Towards A Robust Deepfake Detector: 
Common Artifact Deepfake Detection Model

Shichao Dong, Jin Wang, Renhe Ji*, 
Jiajun Liang, Haoqiang Fan, and Zheng Ge 
MEGVII Technology
{dongshichao,wangjin,liangjiajun,fanhaoqiang,jirenhe,gezheng}@megvii.com

Abstract. Existing deepfake detection methods perform poorly on face forgeries generated by unseen face manipulation algorithms. The generalization ability of previous methods is mainly improved by modeling hand-crafted artifact features. Such properties, on the other hand, impede their further improvement. In this paper, we propose a novel deepfake detection method named Common Artifact Deepfake Detection Model, which aims to learn common artifact features in different face manipulation algorithms. To this end, we find that the main obstacle to learning common artifact features is that models are easily misled by the identity representation feature. We call this phenomenon Implicit Identity Leakage (IIL). Extensive experimental results demonstrate that, by learning the binary classifiers with the guidance of the Artifact Detection Module, our method effectively reduces the influence of IIL and outperforms the state-of-the-art by a large margin, proving that hand-crafted artifact feature detectors are not indispensable when tackling deepfake problems.

Keywords: deepfake detection, face forgery, generalization.

1 Introduction

Recently, face-swap abusers use different face manipulation methods to generate fake images/videos. Those images/videos are then used to spread fake news, make malicious hoaxes, and forge judicial evidence, which has caused severe consequences. In order to alleviate such a situation, an increasing number of deepfake detection methods have been proposed to filter out manipulated images/videos from massive online media resources, ensuring filtered images/videos are genuine and reliable. Most of the deepfake detection methods have achieved high detection accuracy in detecting the seen attacks learned in the training datasets, (i.e. the in-dataset evaluations). However, when confronted with datasets generated from newly-proposed deepfake methods (i.e. the cross-dataset evaluations), previous deepfake detection models often suffer from significant performance drop.

* Corresponding author
Different from the Vanilla Binary Classifiers (VBC), our model (CADDM) also predicts the position of local artifact areas on fake images. To this end, we propose the Multi-scale Facial Swap (MFS) method to generate fake images with the annotations of artifact areas, which facilitates the training of our model. DSSIM indicates the differences on the artifact ground truth areas between the source image and the MFS-generated fake image. The heatmap shows that VBC focus on non-manipulated areas of the MFS-generated fake images, indicating the potentially biased representation of face forgeries inside the VBC.

Previous methods on deepfake detection can be roughly divided into two categories. Some of the researchers deal with the deepfake detection task with Vanilla Binary Classifiers (VBC), which are only supervised by binary labels. These methods often fail to perform well on the cross-dataset evaluation. To improve the model generalization, other recent works design models to learn hand-crafted artifacts (e.g., blending artifacts) and consider such artifacts commonly exist among various manipulated forgeries. These methods successfully achieve great model generalization, but limit their further improvements, since the assumed artifacts may not always exist on all face forgeries. In particular, researchers are constantly proposing new and high-quality face-swap algorithms. It is almost impossible and unrealistic to enumerate all of the subtle and imperceptible artifact representations. Therefore, it is of great importance to design a robust deepfake detection model, which can automatically learn common artifact representations from face forgeries.

In this paper, we come back to the start and take a deep look to explore why Vanilla Binary Classifiers (VBC) fail to perform well on the cross-dataset evaluation. Theoretically, VBC is of great potential to automatically learn generalized artifact representations on images. In this paper, we find that the main reason for the performance drop is due to mistakenly learned identity representation on images. As shown in Figure 2, we argue that when the face of the target image is swapped to the source image, the identity of the generated fake image is not the same as either its source image or its target image, which means there exist a decision boundary between fake images and genuine images based on identities. During the training phase, VBC accidentally tend to consider certain groups of identities as genuine identities and other groups of identities as fake identities. Therefore, when tested on the cross-dataset evaluation, such biased
The Implicit Identity Leakage (IIL) phenomenon. Since the fake image remains some features of its source image, its identity should not be completely regarded as its target image. As a consequence, there is a 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.

representation may be mistakenly used by VBC, thus making false judgments based on the facial appearance of images. In this paper, we have qualitatively and quantitatively verified such phenomenon (named as the Implicit Identity Leakage (IIL)) in VBC. Please see section 4.2 and supplementary materials for more analysis.

Based on such understanding, we propose a novel model named Common Artifact Deepfake Detection Model (CADDM) to reduce the influence of IIL. Intuitively, by forcing models to only focus on local areas of images, less attention will be paid to the global identity information. Therefore, we propose an anchor-based detector named Artifact Detection Module (ADM) to guide our model to detect the local artifact areas. ADM detects artifact areas on images with multi-scale anchors, each of which is assigned with a binary label, indicating whether the artifact exists. By localizing artifact areas and classifying multi-scale anchors, ADM learns to distinguish the differences between local artifact areas and local genuine areas more precisely, thus the misusage of the global identity information can be reduced. As shown in Figure 1, our ADM can also predict the position of artifact areas as a by-product.

Extensive experimental results show that our model accurately predicts the position of artifact areas and identifies deepfakes by responding to common artifacts in different face manipulated algorithms, outperforming SOTA methods by a large margin. 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 we call the Implicit Identity Leakage (IIL).
We propose a novel deepfake detector named Common Artifact Deepfake Detection Model (CADDM), outperforming other state-of-the-art methods by a large margin.

We conduct extensive experiments to verify the IIL phenomenon and demonstrate the effectiveness of CADDM.

2 Related work

With the development of Generative Adversarial Network (GAN) techniques, forgery images/videos have become more realistic and indistinguishable. To deal with attacks based on different face manipulation algorithms, researchers try to improve their deepfake detectors from different perspectives, such as designing different loss functions, extracting richer features, and analyzing the continuity between consecutive frames. Most of these deepfake detection methods can be summarized into two categories.

2.1 Vanilla Binary Classifiers

Many researchers treat the deepfake detection task as a vanilla binary classification problem. They use a backbone encoder to extract high-level features and a classifier to detect whether the input image has been manipulated. Durall et al. first propose a model analyzing the frequency domain for face forgery detection. Masi et al. use 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 achieves good performance on multiple public datasets. Li et al. design a single-center loss to compress the real sample classification space to further improve the detection rate of forged samples. Vanilla binary classifiers achieve high detection accuracy on in-dataset evaluation, but they can not maintain good performance when facing unseen forged images.

2.2 Hand-crafted Deepfake Detectors

Many works attempt to improve the generalization capability of deepfake detectors by modeling specific hand-crafted artifacts among different face manipulation methods. Li et al. believe that some physical characteristics of a real person cannot be manipulated in fake videos. They design an eye blinking detector to identify the authenticity of the video through the frequency of eye blinking. Since 3D data cannot be reversely generated from the fake image, Yang et al. do the face forgery detection task from the perspective of non-3D projection generation samples. Sun et al. and Li et al. focus on precise geometric features and blending artifacts respectively when detecting forged images. Liu et al. equip the model with frequency domain information since the frequency domain is very sensitive to upsampling.
operations (which are often used in deepfake detection models), and use a shallow network to extract rich local texture information, enhancing the model’s generalization and robustness.

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

3 Approach

To improve the model’s generalization ability across different facial forgery algorithms, we expect to build a robust common artifact detection model. This model is designed to learn common features from different artifact areas instead of capturing global identity information on images. But firstly, there are three key problems to tackle when automatically learning such common artifact features:

1. Deepfake detection models are prone to fail in concentrating on artifact areas because of the Implicit Identity Leakage (IIL) (referring to Sec. 4.2 and supplementary materials for more details).
2. There is no current method to annotate the ground truth of artifact area positions in manipulated images.
3. Artifact features in the training set are not rich enough for detectors to achieve good generalization ability.

In this section, we propose the Common Artifact Deepfake Detection Model (CADD). In CADD, the Artifact Detection Module (ADM) is designed to reduce the phenomenon of IIL. Moreover, we propose a Multi-scale Facial Swap (MFS) method to generate fake images with the ground truth of artifact area positions and also enrich artifact features in the training phase.

3.1 Artifact Detection Module

Inspired by the fact that local areas usually do not reflect the identity of images, we propose the Artifact Detection Module (ADM) to guide our model to indicate fake images based on local artifact areas. By preventing our model from learning the global identity representation of images, the influence of IIL can be reduced.

The overall architecture of ADM is shown in Figure 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. Specifically, each feature map grid is associated with multiple default anchors...
Fig. 3. The overall framework of CADDM. Here $N$ and $C$ denote the number of images and channels. The backbone is followed by an Artifact Detection Module (ADM), which aims to predict the positions of local artifact areas on images. Specifically, it uses the Multi-scale Detection Module to locate and classify artifact areas, as shown on the rightest column. Similar to SSD [32], 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 ($N \times 2$) for each default anchor on images. Note that the final $1 \times 1$ feature maps of ADM are not used for the Multi-scale Detection Module, but create a short connection with the end of the backbone. Its output is then fed into a fully connected layer to generate the final prediction.

with different scales on the input images. The Multi-scale Detection Module outputs the position offset and the predicted category (i.e. the fake or genuine anchor) for each anchor on images based on multi-scale feature maps with a set of convolutional filters. 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 0.9. In this way, ADM can predict the position of artifact areas, guiding our model to indicate images based on the local representation of images. Moreover, to make the final prediction, we also add a global average pooling layer after the backbone, which further enriches the artifact features learned by the ADM.

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

3.2 Multi-scale Facial Swap

Although the ADM reduces the IIL by predicting the positions of artifact areas, there are no position annotations in deepfake datasets to support the training. To this end, we propose the Multi-scale Facial Swap (MFS) method. MFS uses multi-scale sliding windows and different blending functions to create new fake images with the position annotations of artifact areas. In this way, MFS uses the positions of the sliding windows as the ground truth of artifact areas and further enriches artifact features in the training set.

The overall procedure of MFS is shown in Figure 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
Fig. 4. Overview of the Multi-scale Facial Swap (MFS). The 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. Specifically, MFS selects the local areas of the fake image via sliding windows of random sizes. A mask is then calculated in two alternative ways based on the size of the selected sliding window. With the mask generated above, MFS blends the fake image and its corresponding source image via Poisson blending [40] or alpha blending [26] to get the new fake image with the position annotations of artifact areas.

Specifically, MFS selects a sliding window of a random size to locate the artifact area. The new fake image is generated based on the mask \( M \) of the located area afterward. For the global swap, the sliding window size is equal to the source image. MFS generates the new fake image similar to Face-x-ray [26]. For the partial swap, we select an appropriate sliding window of other sizes, indicating the position of manipulated areas. The new fake image is generated with the mask based on the selected sliding window.

The main challenge is to select the appropriate sliding window, which shows the position of local artifact areas. Specifically, we calculate the structural dissimilarity (DSSIM) [53] between the source image and fake image to find manipulated areas. Typically, a larger value of DSSIM usually suggests that the areas are more probable to contain artifacts. In order to find the local area where artifacts most likely exist, we select the sliding window of random size by the following equation:

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

where \( x, y \) 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 calculate a mask to generate the new fake image for partial swap. 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:

\[
I'_F = \text{BLENDING}(I_F, I_S, M).
\]

\( I'_F = \text{BLENDING}(I_F, I_S, M) \).
where BLENDING(·) denotes different blending methods (e.g. Poisson blending \cite{40} and alpha blending \cite{26}). Take alpha blending \cite{40} as an example: \( I'_F = I_F * M + I_S * (1 - M) \). The artifact area position of \( I'_F \) is \([x_t, y_t, x_t + h, y_t + w]\).

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

### 3.3 Loss Function

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

\[
L = \beta L_{det} + L_{cls},
\]

where \( \beta \) is a hyper-parameter which controls the trade-off between the detection loss \( L_{det} \) and global classification loss \( L_{cls} \).

\( L_{cls} \) 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 SSD \cite{32}, 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 \cite{14} to measure the position offsets between ADM predictions and the ground truth of artifact areas,

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

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.

### 4 Experiment

In this section, we first introduce our experimental settings. Then, we verify the existence of the Implicit Identity Leakage (IIL) issue in Vanilla Binary Classifiers (VBC), as well as our method’s effectiveness in alleviating such an issue. Moreover, the verification of the common artifact shows that our model learns common artifact features from different face forgery algorithms. After that, we explore the contribution of each component in our model. Finally, we compare our approach with other state-of-the-art deepfake detection methods.

#### 4.1 Experiment Setting

**Datesets.** We make use of two training datasets, FaceForensics++ (FF++) \cite{44} and Deepfake detection Challenge (DFDC-V2) dataset \cite{10}. FF++ contains 4320 videos, including 720 original videos collected from YouTube and 3600 fake videos generated by FaceShifter (FSF) \cite{25}, FaceSwap (FS) \cite{23}, Face2Face (F2F) \cite{50}, Deepfake (DF) \cite{13} and NeuralTextures (NT) \cite{49}.
Fig. 5. The Feature space with \( L_2 \) normalization of the Vanilla Binary Classifiers (VBC) and Common Artifact Deepfake Detection Model (CADDM). We use t-SNE \([33]\) to visualize the features of different images extracted from VBC and CADDM respectively in 2D. Each point represents the features of an image. Different markers of points represent features of images with different identities. For the VBC, features of different identities are scattered and visually separable, which shows that the VBC learns the facial appearance of different identities. In contrast, for our model (CADDM), features of different identities overlap with each other and are visually inseparable, which shows the Artifact Detection Module (ADM) alleviates the influence of Implicit Identity Leakage (IIL).

We evaluate our approach performance on the following datasets: (1) FF++ \([44]\), which contains 140 original videos and 700 fake videos. (2) DFDC-V2 \([10]\), 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 \([29]\) 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, image size to \( 224 \times 224 \). MFS sliding window scale is randomly selected from [40, 80], [80, 120], [120, 160], [224, 224]. All images are aligned by face landmarks which are extracted by a detector \([56]\). Similar to \([58, 26]\), we also use 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 \([8]\) as the model backbones, including ResNet-34 \([16]\) and EfficientNet-b3 \([47]\). We set the number of total epochs to 200, each of which has 512 randomly selected mini-batches. \( \alpha \) and \( \beta \) in the loss function are set to 1 and 0.1 by default. The learning rate (LR) is set to \( 3.6 \times 10^{-4} \) at initialization and decreases to \( 1 \times 10^{-4} \) and \( 5 \times 10^{-5} \) at epoch 10 and epoch 20 respectively for fine-tuning. We use Adam \([22]\) as our optimizer. In the inference process, we choose 32 frames at an equal interval from each video, using deepfake detection accuracy (ACC) and video-level AUC following \([14, 26]\) to report detector performance.
4.2 Experimental Verification

Verification of the Implicit Identity Leakage In this part, we design an experiment to confirm the binary classifier suffers from IIL, and our model alleviates this problem. Each model uses the EfficientNet-b3 as the common backbone and is trained on the FF++ dataset. We randomly sample 100 images with 10 identities and use t-SNE 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 Figure 5(a), on the in-dataset evaluation, the VBC successfully distinguish the differences between the fake images and genuine images. However, features of different identities are visually separable, which shows the VBC partially learn the identity information of images. When tested in the Celeb-DF, as shown in Figure 5(c) such unnecessary knowledge about identity information tends to be misused by the VBC, hindering its further performance.

In contrast, on the in-dataset evaluation for our model (Figure 5(b)), features of different identities are visually inseparable and overlap with each other, which shows that our model reduces the influence of IIL. When tested in the Celeb-DF (Figure 5(d)), our model indicates fake images by detecting artifact areas without the influence of the identity information and still roughly distinguishes the differences between fake images and genuine images. Such results show that ADM helps the model alleviate the IIL. Moreover, we also conduct further experiments to quantitatively verify the phenomenon of IIL. Please see Sec. 1.1 in supplementary materials for more analysis.

Verification of the Common Artifact By reducing the influence of IIL, the anchor-based detection architecture of ADM helps the model to capture common features from different artifact areas. Thanks to the division of training data in the FF++ dataset according to different face forgery methods, we split...
Table 1. Experimental results for the effect of different components of our model. Here DA denotes the Data Augmentations. Each model is trained by FF++ [44] and tested on FF++ [44], Celeb-DF [29], and DFDC-V2 [10]. Our model shows a significant improvement on cross-dataset evaluation.

| Models   | DA | MFS | ADM | Test Set (AUC (%)) |
|----------|----|-----|-----|--------------------|
|          |    |     |     | FF++ [44] | Celeb-DF [29] | DFDC-V2 [10] |
| ResNet-34 | ×  | ×   | ×   | 99.88 | 64.05 | 48.73 |
|          | ×  | ✓   | ×   | 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 | 91.15 (↑27.10) | 71.49 (↑22.76) |

five sub-training sets following Figure 6(a) and further explore the relationship between the number of manipulation algorithms and the model performance in the in-dataset and cross-dataset evaluation. Both our model and the VBC use ResNet-34 [16] as the backbone and test the performance on FF++ [44] and Celeb-DF [29] respectively.

As shown in Figure 6(b), due to the IIL, even if the number of manipulation algorithms in the training set increases, the binary classifier still maintains poor performance on the cross-dataset evaluation. In contrast, common artifacts captured by our model are more generalized. Moreover, our approach also achieves higher performance in in-dataset evaluation when using less training data. Figure 6(c) indicates that when only using the Face2Face [50] sub-dataset as the training set, compared with the binary classifier, our approach gets 17% AUC improvements in FF++ [44].

4.3 Ablation Studies

Effect of CADDM To confirm the effectiveness of our model (CADDMM), we evaluate how data augmentations (DA), MFS, and ADM affect the accuracy of our model. We train models in FF++ [44] and test the model performance on FF++ [44], Celeb-DF [29], and DFDC-V2 [10]. We denote the model without DA, MFS, and ADM as the baseline. As shown in Table 1, the baseline achieves the best performance in the in-dataset evaluation. However, the baseline only achieves 64.05% and 48.73% of AUC on Celeb-DF and DFDC-V2.

To make a fair comparison between our model and other methods, we also add regular data augmentations, such as Random Crop, Gaussian Blur/Noise, and JPEG Compression, to the baseline, similar to [20]. As shown in Table 1 when DA are added, the model gets AUC improvements of 16.02% and 13.73% on Celeb-DF and DFDC-V2 compared with the baseline. We argue that these augmentations also weaken the identity information of the data and further improve the model generalization.

Besides, ADM and MFS further improve the cross-dataset evaluation performance supervised by local artifact regions. ADM guides our model to learn artifact representations of local areas, thus reducing the influence of IIL.
Table 2. Model performance based on different backbones. Our proposed CADDM achieves better performance when using a stronger classification model as the backbone.

| Models       | Training Set | FF++ [44] | Celeb-DF [29] | DFDC-V2 [10] |
|--------------|--------------|-----------|--------------|--------------|
|              |              | ACC       | AUC          | ACC          | AUC          | ACC          | AUC          |
| ResNet-34 [16] | FF++ [44]    | 96.71     | 99.70        | 81.78        | 91.15        | 62.68        | 71.49        |
| EfficientNet-b3 [47] | FF++ [44]    | 98.36     | 99.77        | 85.14        | 90.69        | 63.65        | 73.57        |
| ResNet-34 [16] | DFDC-V2 [10] | 69.79     | 77.32        | 83.59        | 91.45        | 87.01        | 94.85        |
| EfficientNet-b3 [47] | DFDC-V2 [10] | 77.29     | 84.43        | 87.84        | 93.46        | 87.99        | 95.67        |

Table 3. Comparison with the state-of-the-art in FF++ [44] and Celeb-DF [29]. Some numbers are missing because these methods do not provide training codes or pre-trained models.

| Models       | Backbones       | Test Set (AUC (%)) |
|--------------|-----------------|--------------------|
|              |                 | FF++ [44] | Celeb-DF [29] |
| Xception     | Xception [44]   | 99.58     | 49.03        |
| MMMS         | Transformer [21] | 99.50     | 65.70        |
| SPSL         | Xception [40]   | 96.91     | 76.88        |
| Local-Relation | -                | -         | 78.26        |
| Two-branch   | DenseNet [19]   | 93.20     | 73.40        |
| Multi-task   | Self-design     | 76.30     | 54.30        |
| DSP-FWA      | ResNet-50 [16]  | 93.00     | 64.60        |
| F3-Net       | Xception [41]   | 98.10     | 65.17        |
| MAT          | Efficient-b4 [47] | 99.61    | 68.44        |
| Face-x-ray   | HRNet [45]      | 99.17     | 80.58        |
| PCL+12G      | ResNet-34 [10]  | 99.11     | 90.03        |
| CADDM (ours) | ResNet-34 [10]  | 99.70     | 91.15 (↑1.12) |
|              | Efficient-b3 [47] | 99.77     | 90.69 (↑0.66) |

generated images share similar identity information with source images (see Fig 1 for example), which also helps to reduce the influence of IIL. As shown in Table 1, ADM and MFS show a significant improvement in cross-dataset evaluation.

Effect of Different Backbones A robust deepfake detection model with a stronger backbone should have better performance in both the in-dataset and cross-dataset evaluations. To this end, we design a different-backbones experiment to further verify the generalization of our model. In this experiment, we choose ResNet-34 [16] and EfficientNet-b3 [47] as backbones and compare their performances on unseen datasets. As shown in Table 2, when the training set is FF++ [44], EfficientNet-b3 based CADDM (ours) gets 2.08% AUC improvements on DFDC-V2 [10]. Moreover, using EfficientNet-b3 [47] as the backbone, our model trained by DFDC-V2 [10] improves AUC from 94.85% to 95.67% in the in-dataset evaluation and averaged AUC from 84.39% to 88.95% in the cross-dataset evaluation. This phenomenon verifies that our model has a high generalization ability when using different classification models as backbones. Moreover,
4.4 Comparison with state-of-the-art methods

As shown in Table 3 and Table 4, we compare our model with other deepfake detection methods on three public deepfake detection datasets. In Table 3, all methods are trained on FF++ [44] and tested on FF++ [44] and Celeb-DF [29]. In Table 4, all methods are trained on FF++ [44] and tested on DFDC-V2 [10].

**In-dataset evaluation** Previous methods exploiting Vanilla Binary Classifiers (VBC) usually achieve great performances on the in-dataset evaluation (e.g. 99.58% on FF++ for Xception [44]). Meanwhile, hand-crafted methods force models to learn specific artifact features on images, which limits model performance on the in-dataset evaluation a bit (e.g. 99.17% on FF++ for Face-x-ray [29]). Compared with the above methods, by reducing the influence of IIL, our method automatically learns common artifact features on images and achieves even better performance on the in-dataset evaluations. Specifically, compared with the best performing method MAT [57], our approach (Efficient-b3 based) still improves AUC 0.16% on the FF++ for the in-dataset evaluation.

**Cross-dataset evaluation** In this paper, we find that Vanilla Binary Classifiers (VBC) partially learn the identity information on images, thus overfitting certain datasets. When faced with unseen attacks, VBC tend to misuse the identity information of images to predict the results, thus performing poorly on the cross-dataset evaluation (49.03% on Celeb-DF for Xception [44]). To this
end, hand-crafted methods force 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 [26]). These methods can also be seen as reducing the influence of IIL by guiding models to learn hand-crafted artifact features, instead of the identity-relevant information. Even though such methods improve the generalization to some degree, such hand-crafted artifact features usually represent human’s understanding of artifacts and can not reflect the common representations inside the manipulated areas, thus limiting further improvements. In contrast, by reducing the influence of IIL, our method learns common artifact features on face forgeries, achieving better generalization in the cross-dataset evaluations. Specifically, our approach (ResNet-34 based) achieves 1.12% higher AUC than the state-of-the-art method [58] for the cross-dataset evaluation on Celeb-DF. Moreover, in Table 4, our model also achieves 2.58% higher AUC than MAT [57] in the recently released DFDC-V2 [10], which is widely considered as the most challenging dataset.

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

Limitation to video compression Although our model achieves a better performance in cross-dataset evaluation, our method still has some limitations. As shown in Table 5, our framework achieves 97.53% and 83.09% AUC on the C23 and C40 of the FF++ [44] dataset, which is still a gap compared with recent works [57]. We argue that our model pays more attention to local areas. Artifact features on such areas are usually less significant on compressed videos, causing the performance drop.

5 Conclusion

In this paper, we discover the phenomenon called Implicit Identity Leakage (IIL) through experimental verification: the deepfake detection model is very sensitive to the identity information of the data, which reduces the model generalization ability on unseen datasets. To this end, we propose Common Artifact Deepfake Detection Model (CADDM) to alleviate the IIL phenomenon. As shown in Figure 7, the visual results demonstrate that our model indicates fake images manipulated by different face swap algorithms and regresses the bounding box of the artifact areas. Moreover, extensive experiments demonstrate that by reducing the influence of IIL, our model successfully learns common artifact features and outperforms the state-of-the-art by a large margin. In summary, this research provides new insights into the model generalization on deepfake detection, showing that hand-crafted artifact feature detectors are not indispensable.
Fig. 7. Visual results on various facial manipulation algorithms. We use NMS in [32] to select the bounding box position with the highest score as the final prediction. When the score is less than the threshold, no anchor box will be predicted, as is the case for real images. The result shows that our model indicates fake images based on local artifact areas.

References

1. 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)
2. 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)
3. Bondi, L., Cannas, E.D., Bestagini, P., Tubaro, S.: Training strategies and data augmentations in cnn-based deepfake video detection. In: 2020 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–6. IEEE (2020)
4. Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., Ji, R.: Local relation learning for face forgery detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1081–1088 (2021)
5. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8188–8197 (2020)
6. 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)
7. 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)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Ding, X., Raziei, Z., Larson, E.C., Olinick, E.V., Krueger, P., Hahsler, M.: Swapped face detection using deep learning and subjective assessment. EURASIP Journal on Information Security 2020, 1–12 (2020)
10. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Canton Ferrer, C.: The deepfake detection challenge dataset. arXiv e-prints pp. arXiv–2006 (2020)
11. Du, M., Pentyala, S., Li, Y., Hu, X.: Towards generalizable forgery detection with locality-aware autoencoder. arXiv e-prints pp. arXiv–1909 (2019)
12. Durall, R., Keuper, M., Pfreundt, F.J., Keuper, J.: Unmasking deepfakes with simple features. arXiv preprint arXiv:1911.00686 (2019)
13. FaceSwapDevs: Deepfakes. https://github.com/deepfakes/faceswap (2019)
14. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems 27 (2014)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. Hernandez-Ortega, J., Tolosana, R., Fierrez, J., Morales, A.: Deepfakes onphys: Deepfakes detection based on heart rate estimation. arXiv preprint arXiv:2010.00400 (2020)
18. Hsu, C.C., Lee, C.Y., Zhuang, Y.X.: Learning to detect fake face images in the wild. In: 2018 International Symposium on Computer, Consumer and Control (IS3C). pp. 388–391. IEEE (2018)
19. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 2261–2269. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.243
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
21. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Kowalski, M.: FaceSwap. https://github.com/MarekKowalski/FaceSwap (2018)
24. Li, J., Xie, H., Li, J., Wang, Z., Zhang, Y.: 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. pp. 6458–6467 (2021)
25. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457 (2019)
26. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5001–5010 (2020)
27. Li, Y., Chang, M.C., Lyu, S.: In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–7. IEEE (2018)
28. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656 (2018)
29. 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)

30. Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., Yu, N.: Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 772–781 (2021)

31. Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., Wen, S.: Stgan: A unified selective transfer network for arbitrary image attribute editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3673–3682 (2019)

32. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)

33. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research 9(11) (2008)

34. Marra, F., Gragnaniello, D., Cozzolino, D., Verdoliva, L.: Detection of generated fake images over social networks. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 384–389. IEEE (2018)

35. Masi, I., Killekar, A., Mascarenes, R.M., Gurudatt, S.P., AbdAlmageed, W.: Two-branch recurrent network for isolating deepfakes in videos. In: European Conference on Computer Vision. pp. 667–684. Springer (2020)

36. Mo, H., Chen, B., Luo, W.: Fake faces identification via convolutional neural network. In: Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. pp. 43–47 (2018)

37. 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)

38. Nguyen, H.H., Yamagishi, J., Echizen, I.: Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467 (2019)

39. Nirkin, Y., Wolf, L., Keller, Y., Hassner, T.: Deepfake detection based on discrepancies between faces and their context. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)

40. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: ACM SIGGRAPH 2003 Papers, pp. 313–318 (2003)

41. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: European Conference on Computer Vision. pp. 86–103. Springer (2020)

42. Quan, W., Wang, K., Yan, D.M., Zhang, X.: Distinguishing between natural and computer-generated images using convolutional neural networks. IEEE Transactions on Information Forensics and Security 13(11), 2772–2787 (2018)

43. 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)

44. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Forensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–11 (2019)

45. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5693–5703 (2019)
46. Sun, Z., Han, Y., Hua, Z., Ruan, N., Jia, W.: 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. pp. 3609–3618 (2021)

47. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019)

48. Tariq, S., Lee, S., Kim, H., Shin, Y., Woo, S.S.: Detecting both machine and human created fake face images in the wild. In: Proceedings of the 2nd international workshop on multimedia privacy and security. pp. 81–87 (2018)

49. Thies, J., Zollhoefer, M., Neissner, M.: Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics (TOG) 38(4), 1–12 (2019)

50. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Neissner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2387–2395 (2016)

51. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008 (2017), https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

52. Wang, J., Wu, Z., Chen, J., Jiang, Y.G.: M2tr: Multi-modal multi-scale transformers for deepfake detection. arXiv preprint arXiv:2104.09770 (2021)

53. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13(4), 600–612 (2004)

54. Xuan, X., Peng, B., Wang, W., Dong, J.: On the generalization of gan image forensics. In: Chinese conference on biometric recognition. pp. 134–141. Springer (2019)

55. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8261–8265. IEEE (2019)

56. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters 23(10), 1499–1503 (2016)

57. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N.: Multi-attentional deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2185–2194 (2021)

58. Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., Xia, W.: Learning to recognize patch-wise consistency for deepfake detection. arXiv preprint arXiv:2012.09311 (2020)

59. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
6 Supplementary Materials

6.1 Quantitative Analysis of IIL

In this section, we conduct two experiments to evaluate the influence of Implicit Identity Leakage (IIL) quantitatively.

![Graph (a)](image1.png) ![Graph (b)](image2.png)

(a) In-dataset evaluation(FF++). (b) Cross-dataset evaluation(Celeb-DF).

Fig. 8. Quantitative evaluation on the influence of Implicit Identity Leakage (IIL). Results show that compared with VBC, features of different identities in CADDM overlap with each other more. Such results are more prominent on cross-dataset evaluations (Figure 8(b)). Here, we count the number of overlapping identities (IDs) for each identity in the feature space. Specifically, we use principal component analysis (PCA) to project features of images into 2D space. We consider the rectangle area of the projected features for each identity as the region of the identity. Then, two identities are considered to overlap with each other if the Intersection over Union (IoU) between their regions is no less than the threshold. In this figure, given different values of the threshold, each box-and-whisker denotes the distribution for the number of overlapping IDs across all identities in the dataset. Results show that features of different identities in CADDM overlap with each other more, especially on the cross-dataset evaluation (Figure 8(b)). Such results demonstrate that CADDM significantly reduces the influence of IIL.

ID Overlap Experiment When the model extracts identity information from images, we argue that features of different identities tend to be roughly separable. In other words, features of each identity are unlikely to overlap with others. In this way, we expect to measure whether features of different identities are separable in the feature space by counting the number of overlapping identities (IDs) for each identity. To ensure the diversity of images for each identity, we sample 5 images at an equal interval from each video for all identities in the dataset. Then, we use principal component analysis (PCA) to project features of images into 2D space. For each identity, we consider the rectangle area of its
Accuracy of ID Classification

Number of Iterations

(a) Accuracy of linear classification on identities.

Softmax Loss

Number of Iterations

(b) Loss of linear classification on identities.

Fig. 9. Linear classification of identities on normalized features of the Artifact Detection Module (ADM) and Vanilla Binary Classifiers (VBC) for Celeb-DF [29] (56 identities). Results show that linear classification on features of VBC is easier to converge and achieves better accuracy. Such results indicate that features of VBC contain more identity information, verifying the phenomenon of IIL.

projected features as the region of the identity. Two identities are considered to overlap with each other when the Intersection over Union (IoU) of their regions is no less than the threshold.

We evaluate the influence of IIL on Vanilla Binary Classifiers (VBC) and our proposed Common Artifact Deepfake Detection Model (CADMM) quantitatively. Both VBC and CADMM are trained on FF++ [44] and tested on FF++ [44] and Celeb-DF [29]. Figure 8 shows the distributions for the number of overlapping IDs across all identities in the dataset, given different values of the threshold. Compared with VBC, features of different identities in CADMM overlap with each other more, thus less separable. Such results are more prominent on cross-dataset evaluations (Figure 8(b)). Such results demonstrate that our CADMM significantly alleviates the influence of IIL.

ID Linear Classification Experiment Besides, we conduct another experiment to further verify the phenomenon of IIL quantitatively. Specifically, we measure the linear classification accuracy of identities on normalized features of ADM and VBC for Celeb-DF [29] (56 identities) respectively. Figure 9 shows
Table 5. To show that our CADDM learns common artifact features in the training set, we provide more experimental results on the verification of the common artifact. Same as Figure 6(a) in our paper, the number of forgery methods in the training set FF++ [44] increases following the order of A-E. Each model is tested on Celeb-DF [29]. To make a fair comparison with a stronger baseline, we also conduct experiments to add data augmentations to the baseline, similar to [58,26]. We use Frame-level AUC (FAUC) and Video-level AUC (VAUC) as our metrics. Results show that when the training set contains more face manipulation algorithms, CADDM learns more common artifact features, showing higher performances on the cross-dataset evaluation.

| FF++ [44] Sub-dataset | Baseline | + Data Augmentations | CADDM |
|-----------------------|----------|----------------------|-------|
|                       | FAUC     | VAUC                 | FAUC  |
| A                     | 59.01    | 61.16                | 62.84 |
|                       |          |                      | 70.23 |
| B                     | 57.54    | 60.76                | 66.05 |
|                       |          |                      | 70.25 |
| C                     | 58.94    | 63.77                | 68.33 |
|                       |          |                      | 74.18 |
| D                     | 57.24    | 64.38                | 71.35 |
|                       |          |                      | 79.50 |
| E                     | 58.70    | 64.05                | 71.45 |
|                       |          |                      | 80.07 |

that linear classification on features of VBC is easier to converge and achieves better accuracy than features of our model. Such results indicate that features of VBC contain more information of identities than our model, further verifying the phenomenon of IIL.

6.2 More on Verification of the Common Artifact

To confirm that CADDM learns common artifact features in the training set, we provide more experiment results on the verification of the common artifact in Table 5. To make a fair comparison with a stronger baseline, we conduct experiments to add data augmentations, such as Random Crop, Gaussian Blur/Noise, and JPEG Compression, to the baseline, similar to [58,26]. ResNet-34 [16] is used as the backbone, and each model is tested on Celeb-DF [29]. We use Frame-level AUC (FAUC) and Video-level AUC (VAUC) as our metrics. Results show that when the training set contains a single face manipulation algorithm, CADDM tends to overfit a specific type of artifacts and performs worse than the baseline with data augmentations. However, as the number of face manipulation algorithms increases, the training set contains more diverse artifacts. In this way, our CADDM learns common artifact features and significantly performs better on the cross-dataset evaluation.

6.3 More about Effect of Different Backbones

To show that reducing the influence of IIL helps our model perform well on the cross-dataset evaluations, we further explore the effect of different backbones for CADDM to demonstrate the broad applicability of our method. We use Frame-level AUC (FAUC) and Video-level AUC (VAUC) as our metrics. Results in
To show that reducing the influence of IIL helps our model achieve great performances on the cross-dataset evaluations, we conduct experiments on different backbones to show the effectiveness of our method. Here DA denotes the Data Augmentations. Each model is trained by FF++ [44] and tested on FF++ [44], Celeb-DF [29], and DFDC-V2 [10]. We use Frame-level AUC (FAUC) and Video-level AUC (VAUC) as our metrics. Results show that applying our method to different backbones brings a significant improvement on cross-dataset evaluations, which demonstrates the broad applicability of our method.

| Model          | In-dataset Evaluation | Cross-dataset Evaluation |       |       |
|----------------|-----------------------|--------------------------|-------|-------|
|                | FF++ [44]             | Celeb-DF [29]            | DFDC-V2 [10] |       |
| ResNet-18      | ×                     | ×                        | 99.19 | 99.78 |
|                |                       |                          | 59.78 | 65.82 |
|                |                       |                          | 51.34 | 52.23 |
|                | ✓                     | ✓                        | 99.39 | 99.79 |
|                | ✓                     | ✓                        | 99.36 (↑ 0.03) | 99.77 (↑ 0.02) |
|                | ×                     | ×                        | 69.22 | 77.56 |
|                |                       |                          | 60.62 | 63.25 |
| ResNet-34      | ✓                     | ✓                        | 99.41 | 99.74 |
|                | ✓                     | ✓                        | 99.33 (↑ 0.09) | 99.70 (↑ 0.18) |
|                | ×                     | ×                        | 58.69 | 64.05 |
|                |                       |                          | 48.69 | 48.73 |
| ResNet-50      | ✓                     | ✓                        | 99.47 | 99.83 |
|                | ✓                     | ✓                        | 99.46 (↑ 0.01) | 99.76 (↑ 0.07) |
|                | ×                     | ×                        | 61.87 | 69.63 |
|                |                       |                          | 48.84 | 49.49 |
| Xception       | ✓                     | ✓                        | 99.27 | 99.77 |
|                | ✓                     | ✓                        | 99.37 (↑ 0.10) | 99.89 (↑ 0.10) |
|                | ×                     | ×                        | 56.96 | 58.47 |
|                |                       |                          | 46.17 | 45.66 |
| Efficient-b3   | ✓                     | ✓                        | 99.16 | 99.81 |
|                | ✓                     | ✓                        | 99.44 (↑ 0.03) | 99.77 (↑ 0.04) |
|                | ×                     | ×                        | 73.39 | 84.24 |
|                |                       |                          | 64.54 | 68.96 |

Table 6 show that CADDM achieves great performances on the cross-dataset evaluation when using different backbones. On average, our method achieves 5.88% Video-AUC improvement on DFDC-V2 [10] and 9.30% Video-AUC improvement on Celeb-DF [29], compared to the baseline. On the in-dataset evaluation, our method maintains similar performance to the baseline, decreasing only 0.04% Video-AUC on FF++ [44] on average. Such results show that reducing the influence of IIL helps our model achieve great performances on cross-dataset evaluations among various backbones, which sheds new light on the model generalization for deepfake detection.

### 6.4 More details about Loss Function

In this section, we introduce more details about the loss function in the paper. To indicate images based on local artifact areas, the loss of CADDM is designed as follows:

\[
L_{caddm} = \beta L_{det} + L_{cls}
\]
where $\beta$ is a hyper-parameter; $L_{cls}$ denotes the cross entropy loss to measure the accuracy of the final prediction (i.e. whether the image is manipulated); $L_{det}$ denotes the artifact detection loss similar to the loss of SSD [32].

To guide the Multi-scale Detection Module in CADDM to localize the artifact areas and classify multi-scale anchors, $L_{det}$ is designed as follows:

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

(6)

where $\alpha$ denotes a positive weight. $L_{conf}(x, c)$ denotes the confidence loss, which is a binary cross-entropy loss to classify each anchor (i.e. the fake or genuine anchor).

$$L_{conf}(x, c) = -\sum_{i \in Pos} x_{ij} \log \hat{c}^p_i - \sum_{i \in Neg} (1 - x_{ij}) \log \hat{c}^n_i$$

(7)

$$\hat{c}_i^p = \frac{\exp(c_i^p)}{\sum_{p \in \{pos, neg\}} \exp(c_i^p)}$$

(8)

where $x_{ij} \in \{1, 0\}$ denotes the indicator for matching the $i$-th default anchor to the $j$-th ground truth of artifact area. The $i$-th anchor box is regarded as a positive sample (i.e. $x_{ij} = 1$) when the Intersection over Union (IoU) between the anchor box and the $j$-th ground truth of artifact areas is greater than 0.9. $c_i$ denotes the class confidence. $L_{loc}(x, l, g)$ is a Smooth L1 loss [14] between ADM predictions ($l$) and artifact area positions ($g$). In concrete, we regress the offsets for the center ($cx, cy$) of the default anchor ($d$) and for its width ($w$) and height ($h$).

$$L_{loc}(x, l, g) = \sum_{i \in Pos} \sum_{m \in \{cx, cy, w, h\}} x_{ij} \text{smooth}_L1(l_i^m - \hat{g}_j^m)$$

(9)

$$\hat{g}_j^{cx} = (g_j^{cx} - d_i^{cx}) / d_i^w \quad \hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy}) / d_i^h$$

(10)

$$\hat{g}_j^w = \log(g_j^w / d_i^w) \quad \hat{g}_j^h = \log(g_j^h / d_i^h)$$

(11)

6.5 More Visual Results

More Visual Results of MFS Figure 10 and Figure 11 show more visual results for global swap and partial swap of Multi-scale Facial Swap (MFS) respectively.

For global swap in Figure 10, we randomly replace the whole faces of fake images with faces of source images or the other way around with a certain probability. When replacing faces of fake images with faces of sources images, the newly generated MFS images contain similar identity information with source images. In this way, deepfake detection models can learn subtle differences between fake images (i.e. MFS images) and genuine images with less influence of identity information on images, since they are of almost the same identity. When replacing faces of source images with faces of fake images, the newly generated
Fig. 10. More visual results for global swap of MFS on various facial manipulation algorithms. For global swap, we either replace the whole faces of fake images with faces of source images or replace the whole faces of source images with faces of fake images with a certain probability. The column of DSSIM indicates the differences between source images and MFS images.

MFS images contain more blending artifacts than fake images. As demonstrated in [26], such images are also helpful to improve the generalization of deepfake detection models.

For partial swap in Figure 11, MFS exchanges the most significant manipulated areas between source images and fake images, with bounding boxes of different sizes. When using small bounding boxes, the newly generated MFS images also share similar identity information with source images, which helps to reduce the influence of IIL. Moreover, MFS provides the ground truth of local artifact areas, which helps CADDM to concentrate more on the most-likely forged areas, with less influence of other forgery-irrelevant areas on images. Results in Table 1 show that MFS successfully improves the generalization of deepfake detection models by reducing the influence of IIL.

More Visual Results of CADDM Figure 12 shows more visual results of CADDM on different face manipulation algorithms. In practice, there usually exist multiple predicted fake anchors on fake images. During the inference, we use NMS in [32] to select the anchor result with the highest score as the final prediction. When the score is less than a certain threshold, no anchor box will be predicted, as is the case with real images. Results on fake images in Figure 12 show that artifacts on eyes and mouths are more obvious to CADDM.

In summary, our proposed CADDM not only indicates fake images but also regresses the bounding box of local artifact areas. Our method provides new
insights into the model generalization on deepfake detection, showing that hand-crafted deepfake detectors are not indispensable.
Fig. 11. More visual results on partial swap of MFS on various facial manipulation algorithms. For partial swap, we exchange the most significant manipulated areas between fake images and source images, with different sizes of bounding boxes (i.e. 20x40, 40x80, 80x120, 120x160). Specifically, we replace the chosen areas of source images with the corresponding areas of fake images. The exchanged areas between source images and fake images are marked with rectangles. The column of DSSIM indicates the differences between source images and MFS images.
Fig. 12. More visual results on various facial manipulation algorithms. Our CADDM not only indicates fake images but also regresses the bounding box of the artifact area, showing that hand-crafted deepfake detectors are not indispensable when tackling the task of deepfake detection.