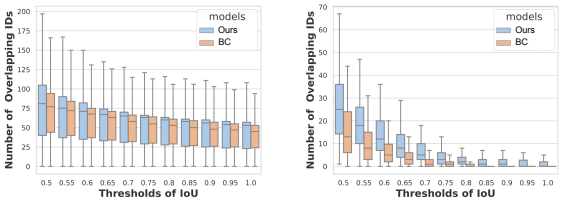
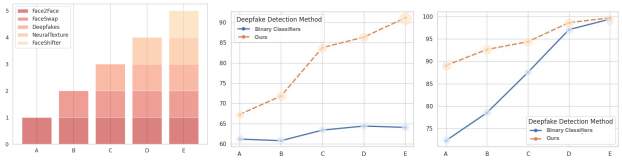


(a) FF++ (b) Celeb-DF

Figure 7. Quantitative evaluation on the influence of Implicit Identity Leakage. Here BC stands for binary classifiers. Results show that compared with the binary classifier, features of different identities in our model overlapped with each other more. Such results were more prominent on cross-dataset evaluations (Figure 7b).



(a) Sub-dataset division (b) AUC on Celeb-DF (c) AUC on FF++

Figure 8. **Verifying that our model learned various artifact features in a data-driven manner.** We split five training sets following Figure 8a to train the model and test the performance respectively. When the number of forgery methods in the training set increased, our model had higher performance in both in-dataset and cross-dataset evaluation.

A. Quantitative Analysis of the Implicit Identity Leakage

In this section, we conduct an experiment to evaluate the influence of Implicit Identity Leakage quantitatively.

A.1. 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 sampled 5 images at an equal interval from each video for all identities in the dataset. Then, we used principal component analysis (PCA) to project the features of images into 2D space. For each identity, we considered the rectangle area of its projected features as the region of the identity. Two identities were considered to overlap with each other when the Intersection over Union (IoU) of their regions was no less than the threshold.

We evaluated the influence of the Implicit Identity Leakage on the binary classifier and our model quantitatively. Both the binary classifier and our model were trained on FF++ [9] and tested on FF++ [9] and Celeb-DF [6]. As shown in Figure 7, each box-and-whisker denotes the distributions for the number of overlapping IDs across all identities in the dataset, given different values of the threshold. Results show that features of different identities in our model overlapped with each other more, especially on the cross-dataset evaluation (Figure 7b). Such results demonstrate that our model reduced the influence of Implicit Identity Leakage.

B. Verification of Learning Various Artifacts

In this part, we conduct an experiment to verify that when reducing the influence of Implicit Identity Leakage, our model can capture various artifact features from different manipulation algorithms in a data-driven manner. Thanks to the division of training data in the FF++ dataset according to different face forgery methods, we split five sub-training sets following Figure 8a, and further explored the relationship between the number of manipulation algorithms and the model performance in the in-dataset and cross-dataset evaluation. Five different face-swap methods are corresponding to 5 sub-datasets in FF++. Both our model and the binary classifier used ResNet-34 [4] as the backbone and were tested on FF++ and Celeb-DF respectively.

As shown in Figure 8b, due to the Implicit Identity Leakage, even if the number of manipulation algorithms in the training set increased, the binary classifier still maintained poor performance on the cross-dataset evaluation. In contrast, artifact features captured by our model were more generalized as the number of manipulation algorithms increased. Such results demonstrate that our model learned various artifact features from forgeries in a data-driven manner. Moreover, our approach also achieved higher performance in the in-dataset evaluation with less training data. Figure 8c indicates that when only using the Face2Face [10] sub-dataset as the training set, compared with the binary classifier, our approach got 17% AUC improvements in FF++.

C. Effect of Different Backbones

In this section, we further explored the effect of different backbones for our model to demonstrate the broad applicability of our method. Each model was trained by FF++ and tested on FF++, Celeb-DF, and DFDC-V2 [2]. We used Frame-level AUC (FAUC) and Video-level AUC (VAUC) as our metrics. Results in Table 7 show that our

Model	DA	Ours	In-dataset Evaluation		Cross-dataset Evaluation			
			FF++		Celeb-DF		DFDC-V2	
			FAUC	*VAUC*	*FAUC*	*VAUC*	*FAUC*	*VAUC*
ResNet-18	×	×	99.19	99.78	59.78	65.82	51.34	52.23
	✓	×	99.39	99.79	69.22	77.56	60.62	63.25
	✓	✓	99.36 (↓ 0.03)	99.77 (↓ 0.02)	76.73 (↑ 7.51)	89.07 (↑ 11.51)	64.54 (↑ 3.92)	69.22 (↑ 5.97)
ResNet-34	×	×	99.42	99.88	58.69	64.05	48.69	48.73
	✓	×	99.41	99.74	71.45	80.07	59.41	62.46
	✓	✓	99.33 (↓ 0.09)	99.70 (↓ 0.18)	79.56 (↑ 8.11)	91.15 (↑ 11.08)	67.04 (↑ 7.63)	71.49 (↑ 9.03)
ResNet-50	×	×	99.47	99.83	61.87	69.63	48.84	49.49
	✓	×	99.32	99.70	68.38	75.92	60.33	62.67
	✓	✓	99.46 (↓ 0.01)	99.76 (↓ 0.07)	76.78 (↑ 8.4)	88.16 (↑ 12.24)	65.52 (↑ 5.19)	69.80 (↑ 7.13)
Xception	×	×	99.27	99.77	56.96	58.47	46.17	45.66
	✓	×	99.27	99.79	71.98	81.53	60.56	64.88
	✓	✓	99.37 (↑ 0.10)	99.89 (↑ 0.10)	74.92 (↑ 2.94)	86.69 (↑ 5.16)	63.74 (↑ 3.18)	67.52 (↑ 2.64)
Efficient-b3	×	×	99.16	99.75	57.49	59.97	52.74	54.12
	✓	×	99.44	99.81	73.39	84.24	64.54	68.96
	✓	✓	99.45 (↑ 0.01)	99.78 (↓ 0.03)	83.02 (↑ 9.63)	93.08 (↑ 8.84)	68.44 (↑ 3.90)	73.74 (↑ 4.78)

Table 7. **Effect of different backbones.** Here DA denotes the Data Augmentations. Results show that applying our method to different backbones brought a significant improvement in cross-dataset evaluations, which demonstrates the broad applicability of our method.

model achieved great performances on the cross-dataset evaluation when using different backbones. On average, our method achieved 5.91% Video-AUC improvement on DFDC-V2 and 9.77% Video-AUC improvement on Celeb-DF, compared to the baseline. On the in-dataset evaluation, our method maintained similar performance to the baseline, decreasing only 0.04% Video-AUC on FF++ on average. Such results show that reducing the influence of the Implicit Identity Leakage helped our model achieve great performances on cross-dataset evaluations among various backbones, which sheds new light on the model generalization for the task of deepfake detection.

D. 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 our model is designed as follows:

$$L = \beta L_{det} + L_{cls}. \tag{1}$$

where β 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 [3, 7, 8].

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

$$L_{det} = \frac{1}{N}(L_{conf}(x, c) + \alpha L_{loc}(x, l, g)). \tag{2}$$

where α 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).

$$\begin{aligned} L_{conf}(x, c) &= -\sum_{i \in Pos} x_{ij} log \hat{c}_i^{pos} - \sum_{i \in Neg} (1 - x_{ij}) log \hat{c}_i^{neg} \\ \hat{c}_i^p &= \frac{exp(c_i^p)}{\sum_{p \in \{pos, neg\}} exp(c_i^p)} \end{aligned} \tag{3}$$

where $x_{ij} \in \{1, 0\}$ denotes the indicator for matching the i-th default anchor to the j-th ground truth of the 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 [3] 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).

$$\begin{aligned} L_{loc}(x, l, g) &= \sum_{i \in Pos} \sum_{m \in \{cx, cy, w, h\}} x_{ij} smooth_{L1}(l_i^m - \hat{g}_j^m) \\ \hat{g}_j^{cx} &= (g_j^{cx} - d_i^{cx})/d_i^w \quad \hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h \\ \hat{g}_j^w &= log(\frac{g_j^w}{d_i^w}) \quad \hat{g}_j^h = log(\frac{g_j^h}{d_i^h}) \end{aligned} \tag{4}$$

E. More Visual Results of MFS

Figure 9 and Figure 10 show more visual results for the global swap and the partial swap of Multi-scale Facial Swap (MFS) respectively.

For the global swap in Figure 9, 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 the faces of fake images with faces of source images, the newly generated MFS images contain similar identity information to 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 the faces of source images with faces of fake images, the newly generated MFS images contain more blending artifacts than fake images. As demonstrated in [5], such images are also helpful to improve the generalization of deepfake detection models.

For the partial swap in Figure 10, 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 Implicit Identity Leakage. Moreover, MFS provides the ground truth of local artifact areas, which helps our model to concentrate more on the most-likely forged areas, with less influence of other forgery-irrelevant areas on images. Results in Table 2 of the paper show that MFS successfully improved the generalization of deepfake detection models by reducing the influence of Implicit Identity Leakage.

F. Discussion for the Utility of ID Representations

Recent method [1] showed the utility of ID representation is effective for the task of deepfake detection, which, nevertheless, is not in conflict with our study. [1] trained their model on real videos of different identities only, and tested the performance based on a distance metric between the test video and a set of genuine videos of the target ID prepared in advance. Such a protocol could also be considered as the reduction of the influence of Implicit Identity Leakage. Firstly, the training process required real videos only. The misguidance of ID representation between real and fake videos we discussed in this paper is naturally eliminated, since all identities are considered as genuine identities. Besides, during the inference process, they calculated the distance metric between the test video and the preset genuine videos of the target ID. In this scenario, the ID representation of the two videos is aligned, which also weakens the distraction of ID representation.

References

[1] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15108–15117, 2021. 3

[2] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv e-prints*, pages arXiv–2006, 2020. 1

[3] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[4] 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. 1

[5] 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. 3

[6] 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. 1

[7] 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. 2

[8] 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. 2

[9] 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. 1

[10] 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

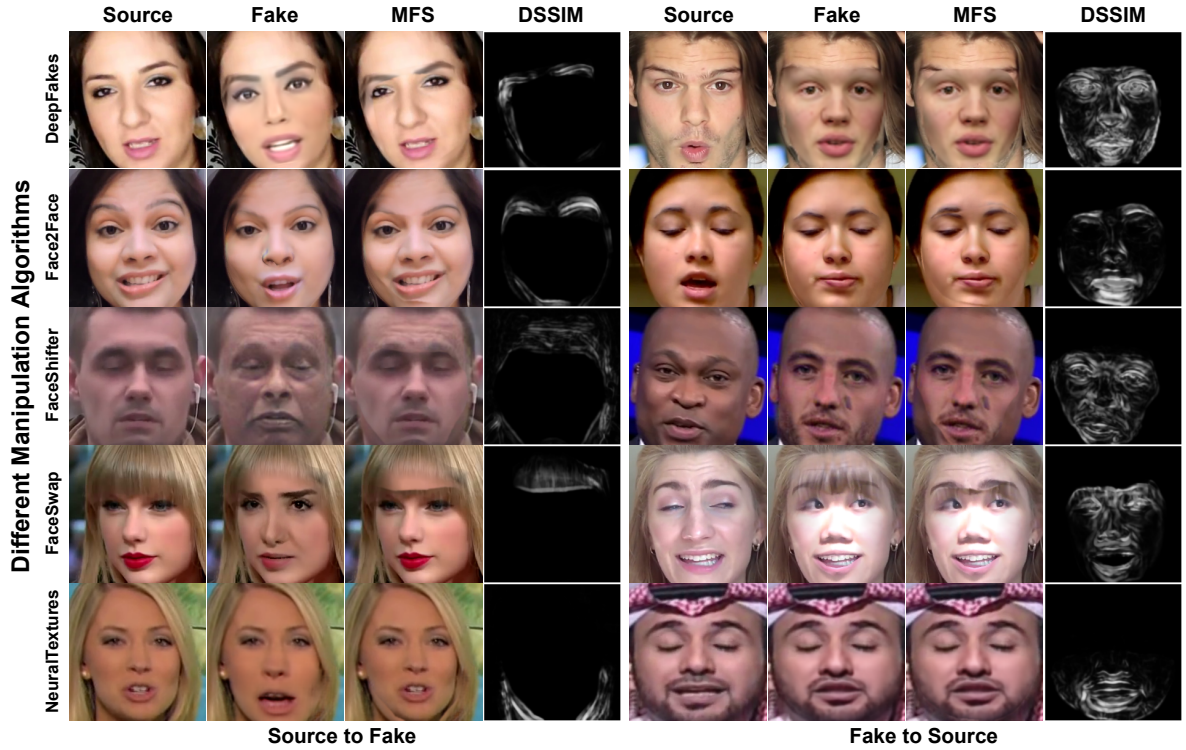


Figure 9. **More visual results for the global swap of MFS on various facial manipulation algorithms.** For the global swap, we either replace the whole faces of fake images with faces of source images or replace the whole faces of source images with faces of fake images with a certain probability. The column of DSSIM indicates the differences between source images and MFS images.

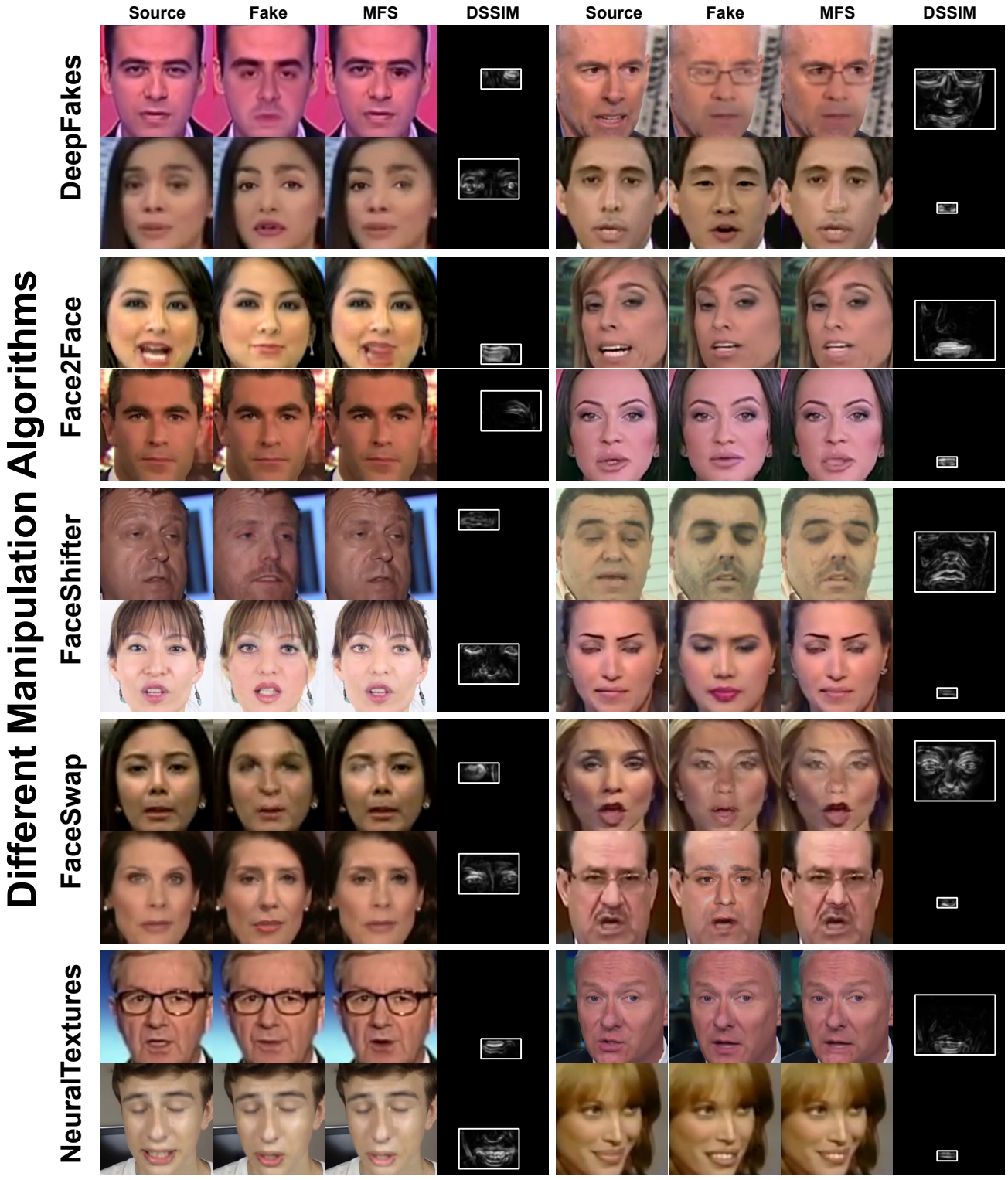


Figure 10. **More visual results on the partial swap of MFS on various facial manipulation algorithms.** For the partial swap, we exchange the most significant manipulated areas between fake images and source images, with different sizes of bounding boxes (*i.e.* 20x40, 40x80, 80x120, 120x160). Specifically, 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.