ground truth of the artifact area by cropping out the sliding window area on the fake image, and generating a new fake image ( $I'_F$ ) as follows:

$$I'_F = \text{BLENDING}(I_F, I_S, M). \quad (3)$$

where  $\text{BLENDING}(\cdot)$  denotes different blending methods (e.g., Poisson blending [48] and alpha blending [32]). Take alpha blending [32] as an example:  $I'_F = I_F * M + I_S * (1 - M)$ . The artifact area position of  $I'_F$  is  $[x_t, y_t, x_t + h, y_t + w]$ . During the procedure of global swap, the sliding window size is equal to the source image. As shown in Fig 4, MFS then generates the new fake image similar to Face-x-ray [32], which also provides more diverse face regions with elastic deformation [55].

Overall, with multi-scale sliding windows and different blending methods, MFS generates fake images with ground truth of artifact area positions. It supports the training of our model and further enriches artifact features in the dataset.

### 4.3. Loss Function

The overall loss function is a weighted sum of the global classification loss  $L_{cls}$  and detection loss  $L_{det}$ ,

$$L = \beta L_{det} + L_{cls}. \quad (4)$$

where  $\beta$  is a positive scalar which controls the trade-off between  $L_{det}$  and  $L_{cls}$ .

$L_{cls}$  is the cross entropy loss to measure the accuracy of the final prediction, i.e. fake or genuine images.  $L_{det}$  is the detection loss to guide the learning of ADM. Similar to [18, 40, 52], it contains confidence loss ( $L_{conf}$ ) and location loss ( $L_{loc}$ ).  $L_{conf}$  is the binary cross-entropy loss to measure the predicted result for each anchor, i.e. the fake or genuine anchors.  $L_{loc}$  is a Smooth L1 loss [18] to measure the position offsets between ADM predictions and the ground truth of artifact areas,

$$L_{det} = \frac{1}{N} (L_{conf}(x, c) + \alpha L_{loc}(x, l, g)). \quad (5)$$

where  $N$  is the number of positive anchor boxes, i.e., fake anchors;  $x \in \{0, 1\}$  is an indicator for matching the default anchor to the ground truth of artifact areas;  $c$  denotes the class confidences;  $l$  and  $g$  denote the ADM predicted box and the artifact area ground truth box;  $\alpha$  denotes a positive weight. Please see supplementary materials for details.

## 5. Experiment

In this section, we first introduced our experimental settings. Then, we compared our model with binary classifiers in terms of the Implicit Identity Leakage, to show the effectiveness of our method. After that, we explored the contribution of each component in our model. Finally, we compared our approach with other SOTA deepfake detection methods.

### 5.1. Experiment Setting

**Datasets.** We trained our models on the widely-used dataset FaceForensics++ (FF++) [56]. FF++ contains 4320 videos, i.e. 720 original videos collected from YouTube and 3600 fake videos generated by FaceShifter [31], FaceSwap [29], Face2Face [65], Deepfakes [17] and NeuralTextures [64].

We evaluated our approach performance on the following datasets: (1) FF++ [56], which contains 140 original videos and 700 fake videos. (2) DFDC-V2 [13], which has 2500 real videos and 2500 fake videos. (DFDC-V2 is widely acknowledged as the most challenging data set, since its real videos are close to life, while the artifact areas in its forgery videos are smaller than other datasets). (3) Celeb-DF [36] includes 178 real videos and all fake videos are generated by only one forgery algorithm.

**Implementation Details.** During the training phase, we set the batch size to 128 and the image size to  $224 \times 224$ . MFS sliding window scale was randomly selected from [40, 80], [80, 120], [120, 160], [224, 224]. All images were aligned by face landmarks which were extracted by a detector [72]. Similar to [32, 75], we also used regular data augmentations (DA) to further improve the model generalization, such as Random Crop, Gaussian Blur/Noise, and JPEG Compression. We select common classification models pre-trained on ImageNet [11] as the model backbones,

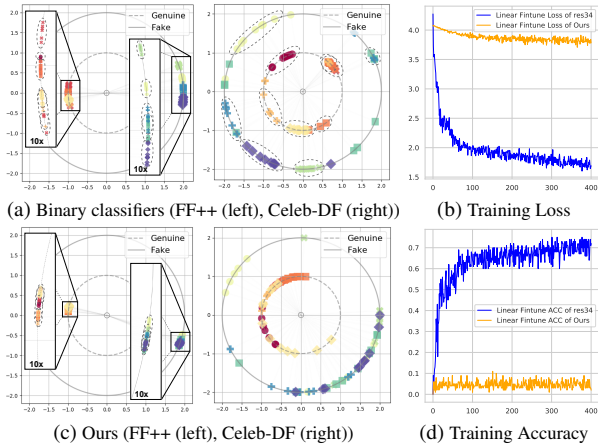


Figure 5. **Comparison in terms of the Implicit Identity Leakage between the binary classifier and our model.** Fig. 5a and 5c show the feature space with  $L_2$  normalization of the binary classifier and our model. Fig. 5b and 5d show the linear classification of identities on features of our model and the binary classifier. Results show that our model paid less attention to the ID representation of images.

including ResNet-34 [22] and EfficientNet-b3/b4 [62]. We set the number of total epochs to 200, each of which had 512 randomly selected mini-batches.  $\alpha$  and  $\beta$  in the loss function were set to 1 and 0.1 by default. The learning rate was set to  $3.6 \times 10^{-4}$  at initialization and decreased to  $1 \times 10^{-4}$  and  $5 \times 10^{-5}$  at epoch 10 and epoch 20 respectively for fine-tuning. We used Adam [28] as our optimizer. In the inference process, we chose 32 frames at an equal interval from each video, using deepfake detection video-level AUC following [32, 56] to report detector performance.

## 5.2. Experimental Analysis

**Comparison of Implicit Identity Leakage.** We designed experiments to compare our model with the binary classifier in terms of the phenomenon of Implicit Identity Leakage, in order to demonstrate that our model alleviates this problem. Each model used the EfficientNet-b3 as the common backbone and was trained on the FF++ dataset. We randomly sampled 100 images with 10 identities and used t-SNE [66] to visualize the high dimensional features extracted from the final layer of different models in 2D. Each point represents the features of an image. Different markers of points represent features of images with different identities.

In Fig. 5a (left), on the in-dataset evaluation, the binary classifier successfully distinguished the differences between the fake images and genuine images. However, features of different identities were visually separable, which shows the binary classifier partially learned the identity information of images. When tested in the Celeb-DF, as shown in Fig. 5a (right), such unnecessary knowledge about identity infor-

Models	DA	MFS	ADM	Test Set (AUC (%))		
				FF++	Celeb-DF	DFDC-V2
	×	×	×	<b>99.88</b>	64.05	48.73
ResNet-34	×	✓	×	98.70	76.35 (↑12.30)	59.97 (↑11.24)
	✓	×	×	99.74	80.07 (↑16.02)	62.46 (↑13.73)
	✓	✓	×	99.75	86.68 (↑22.63)	67.94 (↑19.21)
	✓	✓	✓	99.70	<b>91.15</b> (↑27.10)	<b>71.49</b> (↑22.76)

Table 2. **Experimental results for the effect of different components of our model.** Here DA denotes the Data Augmentations. Each model was trained by FF++ and tested on FF++, Celeb-DF, and DFDC-V2. Our model shows a significant improvement in cross-dataset evaluation.

mation tended to be misused by the binary classifier, hindering its performance.

In contrast, on the in-dataset evaluation for our model (Fig. 5c (left)), features of different identities were visually inseparable and overlapped with each other, which shows that our model reduced the influence of Implicit Identity Leakage. When tested in the Celeb-DF (Fig. 5c (right)), our model indicated fake images by detecting artifact areas with less influence of the identity information and still roughly distinguished the differences between fake images and genuine images. Such results show that ADM helped our model alleviate the phenomenon of Implicit Identity Leakage. Moreover, we also conduct further experiments to quantitatively compare our model with the binary classifier in terms of the phenomenon of Implicit Identity Leakage. Please see supplementary materials for more analysis.

Meanwhile, we conduct the same ID linear classification experiment as before to compare the existence of ID representation in features of our model and the binary classifier. We used ResNet-34 as the backbone. Fig. 5b and 5d show that linear classification on features of binary classifiers was easier to converge and achieved better accuracy than features of our model. Such results indicate that features of binary classifiers contained more information about identities than our model, further verifying the phenomenon of Implicit Identity Leakage (IIL) and the remedy effect of our method to alleviate its negative influence.

## 5.3. Ablation Studies

**Effect of Different Components.** To confirm the effectiveness of our model, we evaluated how data augmentations (DA), MFS, and ADM affected the accuracy of our model. We trained models on FF++ and tested the performance on FF++, Celeb-DF, and DFDC-V2. We denote the model without DA, MFS, and ADM as the baseline. As shown in Table 2, the baseline achieved the best performance in the in-dataset evaluation. However, the baseline only achieved 64.05% and 48.73% of AUC on Celeb-DF and DFDC-V2.

Similar to [32, 75], we also added regular data augmentations (DA), such as Random Crop, Gaussian Blur/Noise,

Models	Backbones	Test Set (AUC (%))						
		FF++	Celeb-DF					
Multi-task [45]	-	76.30	54.30					
Xception [56]	Xception	99.58	49.03					
MMMS [67]	Transformer	99.50	65.70					
SPSL [37]	Xception	96.91	76.88					
Local-Relation [6]	-	-	78.26					
Two-branch [43]	DenseNet	93.20	73.40					
DSP-FWA [35]	ResNet-50	93.00	64.60					
$F^3$ -Net [49]	Xception	98.10	65.17					
MAT [74]	Efficient-b4	99.61	68.44					
SLADD [4]	Xception	98.40	79.70					
Face-x-ray [32]	HRNet	99.17	80.58					
PCL+I2G [75]	ResNet-34	99.11	90.03					
SBI [58]	Efficient-b4	99.64	93.18					
Ours	ResNet-34	<b>99.70</b> ( $\uparrow$ 0.06)	91.15					
	Efficient-b3	<b>99.78</b> ( $\uparrow$ 0.14)	93.08					
	Efficient-b4	<b>99.79</b> ( $\uparrow$ 0.15)	<b>93.88</b> ( $\uparrow$ 0.70)					
Datasets	Ours							
	Xception [56]	SPSL [37]	PCL+I2G [75]	MAT [74]	SBI [58]	Res-34	Effi-b3	Effi-b4
DFDC-V2	45.60	66.16	67.52	70.99	72.42	71.49	<b>73.74</b>	<b>73.85</b> ( $\uparrow$ 1.43)

Table 3. **Comparison with the SOTA in FF++ (top), Celeb-DF (top) and DFDC-V2 (bottom)**. All methods were trained on FF++. Some numbers are missing because these methods do not provide training codes or pre-trained models.

and JPEG Compression to the baseline. As shown in Table 2, when DA were added, our model got AUC improvements of 16.02% and 13.73% on Celeb-DF and DFDC-V2 compared with the baseline. We argue that these augmentations may disrupt the identity information of the data to some degree and improve the model generalization.

Besides, ADM and MFS further improved the cross-dataset evaluation performance. ADM guided our model to learn artifact representations of local areas, thus reducing the influence of Implicit Identity Leakage. MFS-generated images shared similar identity information with source images (see Fig. 6 for example). Such aligned ID representation eliminated the misguidance of ID information between genuine and fake images to some degree, which also helped to reduce the influence of Implicit Identity Leakage. As shown in Table 2, beyond the effect of DA, our proposed method (*i.e.*, MFS and ADM) still led to *further yet remarkable* improvements (*e.g.*, 11.08% on Celeb-DF and 9.03% on DFDC-V2 in Tab.2). Such improvements based on a strong baseline model (*i.e.*, a model trained with DA.) could further validate the efficacy of our proposed method *w.r.t.* a weak baseline model.

#### 5.4. Comparison with state-of-the-art methods

As shown in Tab. 3, we compared our model with other deepfake detection methods on three public deepfake detection datasets. In Tab. 3, models were trained on FF++ and tested on FF++, Celeb-DF and DFDC-V2. Moreover, we also provided visual results of our method in Fig. 6, which demonstrate that our model successfully indicates fake images manipulated by different deepfake algorithms and re-

Method	Saturation	Contrast	Block	Noise	Blur	Pixel	Avg
Xception [56]	99.3	98.6	99.7	53.8	60.2	74.2	81.0
Face-x-ray [32]	97.6	88.5	99.1	49.8	63.8	88.6	81.2
LipForensics [20]	<b>99.9</b>	99.6	87.4	73.8	96.1	95.6	92.1
Ours	99.6	<b>99.8</b>	<b>99.8</b>	<b>87.4</b>	<b>99.0</b>	<b>98.8</b>	<b>97.4</b>

Table 4. **Robustness evaluation on FF++**. Results show that our method achieved better robustness image perturbations.

gresses the bounding box of the artifact areas.

**In-dataset Evaluation.** Previous methods exploiting binary classifiers usually achieved great performances on the in-dataset evaluation (*e.g.* 99.58% on FF++ for Xception [56]). Meanwhile, hand-crafted methods forced models to learn specific artifact features on images, which limited model performance on the in-dataset evaluation a bit (*e.g.* 99.17% on FF++ for Face-x-ray [32]). Compared with the above methods, by reducing the influence of Implicit Identity Leakage, our method automatically learned various artifact features on images and achieved even better performance on the in-dataset evaluations. Specifically, compared with the best performing method SBI [58], our approach (Efficient-b4 based) still improved AUC 0.15% on the FF++ for the in-dataset evaluation.

**Cross-dataset Evaluation.** Hand-crafted methods forced models to learn specific artifact features on images, improving the performance on the cross-dataset evaluation (*e.g.* 80.58% on Celeb-DF for Face-x-ray [32]). These methods can also be seen as reducing the influence of IIL by guiding models to learn hand-crafted artifact features, instead of ID-relevant information. Nevertheless, such human-defined artifact features may fail to reflect the generalized representations inside all manipulated areas, which limits their further improvements. In contrast, by reducing the influence of Implicit Identity Leakage, our method learned generalized artifact features on forgery data, achieving better generalization in the cross-dataset evaluations. In particular, in Table 3 (bottom), our model achieved 1.43% higher AUC than SBI [58] in the recently released DFDC-V2, which is widely considered as the most challenging dataset. Moreover, our approach also achieved 0.70% higher AUC for the cross-dataset evaluation on Celeb-DF in Table 3 (top).

In summary, compared with previous methods, our method significantly improved the performance on both the in-dataset and cross-dataset evaluations, showing the effectiveness of reducing the influence of Implicit Identity Leakage to learn generalized artifact features on face forgeries.

**Robustness Evaluation.** We also evaluated the robustness of our model to different kinds of image perturbations, following the same robustness experiment setting in LipForensics [20]. Results in Tab. 4 show that our method achieved 97.4% of video-level AUC on average, higher than LipForensics (*i.e.*, 92.1%).

Training set	Model	DF	F2F	FS	NT	FF++
DF	Xception [56]	99.38	75.05	49.13	80.39	76.34
	Ours+Xception [56]	<b>100.00</b>	<b>83.94</b>	<b>58.33</b>	<b>68.98</b>	<b>77.81</b> ( $\uparrow 1.47$ )
F2F	Xception [56]	87.56	99.53	65.23	65.90	79.55
	Ours+Xception [56]	<b>99.88</b>	<b>99.97</b>	<b>79.40</b>	<b>82.38</b>	<b>90.41</b> ( $\uparrow 10.86$ )
FS	Xception [56]	70.12	61.70	99.36	<b>68.71</b>	74.91
	Ours+Xception [56]	<b>93.42</b>	<b>74.00</b>	<b>99.92</b>	49.86	<b>79.30</b> ( $\uparrow 4.39$ )
NT	Xception [56]	93.09	84.82	47.98	99.50	83.42
	Ours+Xception [56]	<b>100.00</b>	<b>97.93</b>	<b>86.76</b>	<b>99.46</b>	<b>96.04</b> ( $\uparrow 12.62$ )

Training set	Model	Test Set		Training set	Model	Test Set	
		FF++	DFDC-V2			DFDC-V2	FF++
FF++	ResNet-34	<b>99.88</b>	48.73	DFDC-V2	ResNet-34	92.49	60.56
	Ours+ResNet-34	99.70	<b>71.49</b>		Ours+ResNet-34	<b>94.85</b>	<b>77.32</b>
	Effi-b3	99.75	54.12		Effi-b3	94.31	60.87
	Ours+Effi-b3	<b>99.78</b>	<b>73.74</b>		Ours+Effi-b3	<b>95.67</b>	<b>84.43</b>

Table 5. **Cross-method evaluation in FF++ (top) and generalization evaluations between FF++ and DFDC-V2 (bottom).** Our method achieved better generalization capabilities.

**Cross-method Evaluation.** We also conducted the cross-method experiment on FF++, where our model was trained on one type of manipulated data and tested on the remaining three. Results in Tab. 5 (top) show that our approach demonstrated superior cross-method generalization to binary classifiers [56] under the same setting. Such results indicate that *the impact of the learned ID representations also hindered the cross-method generalization of binary classifiers for deepfake detection to varying degrees*, which complements the proposed hypotheses in Sec. 3. Note that compared with previous hand-crafted methods like [32], which additionally enriched the training data with specific artifact features, there still exists a performance gap.

To point out, one of the most important motivations of our method lies in the goal to *train a robust deepfake detection model, which can automatically learn various yet faithful artifact features from the ever-increasing number of training data*. To this end, by reducing the impact of IIL, Sec. B in supplementary materials shows that our method successfully captured generalized artifact features as the types of training forgeries increased. Compared with previous hand-crafted methods [32, 61, 75] designed for specific artifact features, such a data-driven training scheme helps to *release the pressure to continuously devise new methods for the newly-proposed types of forgeries*, which is of great value for real-life applications. Therefore, results in Tab. 5 (top) are expected due to the limited artifact features in the original single type of forgeries, which, however, could be easily solved by scaling up the training data.

Furthermore, we evaluated the generalization capabilities of our method in a more challenging scenario. FF++ and DFDC-V2 were collected with both different manipulation methods and different original videos, making the generalization between these datasets even more difficult. To this end, Tab. 5 (bottom) demonstrates the efficacy of our method compared with binary classifiers [22, 62].

Models	FF++	Celeb-DF	DFDC-V2
SBI [58]	<b>99.64</b>	93.18	72.42
Ours+SBI [58]	99.33 ( $\downarrow 0.31$ )	<b>94.15</b> ( $\uparrow 0.97$ )	<b>79.57</b> ( $\uparrow 7.15$ )

Table 6. **Combination with the method of SBI [58].** Results show that when combining SBI with our method, the performance on the cross-dataset evaluation was significantly improved.

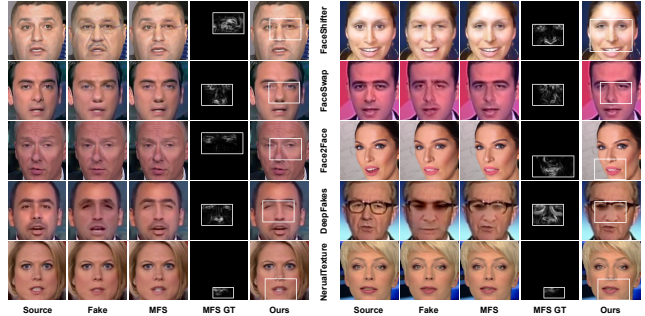


Figure 6. **Visual results on various facial manipulation algorithms.** We used NMS in [18, 40, 52] to select the bounding box position with the highest score as the final prediction. When the score was less than the threshold, no anchor box would be predicted, as was the case for real images. The result shows that our model indicated fake images based on local artifact areas.

**Potential Applicability.** Since the Artifact Detection Module is convenient to add to the end of different backbones, our method can also be potentially applied to other SOTA methods as a plug-and-play module to further boost performance. To simplify the story, we combined our method with SBI [58] for instance. The model of [58] is a binary classifier, which can be easily modified by adding the Artifact Detection Module. As shown in Table 6, when combined with SBI, our method achieved significantly better results during the cross-dataset evaluation. Such results show the potential applicability of our study, which we believe could help advance the field.

## 6. Conclusion

In this paper, we discover the phenomenon termed as Implicit Identity Leakage through experimental verification: the deepfake detection model is sensitive to the identity information of the data, which reduces the model generalization ability on unseen datasets. To this end, we propose ID-unaware Deepfake Detection Model to alleviate the Implicit Identity Leakage phenomenon. Extensive experiments demonstrate that by reducing the influence of Implicit Identity Leakage, our model successfully learns generalized artifact features and outperforms the state-of-the-art methods. In summary, this research provides a new perspective to understand the generalization of deepfake detection models, which sheds new light on the development of the field.



## References

- [1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018. [1](#), [2](#)
- [2] Belhassen Bayar and Matthew C Stamm. 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*, pages 5–10, 2016. [1](#), [2](#)
- [3] Luca Bondi, Edoardo Daniele Cannas, Paolo Bestagini, and Stefano Tubaro. Training strategies and data augmentations in cnn-based deepfake video detection. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2020. [2](#)
- [4] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18710–18719, 2022. [7](#)
- [5] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2003–2011, 2020. [1](#)
- [6] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1081–1088, 2021. [7](#)
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. [3](#)
- [8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. [2](#)
- [9] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. [3](#)
- [10] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. 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*, pages 159–164, 2017. [1](#), [2](#)
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [5](#)
- [12] Xinyi Ding, Zohreh Raziei, Eric C Larson, Eli V Olinick, Paul Krueger, and Michael Hahsler. Swapped face detection using deep learning and subjective assessment. *EURASIP Journal on Information Security*, 2020:1–12, 2020. [1](#)
- [13] Brian Dolhansky, Joanna Bitton, Ben Pfau, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv e-prints*, pages arXiv–2006, 2020. [5](#)
- [14] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9468–9478, 2022. [1](#)
- [15] Mengnan Du, Shiva Pentylala, Yuening Li, and Xia Hu. Towards generalizable forgery detection with locality-aware autoencoder. *arXiv e-prints*, pages arXiv–1909, 2019. [2](#)
- [16] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019. [2](#)
- [17] FaceSwapDevs. Deepfakes. <https://github.com/deepfakes/faceswap>, 2019. [1](#), [5](#)
- [18] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. [5](#), [8](#)
- [19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. [2](#)
- [20] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don’t lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049, 2021. [7](#)
- [21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. [3](#)
- [22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [3](#), [6](#), [8](#)
- [23] Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. *arXiv preprint arXiv:2010.00400*, 2020. [2](#)
- [24] Chih-Chung Hsu, Chia-Yen Lee, and Yi-Xiu Zhuang. Learning to detect fake face images in the wild. In *2018 International Symposium on Computer, Consumer and Control (IS3C)*, pages 388–391. IEEE, 2018. [2](#)
- [25] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in ‘Real-Life’ Images: detection, alignment, and recognition*, 2008. [3](#)
- [26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. [2](#)
- [27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the*

- IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 2
- [28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [29] Marek Kowalski. FaceSwap. <https://github.com/MarekKowalski/FaceSwap>, 2018. 1, 5
- [30] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6458–6467, 2021. 2
- [31] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 1, 5
- [32] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020. 1, 2, 5, 6, 7, 8
- [33] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018. 2
- [34] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6054–6063, 2019. 4
- [35] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018. 7
- [36] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020. 3, 5
- [37] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 772–781, 2021. 2, 7
- [38] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3673–3682, 2019. 2
- [39] Songtao Liu, Di Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 385–400, 2018. 4
- [40] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 4, 5, 8
- [41] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016. 4
- [42] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of gan-generated fake images over social networks. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 384–389. IEEE, 2018. 1
- [43] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *European Conference on Computer Vision*, pages 667–684. Springer, 2020. 2, 7
- [44] Huaxiao Mo, Bolin Chen, and Weiqi Luo. Fake faces identification via convolutional neural network. In *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security*, pages 43–47, 2018. 2
- [45] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019. 7
- [46] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Use of a capsule network to detect fake images and videos. *arXiv preprint arXiv:1910.12467*, 2019. 1, 2
- [47] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2
- [48] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003. 5
- [49] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pages 86–103. Springer, 2020. 7
- [50] Weize Quan, Kai Wang, Dong-Ming Yan, and Xiaopeng Zhang. Distinguishing between natural and computer-generated images using convolutional neural networks. *IEEE Transactions on Information Forensics and Security*, 13(11):2772–2787, 2018. 2
- [51] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2017. 1, 2
- [52] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 5, 8
- [53] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 4
- [54] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 4
- [55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmen-

- tation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5
- [56] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019. 3, 5, 6, 7, 8
- [57] Lloyd S Shapley. A value for n-person games, contributions to the theory of games, 2, 307–317, 1953. 3
- [58] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022. 7, 8
- [59] Luchuan Song, Zheng Fang, Xiaodan Li, Xiaoyi Dong, Zhenchao Jin, Yuefeng Chen, and Siwei Lyu. Adaptive face forgery detection in cross domain. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 467–484. Springer, 2022. 1
- [60] Luchuan Song, Xiaodan Li, Zheng Fang, Zhenchao Jin, Yuefeng Chen, and Chenliang Xu. Face forgery detection via symmetric transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4102–4111, 2022. 1
- [61] Zekun Sun, Yujie Han, Zeyu Hua, Na Ruan, and Weijia Jia. Improving the efficiency and robustness of deepfakes detection through precise geometric features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3609–3618, 2021. 2, 8
- [62] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 3, 6, 8
- [63] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. Detecting both machine and human created fake face images in the wild. In *Proceedings of the 2nd international workshop on multimedia privacy and security*, pages 81–87, 2018. 1
- [64] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 5
- [65] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 1, 5
- [66] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6
- [67] Junke Wang, Zuxuan Wu, Jingjing Chen, and Yu-Gang Jiang. M2tr: Multi-modal multi-scale transformers for deepfake detection. *arXiv preprint arXiv:2104.09770*, 2021. 7
- [68] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5
- [69] Xinsheng Xuan, Bo Peng, Wei Wang, and Jing Dong. On the generalization of gan image forensics. In *Chinese conference on biometric recognition*, pages 134–141. Springer, 2019. 2
- [70] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019. 2
- [71] Hao Zhang, Yichen Xie, Longjie Zheng, Die Zhang, and Quanshi Zhang. Interpreting multivariate shapley interactions in dnns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10877–10886, 2021. 3
- [72] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 5
- [73] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017. 4
- [74] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2185–2194, 2021. 1, 3, 7
- [75] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15023–15033, 2021. 1, 3, 5, 6, 7, 8