Implicit Identity Driven Deepfake Face Swapping Detection

Baojin Huang[†], Zhongyuan Wang[†,*], Jifan Yang[†], Jiaxin Ai[†], Qin Zou[†], Qian Wang[‡], Dengpan Ye[‡]
[†]NERCMS, School of Computer Science, Wuhan University
[‡]School of Cyber Science and Engineering, Wuhan University

Abstract

In this paper, we consider the face swapping detection from the perspective of face identity. Face swapping aims to replace the target face with the source face and generate the fake face that the human cannot distinguish between real and fake. We argue that the fake face contains the explicit identity and implicit identity, which respectively corresponds to the identity of the source face and target face during face swapping. Note that the explicit identities of faces can be extracted by regular face recognizers. Particularly, the implicit identity of real face is consistent with the its explicit identity. Thus the difference between explicit and implicit identity of face facilitates face swapping detection. Following this idea, we propose a novel implicit identity driven framework for face swapping detection. Specifically, we design an explicit identity contrast (EIC) loss and an implicit identity exploration (IIE) loss, which supervises a CNN backbone to embed face images into the implicit identity space. Under the guidance of EIC, real samples are pulled closer to their explicit identities, while fake samples are pushed away from their explicit identities. Moreover, IIE is derived from the margin-based classification loss function, which encourages the fake faces with known target identities to enjoy intra-class compactness and inter-class diversity. Extensive experiments and visualizations on several datasets demonstrate the generalization of our method against the state-of-the-art counterparts.

1. Introduction

The development of deep learning has promoted the continuous progress of face forgery technology [5, 16, 48]. Especially for face swapping, it can replace the target face with the source face to generate a fake face that is not distinguishable by the human eyes. With this technology, attackers can easily forge high-quality videos of public celebrities and political figures to achieve illegal political or commercial purposes. To alleviate the abuse of face swapping, it is urgent to exploit corresponding detection methods.

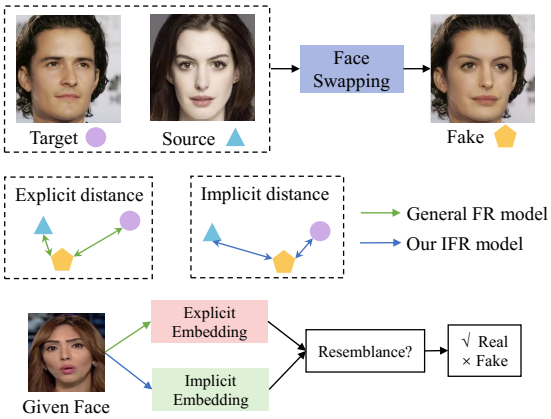


Figure 1. Motivation of our approach. The target face is replaced by the source face through face swapping to generate a fake face. In appearance, the fake face looks like the source face instead of the target face. We resort the general face recognition (FR) model CosFace [51] to obtain the explicit distance of these faces. Particularly, since the fake face is synthesized from the source face and the target face, we aim to explore a implicit face recognition (IFR) model that can mine the corresponding target face identity based on the fake face. With the similarity between explicit and implicit embeddings of the given face, we can significantly distinguish it as real and fake, which facilitates forgery detection.

Early researches [1, 10, 37, 42] usually treat face swap detection as a binary image classification task. Specifically, face images are fed into an existing deep convolutional neural network (CNN) and then classified as real and fake. Such methods can learn the data distribution of the training set, resulting in considerable performance in intra-domian tests. However, the simple classification guidance cannot incorporate the connotation of face swapping, thus the deep network lacks the understanding of forgery [50]. Recent works are devoted to exploring specific forgery patterns, such as noise analysis [27], local regions [7, 53] and frequency information [19, 41]. In this way, fake traces in fake faces can be better detected. Albeit gaining the benefits, they still revolve around certain manipulation methods and

*Corresponding author.

are not conducive to generalize well to unseen real-world scenarios. Therefore, in practice, many emerging forgery methods as well as unknown environmental factors bring serious performance degradation to existing face swapping detection methods.

To address the above issues, we consider the face swapping detection from the perspective of face identity. As shown in Figure 1, face swapping aims to replace the target face with the source face, further generating a fake face that is even indistinguishable for human eyes. Here, we introduce two new concepts for fake faces, including *explicit* identity and *implicit* identity. Specifically, the explicit identity represents what the fake face looks like, that is, the source face identity. Thus, the explicit distance between the fake face and the real face can be measured by existing general face recognition models [11, 22, 51]. For implicit identity, we believe that the fake face comes from the source face and the target face. Although it looks like the source face, it might contain more or less target face identity information. We call this potential target face information the implicit identity of the fake face. It is worth noting that the implicit identities of the real face are consistent with its explicit identities. Therefore, given a face image, we embed it into the explicit and implicit identity feature spaces, respectively. The distance between its explicit and implicit features is taken as the basis for judging real and fake. Provided the distance is very close, the given image is real, otherwise it is a fake image.

With the above considerations in mind, in this paper, we propose a novel implicit identity driven (IID) framework to detect face swapping. Our key motivation is to explore the implicit identity of the face, which guides deep networks to make more reasonable detection results. To this end, we first employ the generic face recognition model to obtain its explicit identity embedding. Subsequently, we propose the explicit identity contrast (EIC) loss and the implicit identity exploration (IIE) loss to supervise the off-the-shelf CNN backbone, aiming to transform the face image into the implicit identity feature space. Specifically, under the guidance of EIC, real samples are pulled closer to their explicit identities, while fake samples are pushed away from their explicit identities. In this way, the difference between the real and fake samples in the feature space is enlarged. It is worth noting that the real sample feature at this time denotes its implicit identity (close to the explicit identity). Moreover, to further explore the implicit identity of the fake sample, we label the identity of the fake face with its corresponding target face identity. Particularly, for those fake faces whose target faces are unknown but come from the same video, we label their identities as extra and identical to ensure identity consistency. Inspired by general face recognition algorithms [11, 51], our proposed IIE is derived from the margin-based classification loss function, which guides fake faces with known target identities to have small intra-class distances and large inter-class distances. Besides, fake faces with unknown target identities originating from the same video have consistent identity embeddings. Thereby, implicit identities of fake faces can be mined comprehensively. Finally, we use the difference between the implicit identity and explicit identity of the face as the basis for distinguishing real and fake.

In brief, the main contributions are as follows:

- From a completely new perspective, we propose the implicit identity driven framework for face swapping detection, which explores the implicit identity of fake faces. This enhances the deep network to distinguish fake faces with unknown manipulations.
- We specially design explicit identity contrast (EIC) loss and the implicit identity exploration (IIE) loss. EIC aims to pull real samples closer to their explicit identities and push fake samples away from their explicit identities. IIE is margin-based and guides fake faces with known target identities to have small intra-class distances and large inter-class distances.
- Extensive experiments and visualizations demonstrate the superiority of our method over the state-of-the-art approaches.

2. Related Work

2.1. Face Swapping

Recent face swapping methods [2, 17, 23, 34, 35, 39] benefit from advances in deep learning. At the outset, researchers [23] view face swapping as a style transfer problem. Under the guidance of face landmarks, the CNN can transfer a face image to the style of another face image with one specific identity. Since then, the classic DeepFakes [17] proposes an encoder-decoder face swapping framework. Once trained, it can swap faces between the two specified identities but cannot generalize to others. On this basis, several methods [2, 34, 35] combining latent representations have emerged. They extract identity features from the source face and attribute features from the target face. However, the expression of the target face is often not preserved in the output of the decoder. The trickiest problem with the above approaches is requiring training on the pairs of faces to be swapped, which is unfriendly in practice. To overcome the above limitation, Nirkin *et al.* [39] propose a novel recurrent neural network based approach for face reenactment, which can be applied to a single image or a video sequence. Recent reconstruction-based face swapping methods [6, 15, 26] with GANs have also shown success. They are subject agnostic and able to generate high-quality and realistic fake images. Overall, existing learning-based face swapping methods claim to decouple the identity

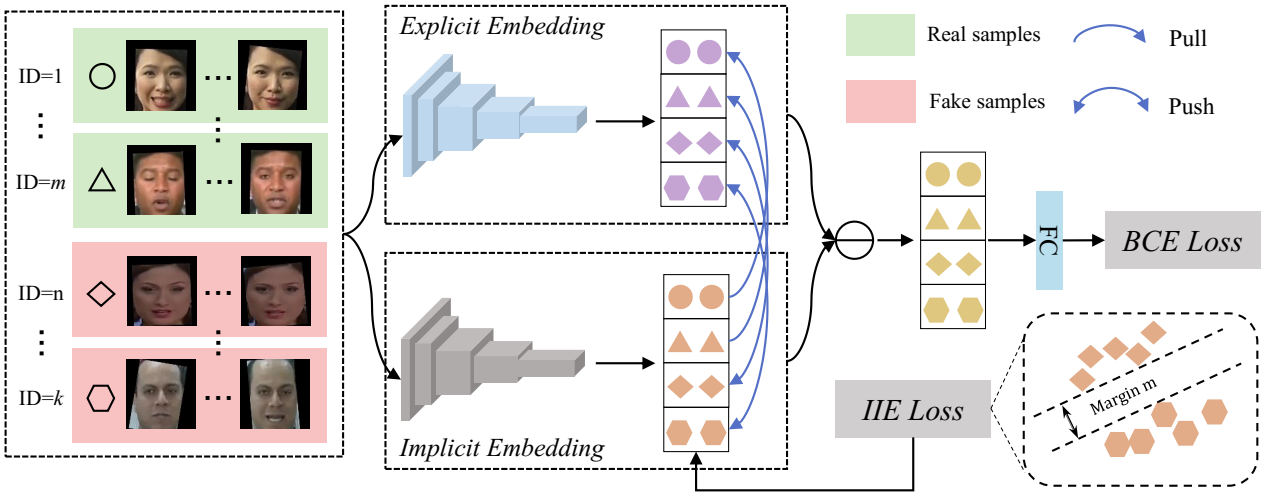


Figure 2. The outline of our proposed implicit identity driven framework for deepfake face swapping detection. We hybridize real face samples (green boxes) and fake face samples (red boxes) as training set. It is worth noting that we label the identity of the fake face with the corresponding target face identity. During training, we employ the generic face recognition model to obtain the explicit identity embedding of training sample as a contrast. The implicit identity embedding extracted by the backbone is supervised by the implicit identity exploration (IIE) loss. Besides, real samples are pulled closer to their explicit identities, while fake samples are pushed away from their explicit identities. The difference between the implicit and explicit identities of face sample is guided by the binary cross-entropy (BCE) to output predicted confidence.

of the original face and assign it to the target face. However, there is no pure decoupling method, thus the fake face contains potential target face identity information. To this end, our method aims to explore this potential cue for face swapping detection.

2.2. Face Forgery Detection

Nowadays, many studies [1, 20, 27, 33, 33, 36, 46, 50, 56] are proposed to boost the performance of face forgery detection. Early works [1, 10, 37, 42] usually utilize existing image classification networks [8, 43] to transform cropped face images into feature vectors and perform binary classification. However, classification methods alone tend to overfit the training data and fail to explore the subtle differences between real and fake images. Therefore, a number of methods based on face forgery patterns have been proposed to discriminate between real and fake. Zhou *et al.* [54] present a two-stream deep network to detect fake faces by focusing on visual appearance and local noise in two branches, respectively. Zhao *et al.* [53] propose a multi-attentional network architecture to capture local discriminative features from multiple face attentive regions. Besides, frequency information [14, 25, 41] is also verified to provide clues for face forgery detection. Recent researches [3, 46, 55] increasingly tend to improve the generalization of detectors for unseen forgeries. Sun *et al.* [46] propose a dual contrastive learning (DCL) for general face forgery detection. Despite the improved performance, DCL mainly rely on the generation of paired images, which is usually unpredictable in practice. To further improve the generalization for the detection model, we consider the face swapping detection from the perspective of face identity. Moreover, we introduce the implicit identity driven method for general face swapping detection.

3. Proposed Method

In this section, we introduce our implicit identity driven (IID) framework for general face swapping detection, which consists of two main schemes, *i.e.*, explicit identity contrast (EIC) and the implicit identity exploration (IIE), as illustrated in Figure 2. EIC loss pulls real samples closer to their explicit identities, while pushing fake samples away from their explicit identities. As such, real samples converge to their implicit identities (same as explicit identities), and fake samples are mined for explicit identities irrelevant features. Moreover, to further clarify the implicit identity for fake samples, the IIE loss constrains the identity of the fake samples to be attributed to their corresponding target faces (implicit identities). Particularly, fake faces with unknown target face originating from the same video are embedded into the consistent identity space. Thereby, the difference between the implicit identity and explicit identity of the face is used as the basis for distinguishing real and fake. In the following, we will elaborate on the individual schemes.

3.1. Explicit Identity Contrast

Since the fake face is derived from the source and the target face, we argue that the fake face contains more or less the identity information of the target face. As such, we propose to use the explicit identity of the face as a contrast to enlarge the difference between the real and fake samples in the feature space. To be specific, given an aligned face image $x_i \in \mathbb{R}^{h\times w\times 3}$, we employ the generic face recognition model F_{ex} to obtain its explicit identity feature, denoted as $F_{ex}(x_i)$. Subsequently, we train a backbone as implicit identity embedding network F_{im}, which transform the input image x_i into the feature vector $F_{im}(x_i)$. Following the characteristics of our proposed implicit identity, the implicit identity of a real face needs to be consistent with its corresponding explicit identity, while the fake face is just the opposite. To this end, we adopt explicit identities as contrasts to initially guide the representation of implicit identities. The designed explicit identity contrast loss is

$$\mathcal{L}_{\text{eic}} = \frac{1}{N_F}\sum_{i\in F}\delta\left(F_{im}(x_i), F_{em}(x_i)\right) - \frac{1}{N_R}\sum_{i\in R}\delta\left(F_{im}(x_i), F_{em}(x_i)\right), \tag{1}$$

where R and F indicate the set of real and fake samples, respectively. N_R and N_F denote the number of real samples and fake samples, respectively. $\delta(\cdot,\cdot)$ represents the cosine similarity calculation function, which is defined as $\delta(u,v) = \frac{u}{\|u\|}\cdot\frac{v}{\|v\|}$.

Our proposed EIC loss works on fake and real face samples, respectively. On the one hand, it encourages fake samples to move away from their explicit identities in the implicit feature space. Because the implicit identity of the fake face corresponds to the target face rather than the source face (explicit identity). In this way, it is guaranteed that fake faces are extracted with explicit identity irrelevant features. On the other hand, real samples are guided towards their explicit identities in the implicit feature space. This conforms to the assumption of explicit and implicit identity consistency for real samples. Note that, the existing contrastive-learning based methods [3,4,46] usually directly act on real and fake samples to seek difference, while our proposed loss takes the explicit identity as a reference, which is more reasonable to explore the essential forgery clues.

3.2. Implicit Identity Exploration

The aforementioned EIC loss enlarges the difference between the real and fake samples in the feature space. At this point, the fake samples are only distinguished from their explicit identities in the feature space. To further clarify the implicit identities of fake samples, we design an implicit identity exploration loss, which uses the target face as a guide to refine the implicit identities of fake faces. In particular, for fake samples with unknown target faces and originating from the same video, we maintain their identity consistency to ensure that there is no large detection difference between the frames of the same video during face swapping detection.

Specifically, since the mainstream fake face datasets, such as FF++ [42], contain source images and fake images, we can further label the fake face with its target face identity y_i (implicit identity). Whereas the real face is labeled with its explicit identity. In other words, the fake face and its corresponding target face are labeled as a category and both as the training sample. We define real samples and fake samples with known implicit identities as a set $\mathcal{K}$. Given a face sample $x_i \in \mathcal{K}$, its extracted feature vector $F_{im}(x_i)$ is further normalized to $\frac{F_{im}(x_i)}{\|F_{im}(x_i)\|}$. Subsequently, our designed IIE loss closes the identity distance between the fake face and its corresponding target face, which can be derived as:

$$\mathcal{L}^{+}_{iie} = -\mathbb{E}_{x_i,y_i\sim\mathcal{K}}\left[\log\frac{e^{s(\cos(\theta y_i)-m)}}{e^{s(\cos(\theta_{y_i})-m)}+\sum_{j\neq y_i}e^{s\cos\theta_j}}\right]. \tag{2}$$

Here, θ_j represents the angle between normalized $F_{im}(x_i)$ and the normalized proxy of j-th identity on the hypersphere. s and m stand for feature rescale and margin hyperparameter, respectively. The margin can simultaneously enhance the intra-class compactness and inter-class discrepancy. Different from the popular face recognition loss CosFace [51] which sets a fixed margin, we assign different margin values to real and fake samples respectively. Specifically, the margin m_{real} for the real sample is set to a fixed value of 0.4. Particularly, we use the identity fitting progress of real samples to obtain a progressive margin for fake samples, calculated as

$$m_{fake} = \alpha\cdot\frac{1}{N_r}\sum_{i\in R_{mini}}\cos(\theta y_i), \tag{3}$$

where R_{mini} denotes the set of real samples for a mini-batch. N_r represents the number of samples in R_{mini}. α is a hyperparameter to limit the maximum value of the margin, which is empirically set to 0.5.

It can be observed that the m_{fake} at this time varies with the fit of the real samples. Therefore, in the early stage of model training, m_{fake} is so small that the real samples are concerned. After the real samples (target faces) are fully fitted, the margin for the fake samples starts to work. Furthermore, the fake face keeps approaching its corresponding target face. With this progressive learning strategy, the deep network first explores the implicit identity (same as the explicit identity) of the real face (target face), and then fits the fake face. Compared to fitting the implicit identities of real

and fake samples in one go, progressive learning makes it easier for the network to converge and achieve better performance.

To comprehensively cover the actual situation, we take the fake samples of unknown target faces into consideration. The key idea is to maintain the identity consistency of frames in the same fake video. Specifically, the set of unknown fake samples is defined as $\mathcal{U}$. For a fake sample $x_i \in \mathcal{U}$, we label its unknown implicit identity as y_i^*. Meanwhile, other frames from the same video as x_i also have the same implicit identity. We embed x_i into the feature space $F_{im}(x_i) \in \mathbb{R}^D$ by the implicit identity embedding network, where D is the feature dimension. We establish a lookup table $V \in \mathbb{R}^{D \times Q}$ to store the normalized features of all the unknown implicit identities. During the implicit identity embedding network forward propagation, we calculate the distance between sample x_i and unknown identities in the lookup table by cosine similarity, denoted as $V^T F_{im}(x_i)$. During backward, we update the y_i^*-th column in the lookup table by $v_{y_i^*} \leftarrow \beta v_{y_i^*} + (1-\beta) F_{im}(x_i)$, where $\beta \in [0, 1]$. Moreover, we define the probability that sample x_i is classified as y_i^* by the Softmax function and maximize the expected log-likelihood

$$\mathcal{L}_{iie}^{-} = -\mathbb{E}_{x_i, y_i^* \sim \mathcal{U}} \left[\log \frac{e^{\left(v_{y_i^*}^T F_{im}(x_i)/\tau\right)}}{\sum_{j=1}^{Q} e^{\left(v_j^T F_{im}(x_i)/\tau\right)}} \right]. \quad (4)$$

The higher temperature τ leads to softer probability distribution.

It can be seen that our $\mathcal{L}_{iie}^{-}$ effectively compares the mini-batch unknown fake sample with all the unknown implicit identities, driving the identity consistency between different frames of the same video. The overall IIE loss $\mathcal{L}_{iie}$ can be derived as

$$\mathcal{L}_{iie} = \mathcal{L}_{iie}^{+} + \mathcal{L}_{iie}^{-}. \quad (5)$$

3.3. Overall Loss Function

With the difference between the implicit and explicit identities of face samples, we insert a fully connected classifier to perform classification and make full use of label information. Therefore, the overall loss function $\mathcal{L}$ of the IID framework includes the EIC and the IIE loss, as well as the binary cross-entropy loss:

$$\mathcal{L} = \mathcal{L}_{bce} + \lambda_1 \mathcal{L}_{eic} + \lambda_2 \mathcal{L}_{iie}, \quad (6)$$

where λ_1 and λ_2 are weight parameters for trading off the losses.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our proposed method on five challenging datasets, including FaceForensics++ (FF++) [42], Celeb-DF [29], FaceShifter [26], DFD [38] and DFDC [13]. **FF++** is the most widely used forgery dataset, covering 720 videos for training and 280 videos for validation or testing. It contains four manipulation methods, including identity swapping methods (DeepFakes [17], FaceSwap [18]) and expression swapping methods (Face2Face [49], and NeuralTexture [48]), which is suitable for evaluating the generalization of the model. Note that Face2Face and NeuralTexture are based on expression swapping rather than identity swapping, thus in the following experiments, we only use IIE constraints instead of EIC constraints for the samples in Face2Face and NeuralTexture. Particularly, there are two types of video quality in FF++, including high quality (C23) and low quality (C40). **Celeb-DF** is generated by face swapping for 59 pairs of subjects, it contains 590 real videos and 5,639 high-quality fake videos. **FaceShifter** is a new forgery dataset obtained by applying the FaceShifter [26] manipulation method to the original video of FF++, which is more realistic and more difficult to detect real and fake. **DFD** is a Deepfake based dataset that has 363 real videos and 3,068 fake videos. **DFDC** is currently the largest publicly available face swapping video dataset, containing 1,133 real videos and 4,080 fake videos for testing. It is very challenging for existing forgery detection due to the diverse and unknown manipulation methods. For all datasets, we randomly select 50 frames by FFmpeg from each video as training and testing. Particularly, we adopt open-source RetinaFace [12] to detect and align faces from raw images. In this way, all face images are cropped and normalized to 224×224.

Evaluation Metrics. We employ common metrics to evaluate our method, including Area Under the Receiver Operating Characteristic Curve (AUC), Equal Error Rate (EER) and Accuracy (ACC).

Implementation Details. Our proposed IID method is implemented by Pytorch deep learning framework [40], with the batch size of 64 on two NVIDIA GTX 3090 GPUs. To improve the robustness of the model, we perform data augmentation such as flipping on the training set. The initial learning rate is set to 0.1 and divided by 10 at the 8-th and 14-th epochs. The entire deep network is optimized by the SGD with momentum 0.9 and weight decay 5e-4. Moreover, we use CosFace [51] trained on the WebFace dataset to extract face explicit identity features during training. The implicit identity embedding network is based on ResNet18 [21]. The temperature parameter τ in Equation 4 is set to 0.1. Besides, λ_1 and λ_2 in Equation 6 are empirically set to 0.05 and 0.1, respectively.

Model	$\mathcal{L}_{eic}$	$\mathcal{L}_{iie}$	Celeb-DF		DFDC	
			ACC (%)	AUC (%)	ACC (%)	AUC (%)
A			70.34	74.09	69.85	72.65
B	✓		77.76	82.24	76.39	78.80
C		✓	76.40	81.46	74.95	77.22
D	✓	✓	**79.16**	**83.80**	**79.37**	**81.23**

Table 1. Effectiveness of the proposed constraints in our method on the Celeb-DF and DFDC datasets. Specifically, $\mathcal{L}_{eic}$ and $\mathcal{L}_{iie}$ denote the EIC loss and IIE loss, respectively.

4.2. Ablation Study

Since our proposed framework is composed of several collaborative components, including the EIC loss and IIE loss, we conduct ablation experiments on Celeb-DF and DFDC datasets to verify the effects of these strategies. Specifically, we first construct the baseline model A without the EIC and IIE, which is actually a simple binary classification model. Subsequently, several variants are designed as: 1) baseline with the EIC, 2) baseline with the IIE, 3) baseline with the EIC and IIE.

The quantitative results on Celeb-DF and DFDC are reported in Table 1. Compared with the model A, model B achieves 7.42%, 6.54% ACC and 8.15%, 6.15% AUC gains on Celeb-DF and DFDC, respectively. This is attributed to EIC taking the explicit identities of real and fake samples as clues rather than simple classification. It is feasible for real samples to be close to their explicit identities and fake samples away from their explicit identities. Particularly, the drop in the accuracy of model C demonstrates that the supervision of IIE is meaningless without EIC. Because IIE binds the explicit and implicit identities of real samples together, without it the implicit identities of real samples are unknown. IIE aims to explore the implicit identities of fake faces and therefore cannot distinguish the real and fake well alone. The best performance is achieved when combining all the proposed constraints with 79.16%, 83.80% ACC and 79.37%, 81.23% AUC on Celeb-DF and DFDC, respectively.

4.3. Quantitative Results

Cross-dataset evaluation. To verify the generalizability of our proposed IID for cross-dataset, we conduct comprehensive experiments on representative datasets. Specifically, the models are trained on the FF++(C23) and evaluated on the Celeb-DF, DFD and DFDC, respectively. Besides, we select the classic and recent state-of-the-art methods for comparison, including Xception [42], Face X-ray [27], F3-Net [52], DCL [46] and UIA-ViT [55], etc.

Quantitative evaluation results of the above models are tabulated in Table 2. From the table, we can see that our proposed IID generally outperforms all counterparts on unseen test data, even achieving significant improvements on some datasets. For instance, the AUC scores of previous methods drop significantly on the unknown dataset DFDC. In contrast, IID reaches an AUC of 81.23%, which exceeds DCL [46] by 4.52%. The gains mainly benefits our proposed IID framework, which learns the implicit identities of real and fake samples under the guidance of EIC and IIE. It is worth noting that our model is not the most superior on the Intra-testing dataset (FF++). That is because our IID focuses more on exploring generalization differences between real and fake samples rather than simply fitting the training data distribution.

We further conduct a low-quality cross-dataset experiment by training on FF++(C40) and testing on Deepfakes class and Celeb-DF. We compare our model with state-of-the-art approaches in Table 3. Similar to high-quality cross-datasets, our IID achieves sub-superior performance on intra-testing, but outperforms by 4.69% compared with the recent ITA [55] on Celeb-DF.

Cross-manipulation evaluation. We further conduct experiments across manipulation methods to further explore the generalization ability of the model for different manipulation methods. Specifically, we choose the DeepFakes (DF) and FaceSwap (FS) methods of FF++(C23), and the FaceShifter (FST) dataset, which have the same face swapping objects. The model is trained on one of the datasets and tested on the other two.

As tabulated in Table 4, our method generally outperforms competitors in terms of mean AUC on unseen manipulation types. Specifically for our model trained on DF and tested on FST, it achieves an AUC gain of 5.04% versus DCL. In contrast, DF requires training on the pairs of faces to be swapped while FST can arbitrarily swap faces for a single face image. In principle they are extremely different manipulation methods. Therefore, this case of cross-manipulation methods requires detection methods to mine the most essential differences between real and fake faces. Experiments across manipulation methods effectively demonstrate the strong generalization ability of our method, which takes the implicit identity of the face as a clue to exploit fake-invariant features for discriminating real and fake faces.

Multi-source manipulation evaluation. In practice, it is usually necessary to train on multiple manipulated datasets and test on unknown samples. To demonstrate the effectiveness of our model in this multi-source manipulation scenario, we conduct experiments on the benchmark proposed by Sun *et al.* [45, 46]. Specifically, the model is trained on the three manipulated methods of FF++ and tested on the other one. In particular, we use EfficientNet-b0 as the backbone to ensure fair comparisons. The results are presented in Table 5. Our method generally outperforms others in terms of ACC and AUC on both high-quality and low-quality evaluations. The performance mainly benefits

Method	FF++		Celeb-DF		DFD		DFDC	
	AUC (%)	EER (%)	AUC (%)	EER (%)	AUC (%)	EER (%)	AUC (%)	EER (%)
Xception [42]	99.09	3.77	65.27	38.77	87.86	21.04	69.90	35.41
EN-b4 [47]	99.22	3.36	68.52	35.61	87.37	21.99	70.12	34.54
Face X-ray [27]	87.40	-	74.20	-	85.60	-	70.00	-
MLDG [24]	98.99	3.46	74.56	30.81	88.14	21.34	71.86	34.44
F3-Net [52]	98.10	3.58	71.21	34.03	86.10	26.17	72.88	33.38
MAT(EN-b4) [53]	99.27	3.35	76.65	32.83	87.58	21.73	67.34	38.31
GFF [32]	98.36	3.85	75.31	32.48	85.51	25.64	71.58	34.77
LTW [45]	99.17	3.32	77.14	29.34	88.56	20.57	74.58	33.81
Local-relation [7]	**99.46**	3.01	78.26	29.67	89.24	20.32	76.53	32.41
DCL [46]	99.30	3.26	82.30	26.53	91.66	16.63	76.71	31.97
UIA-ViT [55]	99.33	-	82.41	-	**94.68**	-	75.80	-
Ours	99.32	**2.99**	**83.80**	**24.85**	93.92	**14.01**	**81.23**	**26.80**

Table 2. Cross-database evaluation from FF++(C23) to Celeb-DF, DFD, and DFDC in terms of AUC and EER. The FF++ belongs to the intra-testing results while others represent to the unseen dataset testing.

Method	FF++ (%)	Celeb-DF (%)
Meso-4 [1]	84.70	54.80
Mesoinception4s [1]	83.00	53.60
FWA [28]	80.10	56.90
Xception [42]	95.50	65.50
Multi-task [36]	76.30	54.30
SMIL [31]	96.80	56.30
Two Branch [33]	93.18	73.41
EN-b4 [47]	96.39	71.10
MAT [53]	96.41	72.50
GFF [32]	95.73	74.12
SPSL [20]	96.91	76.88
ITA [44]	**96.94**	77.35
Ours	96.79	**82.04**

Table 3. Cross-dataset evaluation from FF++(C40) to deepfake class of FF++ and Celeb-DF in terms of AUC.

Train	Method	DF	FS	FST	Mean
DF	EN-b4	99.97	46.24	51.26	65.82
	MAT	99.92	40.61	45.39	61.97
	GFF	99.87	47.21	51.93	66.34
	DCL	**99.98**	61.01	68.45	76.48
	Ours	99.51	**63.83**	**73.49**	**78.94**
FS	EN-b4	69.25	99.89	60.76	76.63
	MAT	64.13	99.67	57.37	73.72
	GFF	70.21	99.85	61.29	77.12
	DCL	74.80	**99.90**	64.86	79.85
	Ours	**75.39**	99.73	**66.18**	**80.43**
FST	EN-b4	61.11	56.19	99.52	72.27
	MAT	58.15	55.03	99.16	70.78
	GFF	61.48	56.17	99.41	72.35
	DCL	63.98	58.43	99.49	73.97
	Ours	**65.42**	**59.50**	**99.50**	**74.81**

Table 4. Cross-manipulation evaluation in terms of AUC. Diagonal results indicate the intra-testing performance. DF, FS and FST denote the DeepFakes, FaceSwap and FaceShifter datasets, respectively.

from the unique perspective of our proposed IID framework, which explores essential forgery clues so as to be robust to multiple manipulation methods.

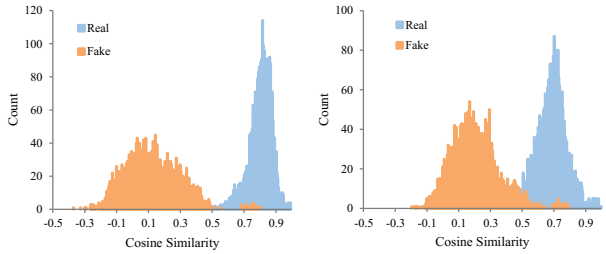
4.4. Visualization

Visualization of explicit identity contrast. To visualize the effectiveness of our proposed EIC, we conduct visualization analysis on FF++(C23) and FaceShifter. Specifically, given a face image (real or fake), we use our implicit identity embedding network trained on FF++(C23) to extract its features as an implicit identity. Besides, CosFace [51] is employed to extract the features of its face images as explicit identities. The cosine similarity between explicit and implicit identities serves as the explicit-implicit identity similarity (EIIS) for such images. For the preprocessed DeepFakes of FF++(C23) and FaceShifter video frames, we calculate the EIIS for each face image in them according to the above method. The cosine similarity distribution is shown in Figure 3. Overall, real and fake faces have a distinct boundary in terms of EIIS. Specifically, the EIIS of fake faces is about -0.3 to 0.5, while the EIIS of real faces is about 0.5 to 1.0. Note that there are still some samples that are obfuscated, especially for FaceShifter as it is across datasets.

Visualization of implicit identity exploration. To demonstrate that our method effectively explores the implicit identity, we also conduct a visual analysis of FF++(C23) and FaceShifter. Specifically, the fake face dataset consists of

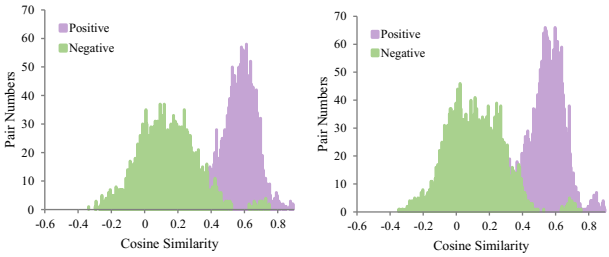
Method	GID-DF (C23)		GID-DF (C40)		GID-F2F (C23)		GID-F2F (C40)	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)
EfficientNet [47]	82.40	91.11	67.60	75.30	63.32	80.10	61.41	67.40
Focalloss [30]	81.33	90.31	67.47	74.95	60.80	79.80	61.00	67.21
ForensicTransfer [9]	72.01	-	68.20	-	64.50	-	55.00	-
Multi-task [36]	70.30	-	66.76	-	58.74	-	56.50	-
MLDG [24]	84.21	91.82	67.15	73.12	63.46	77.10	58.12	61.70
LTW [45]	85.60	92.70	69.15	75.60	65.60	80.20	65.70	72.40
DCL [46]	87.70	94.9	75.90	83.82	68.40	82.93	67.85	**75.07**
Ours	**88.21**	**95.03**	**76.90**	**84.55**	**69.36**	**84.37**	**67.99**	74.80

Table 5. Performance on multi-source manipulation evaluation. GID-DF means traning on the other three manipulated methods of FF++ and test on DeepFakes. The same for the others.



(a) DeepFakes (b) FaceShifter

Figure 3. Cosine similarity distribution of explicit and implicit identities for real and fake samples. We respectively extract the explicit and implicit identity features of face images and calculate the cosine similarity between them.



(a) DeepFakes (b) FaceShifter

Figure 4. Cosine similarity distribution for positive and negative samples. For a fake face dataset, the fake face and the target face are organized as positive sample (same implicit identity), the fake face and source face are organized as negative sample (different implicit identities). We resort our implicit identity embedding network trained on FF++(C23) to extract the implicit identity features of each pair of samples respectively.

several source-target-fake face video triples. We organize the fake face and the target face as positive sample (same implicit identity), the fake face and source face as negative sample (different implicit identities). In this way, we construct the corresponding fake face verification datasets for DeepFakes of FF++(C23) and FaceShifter. Subsequently, we resort our implicit identity embedding network trained on FF++(C23) to extract the implicit identity features of each pair of samples respectively. The cosine distance of each pair of face features is used as their identity similarity. The cosine similarity distribution is shown in Figure 4, which basically conforms to the normal distribution for positive and negative samples. We observe that positive sample pairs and negative sample pairs are distinguishable in terms of cosine similarity. Moreover, the distribution variance of positive samples is significantly smaller than that of negative samples, which implies that the implicit identities we extract are relatively stable. The results explain the effectiveness of our IID from the implicit identity perspective.

5. Conclusion

In this paper, we consider a new perspective for face swapping detection that focuses on the implicit identity of face. Specifically, we propose a novel implicit identity driven framework for face swapping detection. Particularly, we design an explicit identity contrast (EIC) loss and an implicit identity exploration (IIE) loss to guide a CNN backbone, which can embed face images into the implicit identity space. EIC aims to pull real samples closer to their explicit identities and push fake samples away from their explicit identities. Moreover, IIE is margin-based and guide fake faces with known target identities to have small intra-class distances and large inter-class distances. Extensive experiments and visualizations on several datasets demonstrate the superiority and generalization capability of our method over the state-of-the-art competitors.

6. Acknowledgement

This work is supported by National Key Research and Development Program of China (2021YFF0602102) and National Natural Science Foundation of China (U1903214, 62171324, U20B2049, U21B2018), and Key R&D Program of Hubei Province (2020BAB018, 2022BAA079).

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE WIFS*, pages 1–7, 2018. 1, 3, 7

[2] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *IEEE CVPR*, pages 6713–6722, 2018. 2

[3] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *IEEE CVPR*, pages 4113–4122, 2022. 3, 4

[4] Shenhao Cao, Qin Zou, Xiuqing Mao, Dengpan Ye, and Zhongyuan Wang. Metric learning for anti-compression facial forgery detection. In *ACM MM*, pages 1929–1937, 2021. 4

[5] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *IEEE ICCV*, pages 5933–5942, 2019. 1

[6] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *ACM MM*, pages 2003–2011, 2020. 2

[7] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *AAAI*, volume 35, pages 1081–1088, 2021. 1, 7

[8] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE CVPR*, pages 1251–1258, 2017. 3

[9] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 8

[10] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *IEEE CVPR*, pages 5781–5790, 2020. 1, 3

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE CVPR*, pages 4685–4694, 2018. 2

[12] Jiankang Deng, Jia Guo, Zhou Yuxiang, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv: Computer Vision and Pattern Recognition*, 2019. 5

[13] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 5

[14] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *ICML*, pages 3247–3258. PMLR, 2020. 3

[15] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *IEEE CVPR*, pages 3404–3413, 2021. 2

[16] Yue Gao, Fangyun Wei, Jianmin Bao, Shuyang Gu, Dong Chen, Fang Wen, and Zhouhui Lian. High-fidelity and arbitrary face editing. In *IEEE CVPR*, pages 16115–16124, 2021. 1

[17] Deepfakes github. Deepfakes. http://github.com/deepfakes/faceswap, 2017. 2, 5

[18] FaceSwap github. Faceswap. https://github.com/MarekKowalski/FaceSwap, 2017. 5

[19] Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. In *AAAI*, volume 36, pages 735–743, 2022. 1

[20] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *ACM MM*, pages 3473–3481, 2021. 3, 7

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. 5

[22] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *IEEE CVPR*, pages 5901–5910, 2020. 2

[23] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *IEEE ICCV*, pages 3677–3685, 2017. 2

[24] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, volume 32, 2018. 7, 8

[25] 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 *IEEE CVPR*, pages 6458–6467, 2021. 3

[26] 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. 2, 5

[27] 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 *IEEE CVPR*, pages 5001–5010, 2020. 1, 3, 6, 7

[28] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018. 7

[29] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE CVPR*, pages 3207–3216, 2020. 5

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE CVPR*, pages 2980–2988, 2017. 8

[31] 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 *IEEE CVPR*, pages 772–781, 2021. 7

[32] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *IEEE CVPR*, pages 16317–16326, 2021. 7

[33] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *ECCV*, pages 667–684. Springer, 2020. 3, 7

[34] R Natsume, T Yatagawa, and S Morishima. Rsgan: Face swapping and editing using face and hair representation in latent spaces. arxiv 2018. *arXiv preprint arXiv:1804.03447*. 2

[35] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Fsnet: An identity-aware generative model for image-based face swapping. In *ACCV*, pages 117–132. Springer, 2018. 2

[36] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–8, 2019. 3, 7, 8

[37] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *IEEE ICASSP*, pages 2307–2311, 2019. 1, 3

[38] Google Research Nick Dufour and Jigsaw Andrew Gully. Deep fake detection dataset. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html, 2019. 5

[39] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *IEEE ICCV*, pages 7184–7193, 2019. 2

[40] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *NeurIPS Workshop*, 2017. 5

[41] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, pages 86–103. Springer, 2020. 1, 3

[42] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE ICCV*, pages 1–11, 2019. 1, 3, 4, 5, 6, 7

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[44] Ke Sun, Hong Liu, Taiping Yao, Xiaoshuai Sun, Shen Chen, Shouhong Ding, and Rongrong Ji. An information theoretic approach for attention-driven face forgery detection. In *ECCV*, pages 111–127. Springer, 2022. 7

[45] Ke Sun, Hong Liu, Qixiang Ye, Yue Gao, Jianzhuang Liu, Ling Shao, and Rongrong Ji. Domain general face forgery detection by learning to weight. In *AAAI*, volume 35, pages 2638–2646, 2021. 6, 7, 8

[46] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *AAAI*, volume 36, pages 2316–2324, 2022. 3, 4, 6, 7, 8

[47] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019. 7, 8

[48] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM TOG*, 38(4):1–12, 2019. 1, 5

[49] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE CVPR*, pages 2387–2395, 2016. 5

[50] Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *IEEE CVPR*, pages 14923–14932, 2021. 1, 3

[51] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE CVPR*, pages 5265–5274, 2018. 1, 2, 4, 5, 7

[52] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *AAAI*, volume 34, pages 12321–12328, 2020. 6, 7

[53] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *IEEE CVPR*, pages 2185–2194, 2021. 1, 3, 7

[54] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *IEEE CVPRW*, pages 1831–1839, 2017. 3

[55] Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *ECCV*. Springer, 2022. 3, 6, 7

[56] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *ACM MM*, pages 2382–2390, 2020. 3