Learning Real Facial Concepts for Independent Deepfake Detection

Ming-Hui Liu[1] , **Harry Cheng**[2] , **Tianyi Wang**[3] , **Xin Luo**[1] , **Xin-Shun Xu**[1] *
[1]School of Software, Shandong University
[2]School of Computing, National University of Singapore
[3]College of Computing and Data Science, Nanyang Technological University
liuminghui@mail.sdu.edu.cn, xaCheng1996@gmail.com, terry.ai.wang@gmail.com,
luoxin.lxin@gmail.com, xuxinshun@sdu.edu.cn

Abstract

Deepfake detection models often struggle with generalization to unseen datasets, manifesting as misclassifying real instances as fake in target domains. This is primarily due to an overreliance on forgery artifacts and a limited understanding of real faces. To address this challenge, we propose a novel approach RealID to enhance generalization by learning a comprehensive concept of real faces while assessing the probabilities of belonging to the real and fake classes independently. RealID comprises two key modules: the Real Concept Capture Module (RealC^2) and the Independent Dual-Decision Classifier (IDC). With the assistance of a Multi-Real Memory, RealC^2 maintains various prototypes for real faces, allowing the model to capture a comprehensive concept of real class. Meanwhile, IDC redefines the classification strategy by making independent decisions based on the concept of the real class and the presence of forgery artifacts. Through the combined effect of the above modules, the influence of forgery-irrelevant patterns is alleviated, and extensive experiments on five widely used datasets demonstrate that RealID significantly outperforms existing state-of-the-art methods, achieving a 1.74% improvement in average accuracy.

1 Introduction

With the rapid development of generative algorithms, the barrier to the synthesis of facial forgeries has decreased significantly, fostering online fraud, opinion manipulation, and pornographic content dissemination. To mitigate the abuse of face forgeries, a considerable number of effective deepfake detection methods [Xia *et al.*, 2024b; Sun *et al.*, 2024] have been proposed, and they have achieved significant success when training and testing on the same dataset, *i.e.*, evaluating under the within-dataset setting. Most of these methods treat deepfake detection as a binary classification problem, aiming to capture various artifacts in facial images, such as inter-frame correlation [Qiao *et al.*, 2024] and style latent flow [Choi *et al.*, 2024] to distinguish between real and fake.

*Corresponding author.

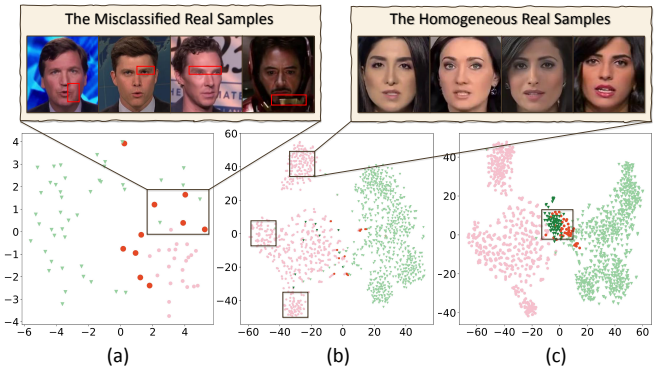

Figure 1: (a) The real samples are misclassified as fake (highlighted in red), due to the local imperfections that are not actual forgery traces; (b) The features of real faces tend to cluster more tightly; (c) Non-overlapping test samples leading to potential misclassification.

However, when these models are applied to unseen datasets, their performance often degrades significantly. This issue, *i.e.*, the lack of generalization, has garnered widespread attention in recent years. The solutions for it can be broadly categorized into two approaches: 1) Expanding the scale or diversity of the training dataset to include more complex forgery traces; 2) Optimizing the model architecture to extract general forgery artifacts that are consistent across various deepfake techniques. Based on the aforementioned approaches, many studies achieve remarkable detection performance. However, as illustrated in Figure 1(a), the existing models counterintuitively misclassify real instances as fake ones, rather than maintaining a balanced misclassification rate between real and fake instances (red circles and green triangles represent real and fake samples, respectively).

We attribute this phenomenon to two key factors: 1) **Data**. As shown in Figure 1(b), real facial features exhibit a clustered distribution, while the distribution of fake ones is relatively uniform. This will cause the model to overfit these homogeneous real samples and limit the feature distribution of the real class. Meanwhile, in Figure 1(c), we find test samples that shift beyond the original training distribution are likely to be misclassified. This indicates that the narrow distribution of the real class makes it difficult to cover the real samples in the test set and classify the extra-distributed real samples accurately. 2) **Classifier**. Deepfake detectors rely on basic binary classifiers to differentiate between real and

fake instances. However, because of the difficulty in obtaining the concept of real faces, the classifier would overly rely on forgery traces to make judgments. In other words, the decision-making process has a built-in priority: the model could first determine whether an image is fake according to forgery artifacts, and only if it is not classified as fake will it be recognized as real. Moreover, due to the limited diversity of the training dataset, forgery-irrelevant patterns, such as noise and blur, will also be entangled with forgery artifacts. When these so-called 'mis-artifacts' appear in real samples during the testing phase, they can easily lead to an incorrect judgment. Given the insufficient exploration of real samples and the overlook of real facial concepts, a natural question motivates us – Could we improve the generalization by learning a more comprehensive concept for real faces and inducing the model to make judgments based on both the real facial concepts and the presence of forgery artifacts?

We address the aforementioned issues by proposing the Real Facial Concepts based Independent Deepfake Detection (RealID) approach. Our approach focuses on exploring and leveraging real facial concepts, enabling the model to make judgments based on both real facial concepts and forgery artifacts. Specifically, we design two novel modules: 1) Real Concept Capture Module (RealC^2), which employs a Multi-Real Memory mechanism to store various real prototypes. It operates over the entire training set, learning a more comprehensive real facial concept with specially designed Prototype Distinction Loss and Prototype Diversity Loss. In this way, the model can focus more on the differences of the real samples and generate a more robust distribution for the real class. 2) Independent Dual-Decision Classification strategy (IDC). We reformulate the binary classification strategy to a novel independent dual-decision strategy. By adding a regularization term and two auxiliary classes, the probability of belonging to the incorrect classes can further decrease after being optimized with the cross-entropy loss. This means that our IDC can adaptively search for an alternative optimization path, *i.e.*, utilizing the real facial concepts learned from RealC^2, and mitigate the misguidance of the overfitting mis-artifacts. Using previously overlooked real facial concepts, our method enables the model to make robust decisions from both real and fake perspectives, thereby enhancing its generalization capability in target domains. We conduct extensive experiments on five widely used datasets and the experimental results demonstrate a 1.74% average accuracy improvement compared to existing state-of-the-art methods.

In summary, our contributions are three-fold:

- We attribute the issue of limited model generalization to the inadequacy in learning comprehensive real facial concepts. Our observations indicate that existing detectors tend to overly rely on forgery artifacts, which often results in the misclassification of real instances as fake.
- We propose the Real Facial Concepts based Independent Deepfake Detection approach, which combines two special modules to facilitate the learning of real facial concepts and enable the classifier to independently assess the probabilities of belonging to real and fake classes.
- Extensive experiments are conducted to demonstrate the effectiveness of our approach which achieves state-of-the-art generalization performance on several datasets.

2 Related Work

2.1 Deepfake Generation

The rapid development of portrait synthesis has propelled deepfake technology [Xu *et al.*, 2022b] into a prominent research area. Autoencoders [Kingma and Welling, 2014] form the foundational architecture for many early deepfake approaches [Thies *et al.*, 2016; Suwajanakorn *et al.*, 2017]. Typically, these methods involve training two separate models on a reconstruction task and then swapping their decoders to change the identities of source faces. While these techniques can produce realistic face-swapped images, they are limited to one-to-one face swapping. To enable more versatile synthesis, GANs [Goodfellow *et al.*, 2014] have become increasingly popular due to their ability to generate arbitrary high-quality facial images. For example, StyleGAN [Karras *et al.*, 2019] allows for the manipulation of high-level facial attributes by using a progressively growing structure combined with adaptive instance normalization. IPGAN [Bao *et al.*, 2018] disentangles the identity and attributes of the source and target faces for synthesizing new faces. Recently, identity-relevant features have been integrated into deepfake generation to enhance identity consistency. Xu *et al.* [Xu *et al.*, 2022a] refined identity consistency by augmenting both local and global identity-relevant features through cross-scale semantic interaction modeling, achieving more coherent face swapping that maintains identity integrity.

2.2 Deepfake Detection

Deepfake detection [Wang *et al.*, 2024; Hong *et al.*, 2024; Yan *et al.*, 2024; Xia *et al.*, 2024a; Guan *et al.*, 2024] is generally cast as a binary classification task. Preliminary efforts often endeavor to detect the specific manipulation traces [Jia *et al.*, 2022]. Masi *et al.* [Masi *et al.*, 2020] utilized optical and frequency artifacts separately. SSTNet [Wu *et al.*, 2020] detects edited faces through spatial, steganalysis, and temporal features. These models have shown certain improvements on some datasets. However, they often encounter inferior performance when applied to different data distributions. To improve generalization [Tan *et al.*, 2024; Li *et al.*, 2023], one manner is to construct datasets that encompass a wider range of forgery methods [Khalid *et al.*, 2021; Le *et al.*, 2021; Cheng *et al.*, 2024a]. For instance, the KoDF dataset [Kwon *et al.*, 2021] comprises over 200,000 videos generated by six distinct algorithms, while DF-Platter [Narayan *et al.*, 2023] involves over 130,000 multi-face forgery videos. These high-quality datasets encompass diverse data sources. However, it is crucial to note that these datasets often place greater emphasis on the fake side, while the distribution of real images tends to be more homogeneous. This imbalance makes it easier for a model to overfit on real features, which in turn leads to misclassification. Another approach is to learn more generalized features of forgeries [Wang and Deng, 2021; Luo *et al.*, 2021; Zhao *et al.*, 2021; Li *et al.*, 2020a; Nirkin *et al.*, 2022]. For instance, RealForensics [Haliassos *et al.*, 2022] exploits the visual and auditory [Cheng *et al.*,

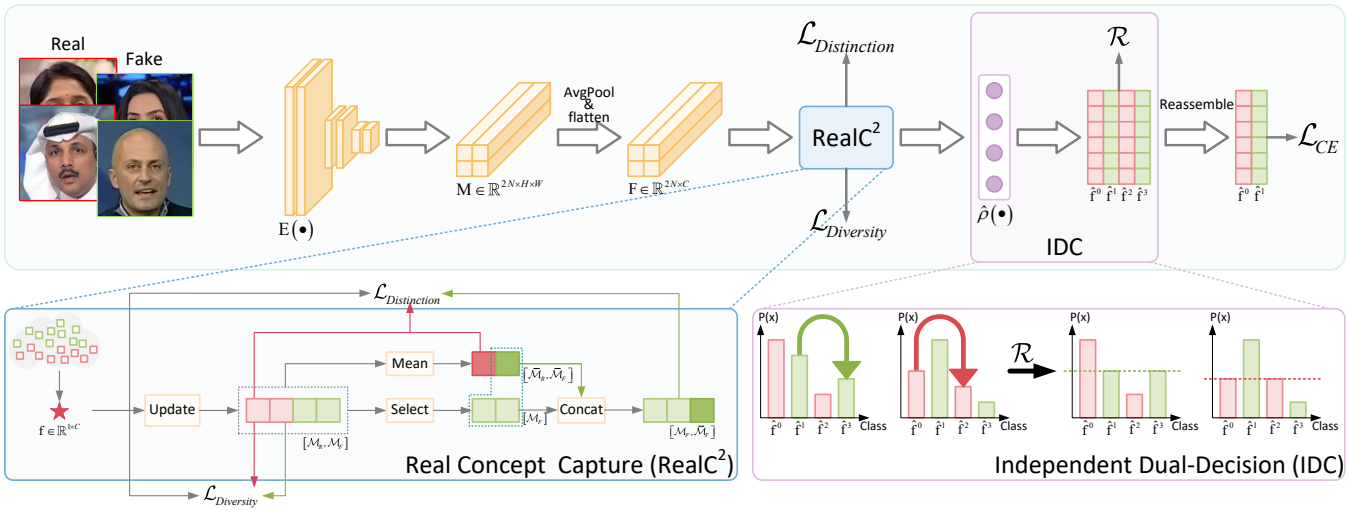

Figure 2: Overall architecture of the RealID framework. RealID consists of two main modules: (i) the Real Concept Capture(RealC^2) module, which uses a multi-real memory mechanism to learn a more comprehensive real facial concept, and (ii) the Independent Dual-Decision(IDC) module, which leverages regularization terms to independently optimize the decision-making process for different categories.

2023] correspondence in real videos. Chen *et al.* [Chen *et al.*, 2022] specified the blending regions and facial attributes to enrich the deepfake dataset with more manipulation types. Shiohara *et al.* [Shiohara and Yamasaki, 2022] introduce a self-blended framework to capture boundary-fusion features. These methods have achieved notable success in capturing forgery artifacts, reaching performance plateaus on multiple datasets. However, they tend to lack sufficient focus on real instances, leading to an imbalanced decision-making process.

3 Methodology

As illustrated in Figure 2, our RealID framework improves generalization through the combination of two novel modules: the Real Concept Capture (RealC^2) module and the Independent Dual-Decision (IDC) module. Specifically, RealC^2 uses a Multi-Real memory mechanism [Park *et al.*, 2020] to maintain several prototypes of real faces, enabling the model to extract more comprehensive features from real samples and preventing the model from overfitting to the homogeneous training sets. IDC modifies the decision logic of the naive deepfake classifier, which makes decisions based solely on the presence or absence of forgery artifacts. By incorporating an additional regularization term and auxiliary classes, it can independently reduce the probability of belonging to the false classes through different optimization paths. This means IDC mitigates the influence of mis-artifacts and improves the generalization of the deepfake detector.

During the training process, we specifically construct mini-batches to ensure the inclusion of both N real samples $\mathbf{X}^{\text{real}}$(with the labels of $\mathbf{Y}=0$) and N fake samples $\mathbf{X}^{\text{fake}}$(with the labels of $\mathbf{Y}=1$). Then, we obtain the feature maps $\mathbf{M} \in \mathbb{R}^{2N\times H\times W}$ via a feature extractor $\mathrm{E}(\cdot)$ and transform them to feature vectors $\mathbf{F} \in \mathbb{R}^{2N\times C}$ by a series of pooling and flattening operations. This feature extraction process can be formalized as follows:

$$\mathbf{F} = \mathrm{Flatten}\left(\mathrm{AvgPool}\left(\mathrm{E}\left(\mathbf{X}\right)\right)\right). \tag{1}$$

3.1 Real Concept Capture Module

As shown in Figure 1(b), the homogeneous real samples in the training set may mislead the detection models into overfitting to a specific cluster while neglecting the concept of real faces. This limitation makes it difficult to address the distribution shift in the testing set. To overcome this issue, we design the RealC^2 module to extract more comprehensive real facial features. In particular, instead of treating the real class as a single homogeneous entity, our model subdivides real facial features with the assistance of diverse real prototypes. In this way, the model can distinguish the fine-grained differences between real instances and acquire a more comprehensive understanding of real faces.

Prototype Initialization. Before training begins, we randomly initialize K vectors $\mathcal{M}_R = \{\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_K\}$ to construct the Multi-Real Memory, and each $\mathbf{m}_i \in \mathcal{M}_R$ can be considered as a real facial prototype. During the training, these prototypes are designed to record real facial patterns and explicitly enforced to maintain diversity. This prevents the model from focusing solely on the commonalities of real faces (specific implementation will be discussed below).

Prototype Updating. In each training iteration, we update the real facial prototypes $\mathcal{M}_R$ with total real facial features in the mini-batch. Specifically, we first calculate the similarity between each prototype $\mathbf{m}_i$ and specific real facial features, resulting in a correlation map of size $K \times N$. Then, we normalize the correlation map using Softmax function along both horizontal and vertical directions. For the vertical one, we perform feature normalization among the K prototypes:

$$w_{i,j} = \frac{\exp\left(\mathbf{m}_i{}^T \mathbf{f}_j\right)}{\sum_{i=1}^{K} \exp\left(\mathbf{m}_i{}^T \mathbf{f}_j\right)}, \tag{2}$$

where $\mathbf{f}_j$ is the feature of a specific real instance. Similarly, the horizontal one is performed on the N real features:

$$v_{i,j} = \frac{\exp\left(\mathbf{m}_i{}^T \mathbf{f}_j\right)}{\sum_{j=1}^{N} \exp\left(\mathbf{m}_i{}^T \mathbf{f}_j\right)}. \tag{3}$$

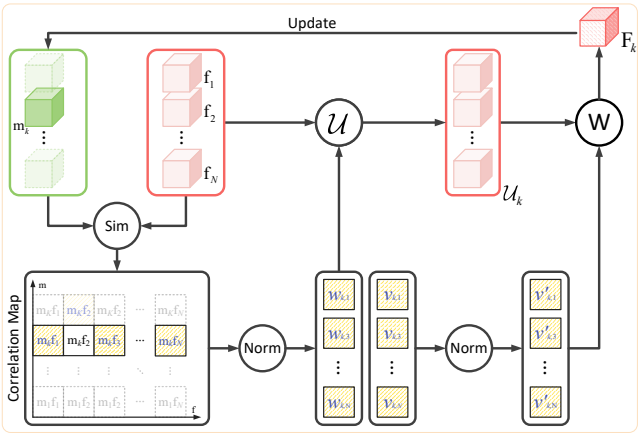

Figure 3: Illustration of the update process for real facial prototypes.

The results in Equation (2) and (3), *i.e.*, $w_{i,j}$ and $v_{i,j}$, represent the vertical and horizontal normalized matching probabilities and guide the selection of the most suitable real facial features for prototypes update. Firstly, each sample $\mathbf{f}_j$ is assigned to its nearest prototype $\mathbf{m}_{p_1}$. And we select the prototype $\mathbf{m}_{p_1}$ corresponding to the highest vertical normalized matching probabilities between $\{w_{1,j}, w_{2,j}, \cdots, w_{K,j}\}$:

$$p_1 = \text{argmax} w_{i,j}. \tag{4}$$

This allocation process means that each sample has exactly one nearest prototype, while the number of samples assigned to each prototype is not fixed. Then, all samples assigned to the same prototype $\mathbf{m}_k$ can be considered as an update set $\mathcal{U}_k$. Based on corresponding horizontal normalized matching probabilities $\{v_{k,1}, v_{k,2}, \cdots, v_{k,n}\}$, the prototype $\mathbf{m}_k$ will be updated using a weighted sum of update set $\mathcal{U}_k$:

$$\mathbf{m}_k \leftarrow \text{Normalize}(\mathbf{m}_k + \sum_{\mathbf{f}_j \in \mathcal{U}_k} v'_{k,j} \mathbf{f}_j), \tag{5}$$

where $\text{Normalize}(\cdot)$ represents the L2 norm function, and $v'_{k,j}$ is the variant of horizontal normalized matching probabilities $v_{k,j}$, which is further regularized with the max value in its respective update set $\mathcal{U}_k$ as follows:

$$v'_{k,j} = \frac{v_{k,j}}{\max\limits_{\mathbf{f}_{j'} \in \mathcal{U}_k} v_{k,j'}}. \tag{6}$$

To have a more clear perspective, we illustrate the computation and update process related to $\mathbf{m}_k$ in Figure 3.

Prototypes Distinction Loss and Diversity Loss. We utilize two specially designed prototype losses to achieve precise prototype learning. First, we employ the Prototypes Distinction Loss ($\mathcal{L}_{\text{Distinction}}$) to align the distribution within the subclass. Specifically, we encourage each real facial feature $\mathbf{f}_j$ to move toward its nearest prototype $\mathbf{m}_{p_1}$ and the mean real facial prototype $\bar{\mathcal{M}}_R$. To enhance the distinction between the real and fake classes, we introduce a set of fake facial prototypes[1] $\mathcal{M}_F$. Then, we further push the real facial feature $\mathbf{f}_j$ away from the fake prototypes and the mean fake prototype $\bar{\mathcal{M}}_F$. This process can be expressed as:

$$\mathcal{L}_{\text{Distinction}} = -\log \frac{\exp(\mathrm{m}_{p_1}{}^T \mathrm{f}_j) + \exp(\bar{\mathcal{M}}_R^T \mathrm{f}_j)}{\sum_{k=K}^{2K} \exp(\mathrm{m}_k{}^T \mathrm{f}_j) + \exp(\bar{\mathcal{M}}_F^T \mathrm{f}_j)}, \tag{7}$$

where the mean prototype $\bar{\mathcal{M}}$ can be calculated as follows:

$$\bar{\mathcal{M}} = \frac{\sum_{k=1}^{K} \mathrm{m}_k}{K}. \tag{8}$$

In addition, due to the lack of explicit supervision for subclasses, the Prototypes Distinction Loss may mislead different real subclasses to collapse into an undesirable whole. To address this, we design a Prototypes Diversity Loss to maintain the diversity of real facial prototypes:

$$\mathcal{L}_{\text{Diversity}} = \sum_{j=1}^{N} [\|\mathbf{f}_j - \mathbf{m}_{p_1}\|_2 - \|\mathbf{f}_j - \mathbf{m}_{p_2}\|_2 + \alpha]_+ \tag{9}$$
$$p_2 = \underset{i \neq p_1}{\text{argmax}}\, w_{i,j},$$

where $\mathbf{m}_{p_2}$ is the second nearest prototype of $\mathbf{f}_j$. α is a predefined margin. In this way, the features of different sub-classes can maintain a proper distance during training.

3.2 Independent Dual-Decision Module

Because of the difficulty in learning and leveraging real facial concepts, deepfake detection models would classify samples with forgery-related features as fake and classify samples without those features as real. Traditionally, deepfake detectors incorporate a binary classifier and are updated based on a naive cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{2N} \sum_{j=1}^{2N} y_j \log(\rho(\mathbf{f}_j)) + (1 - y_j) \log(1 - \rho(\mathbf{f}_j)), \tag{10}$$

where $2N$ is the total number of samples in a mini-batch. y_i is the ground-truth of feature $\mathbf{f}_i$. $\rho(\cdot)$ is the classifier that outputs the probability of belonging to the real and fake classes.

Due to the inherent properties of binary cross-entropy loss, as $\rho(\mathbf{f}_j)$, the predicted probability of belonging to the fake class *increases*, $1 - \rho(\mathbf{f}_j)$, the probability of belonging to the real class will correspondingly *decrease*, and vice versa. If the detector can abstract accurate criteria for identifying forgery faces, this 'one rises as the other falls' single-decision logic will achieve a satisfactory optimization result. However, in real-world scenarios, the forgery features are often tangled with overfitted 'mis-artifacts' and have poor generalization ability in the target domain. Therefore, we introduce a more reliable Independent Dual-Decision Classification (IDC) strategy and produce the final classification probability based on both real facial concepts and forgery artifacts.

Independent Dual-Decision Classifier. As for a sample $\mathbf{f}_j$, to sever the one-to-one correspondence between the fake probability $\rho(\mathbf{f}_j)$ and the real probability $1 - \rho(\mathbf{f}_j)$, we extend

[1]The fake prototypes are updated using the same method as the real ones and relying on the fake instances in the mini-batch.

the output dimensions of the naive binary classifier to twice, and the classification process is as follows:

$$\hat{\mathbf{f}}_j = \hat{\rho}(\mathbf{f}_j) = \mathrm{Sofmax}(\mathrm{FC}(\mathbf{f}_j)), \tag{11}$$

where the Independent Dual-Decision Classifier $\hat{\rho}(\cdot)$ consists of the fully connected layer and the normalization layer. The output vector $\hat{\mathbf{f}}_j \in \mathbb{R}^{1\times 4}$ can be further represented as:

$$\hat{\mathbf{f}}_j = \left\{\hat{\mathbf{f}}_j^0, \hat{\mathbf{f}}_j^1, \hat{\mathbf{f}}_j^2, \hat{\mathbf{f}}_j^3\right\} (s.t. \sum_{m=0}^{3} \hat{\mathbf{f}}_j^m = 1), \tag{12}$$

where the first and second dimensions of the output vector $\hat{\mathbf{f}}_j$ represent the probabilities of belonging to the real and fake classes, respectively. The third and fourth dimensions serve as the auxiliary components for our IDC strategy, whose function will be elaborated in the next section. In this setup, the constraints are imposed on all four dimensions, so that $\hat{\mathbf{f}}_j^0$ no longer entirely follows the variations of $\hat{\mathbf{f}}_j^1$.

Independent Dual-Decision Regularization. After obtaining the relative outputs from $\hat{\rho}(\cdot)$, we try to further reduce the probabilities of belonging to the incorrect classes by appending an extra regularization term $\mathcal{R}$ to the classification loss:

$$\mathcal{R} = \begin{cases} \sum_{j=1}^{2N} \beta \cdot d_j^2 & \text{if } |d| < 1 \\ \sum_{j=1}^{2N} (|d_j| - \beta) & \text{otherwise,} \end{cases} \tag{13}$$

where $2N$ and y_j are consistent with that in Equation (10). β is a predefined value to control the optimization intensity. d_j is used to represent the discrepancy between the incorrect probabilities and their corresponding auxiliary components:

$$d_j = \begin{cases} \hat{\mathbf{f}}_j^1 - \hat{\mathbf{f}}_j^3 & y_j = 0 \\ \hat{\mathbf{f}}_j^0 - \hat{\mathbf{f}}_j^2 & y_j = 1. \end{cases} \tag{14}$$

This way, the probability of belonging to the incorrect class will progressively approach the lower auxiliary component without being influenced by the correct probability. For instance, for a real sample $\hat{\mathbf{f}}_j$ (with the label of $y_j = 0$), $\mathcal{R}$ balances the value of $\hat{\mathbf{f}}_j^1$ and $\hat{\mathbf{f}}_j^3$ which consistently maintains a small value. This means combining the classifier $\hat{\rho}(\cdot)$ and regularization term $\mathcal{R}$, our Independent Dual-Decision strategy can rely on another optimization path (*i.e.*, leveraging the real facial concepts) to make robust judgments and implicitly reduces the reliance on forgery artifacts.

3.3 Training Strategy

Combining all modules, the network framework is optimized in an end-to-end manner based on the following loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{Diversity}} + \lambda_2 \mathcal{L}_{\text{Distinction}} + \lambda_3 \mathcal{R}, \tag{15}$$

where $\mathcal{L}_{\text{CE}}$ is a binary cross-entropy loss. $\mathcal{L}_{\text{Diversity}}$ and $\mathcal{L}_{\text{Distinction}}$ are used to learn comprehensive real prototypes. $\mathcal{R}$ is a regularization term for achieving the independent dual-decision strategy. λ_1, λ_2, and λ_3 are the hyperparameters that balance the contribution of each term. Through this holistic loss function, the model optimizes classification results based on both comprehensive real concepts and forgery artifacts, thereby enhancing generalization in unseen target domains.

4 Experiment

4.1 Implementation

We utilized Dlib[2] to extract faces and resize them to 256 × 256 pixels for both the training and testing sets. We conducted the experiments on a single RTX 3090 GPU with a batch size of 16. The backbone we employed is EfficientNet [Tan and Le, 2019] and we also switched the backbone to other networks, *e.g.*, ViT [Dosovitskiy *et al.*, 2021], to demonstrate the robustness of our proposed method in Section 4.2. The hyperparameters λ_1, λ_2, and λ_3 in Equation (15) are selected via grid search and set to 0.6, 1.0, and 1.0, respectively. Similar to the common setup for generalizable deepfake detection [Wang and Deng, 2021; Fei *et al.*, 2022; Cao *et al.*, 2022], we trained our model with the FF++ dataset [Rössler *et al.*, 2019]. This dataset includes 1,000 real videos from YouTube, as well as five types of manipulated videos yielding a total of 6,000 videos. Finally, to evaluate the generalization capability of our model, we performed the cross-dataset testing on five widely used deepfake datasets, *i.e.*, Celeb-DF [Li *et al.*, 2020b], DFD [Dufour and Gully, 2020], DFDC [Dolhansky *et al.*, 2020], DFDCp [Dolhansky *et al.*, 2020], and UADFV [Li *et al.*, 2018].

4.2 Main Experimental Results

Performance on Cross-Dataset Evaluation. We reported the generalization capability of several SoTA baselines and our RealID in Table 1. All of them are trained on FF++ and evaluated on other five testing datasets. This cross-dataset setup is challenging since neither the testing pristine/forged videos nor the manipulated techniques are visible in the training dataset. We utilized the AUC (area under the receiver operating characteristic curve) metric to quantify the performance. In addition to reporting the AUCs on each individual dataset, we also calculated the average AUC over the five testing datasets. From Table 1, we have three main observations: 1) Our method RealID significantly improves the generalization ability compared to several SoTA baselines. For instance, our method achieves an AUC improvement of nearly 20% on the Celeb-DF dataset compared to FoCus [Tian *et al.*, 2024] and an average AUC improvement of approximately 18% across five datasets compared to EfficientNet. 2) Our method is more robust. Most of the baselines are only effective on specific datasets, whereas our method achieves significant improvements across all datasets. For instance, RECCE, which demonstrates competitive performance on other datasets, unexpectedly suffers a significant performance drop on Celeb-DF. In contrast, our RealID approach consistently achieves effective improvements. 3) Exploring detailed image features can enhance performance. For instance, training with self-blended images enables SBI to achieve an impressive average AUC of 89%, and UCF leverages the common forgery features to achieve an AUC of 97% on the UAVDF dataset, outperforming other baselines.

Performance with Different Backbones. By switching the backbone, we further demonstrated the generalizability of our method. Specifically, we selected three widely employed

[2]http://dlib.net/.

Method	Venue	Celeb-DF	DFD	DFDC	DFDCp	UADFV	AVG
‡EfficientNet [Tan and Le, 2019]	ICML'19	64.59	92.31	65.43	80.27	63.19	73.16
‡Face X-ray [Li *et al.*, 2020a]	CVPR'20	74.76	93.47	61.57	71.15	64.34	73.06
‡CORE [Ni *et al.*, 2022]	CVPRW'22	79.45	93.74	62.60	75.74	65.41	75.39
‡RECCE [Cao *et al.*, 2022]	CVPR'22	69.71	93.15	62.82	74.19	78.61	75.70
‡SBI [Shiohara and Yamasaki, 2022]	CVPR'22	93.18	97.56	72.42	86.15	97.28	89.32
‡UCF [Yan *et al.*, 2023]	ICCV'23	81.90	93.09	66.21	80.94	97.15	83.86
FoCus [Tian *et al.*, 2024]	TIFS'24	76.13	-	68.42	76.62	-	-
Qiao et al. [Qiao *et al.*, 2024]	TPAMI'24	70.00	94.00	-	-	78.00	-
GRU [Choi *et al.*, 2024]	CVPR'24	89.00	96.10	-	-	-	-
ProDet [Cheng *et al.*, 2024b]	NeurIPS'24	84.48	-	72.40	81.16	-	-
RealID	-	**95.16**	**98.32**	**74.67**	**88.80**	**98.34**	**91.06**

Table 1: Performance comparison (%). All models are trained on the FF++ dataset. The best performance is marked as bold. ‡: We re-implemented this detector. -: The authors did not report the results on this dataset in their original paper.

Backbone	Celeb-DF	DFDC	DFDCp	UADFV	AVG
Xception	56.75	64.19	74.17	62.05	64.29
Xception+RealID	87.44	73.21	81.89	99.23	85.44
ViT-L	77.27	71.83	84.05	76.13	77.32
ViT-L+RealID	90.00	81.80	91.60	98.92	90.77
ViT-B	89.63	73.00	87.04	96.17	86.46
ViT-B+RealID	93.56	80.92	92.30	98.85	91.40

Table 2: Validate the effectiveness of RealID on different backbones. ViT-L and Vit-B represent different scales of the ViT.

backbones: Xception [Rössler *et al.*, 2019], ViT-L-32 [Dosovitskiy *et al.*, 2021], and ViT-B-16 [Dosovitskiy *et al.*, 2021] to evaluate the performance of combining them with our RealID. The training details remain consistent with those that utilize EfficientNet. It is worth noting that RealID can be seamlessly integrated into any deepfake backbone since it does not require modifications to the original architecture.

We have reported the experimental results in Table 2. From this table, we have the following observations: 1) Our RealID remains effective across different backbones. For instance, Xception+RealID achieves a 21.15% improvement in average AUC compared to the Xception alone, while ViT-L+RealID and ViT-B+RealID achieve 12.73% and 3.93% improvements on the Celeb-DF dataset, respectively. The above results fully demonstrate the generalization of our approach. 2) Stronger backbone architectures yield better performance but are harder to improve. For instance, the ViT-B-based model outperforms both EfficientNet and Xception in terms of average AUC. However, the performance improvement ratio of ViT-B+RealID is the smallest. 3) Appropriate patch size is more important than the model scale. For instance, ViT-B-16 with a patch size of 16 outperforms ViT-L-32 with a patch size of 32. This may also indicates the importance of local features in deepfake detection.

4.3 Ablation Studies

Module Analysis. In Table 3, we presented the results of ablation studies on two key modules in RealID, *i.e.*, RealC^2 and IDC. From the results, it can be observed that both components effectively enhance the generalizability of the model. For instance, compared with the backbone on the Celeb-DF dataset, RealC^2 and IDC improve the AUC by 30% and 29%, respectively. Combining the two components, the best result, a 30.57% AUC improvement, can be achieved.

Backbone	RealC^2	IDC	Testing Set			
			Celeb-DF	DFDC	DFDCp	UADFV
✓			64.59	65.43	80.27	63.19
✓	✓		94.92	73.58	88.11	98.09
✓		✓	93.79	74.64	88.44	98.30
✓	✓	✓	**95.16**	**74.67**	**88.80**	**98.34**

Table 3: AUC (%) comparison of different modules in RealID.

$\mathcal{L}_{\mathrm{Diversity}}$	$\mathcal{L}_{\mathrm{Distinction}}$	Testing Set			
		Celeb-DF	DFDC	DFDCp	UADFV
✓		91.83	72.22	75.37	95.60
	✓	93.87	72.49	87.81	97.70

Table 4: AUC (%) comparison of loss functions in RealC^2.

Loss Analysis. To ensure the diversity and distinction of the real prototypes used in RealC^2, we employ two relative loss functions, *i.e.*, $\mathcal{L}_{\mathrm{Diversity}}$ and $\mathcal{L}_{\mathrm{Distinction}}$, to supervise the training process. Specifically, $\mathcal{L}_{\mathrm{Distinction}}$ ensures the consistency of features within a subclass while maintaining the separability between the real and fake classes. $\mathcal{L}_{\mathrm{Diversity}}$ enforces diversity among prototypes, preventing them from converging into the same distribution. In Table 4, we reported the impacts of these two loss functions. It can be seen that, between the two loss functions, $\mathcal{L}_{\mathrm{Distinction}}$ plays a more significant role. For instance, the AUC on the Celeb-DF dataset will drop by 3.3% when $\mathcal{L}_{\mathrm{Distinction}}$ is removed. One critical reason for this is that $\mathcal{L}_{\mathrm{Distinction}}$ updates the real prototypes based on real instances. Without it, the real prototypes cannot undergo global optimization. Furthermore, $\mathcal{L}_{\mathrm{Diversity}}$ is also crucial, contributing approximately a 1.2% improvement. In summary, the combination of these two loss functions in RealC^2 delivers the best performance gains.

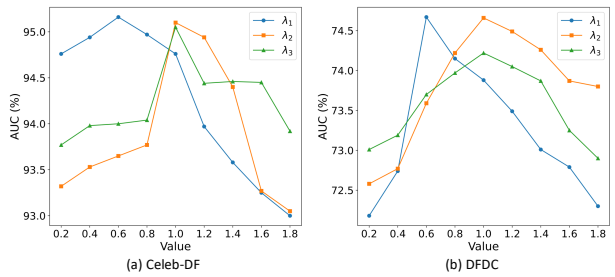

Figure 4: AUC (%) comparison under different hyperparameter combinations. For λ_1, λ_1, and λ_1, we vary one of their values while keeping the other two values fixed at 0.5.

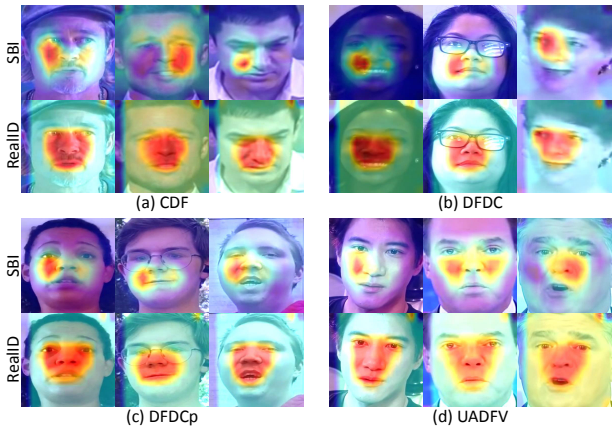

Figure 5: Illustration for no-cherry-pick heatmaps from four datasets, comparing our RealID with the SoTA baseline SBI.

Parameter Analysis. In Equation (15), the three parameters, *i.e.*, λ_1, λ_2, and λ_3, control the influence of the loss functions. Figure 4 illustrates the impacts of these parameters on the performance. Specifically, we progressively increase a certain hyperparameter within the range of [0, 2], while keeping the other two parameters fixed, to retrain the model. The performance is then evaluated on DFDC and Celeb-DF. From the figure, we observed that the impact of the three hyperparameters on the model's performance on both datasets follows the same pattern: as the corresponding λ value increases, the model's performance first improves and then drops sharply. Based on the combined performance of the three parameters, we empirically selected $\lambda_1 = 0.6$, $\lambda_2 = 1.0$, and $\lambda_3 = 1.0$ as the optimal configuration.

4.4 Qualititive Studies

In Figure 5, we presented heatmaps of no-cherry-pick instances from different datasets. It can be observed that, compared to SBI [Shiohara and Yamasaki, 2022], the baseline model, our approach achieves a more evenly distributed attention across the entire face. For example, on the DFDC dataset, SBI focuses primarily on the left cheek, whereas our approach distributes attention more evenly across the lips, nose, and eyes. This indicates that our RealID makes decisions based on more comprehensive facial concepts rather than the presence or absence of local forgery traces.

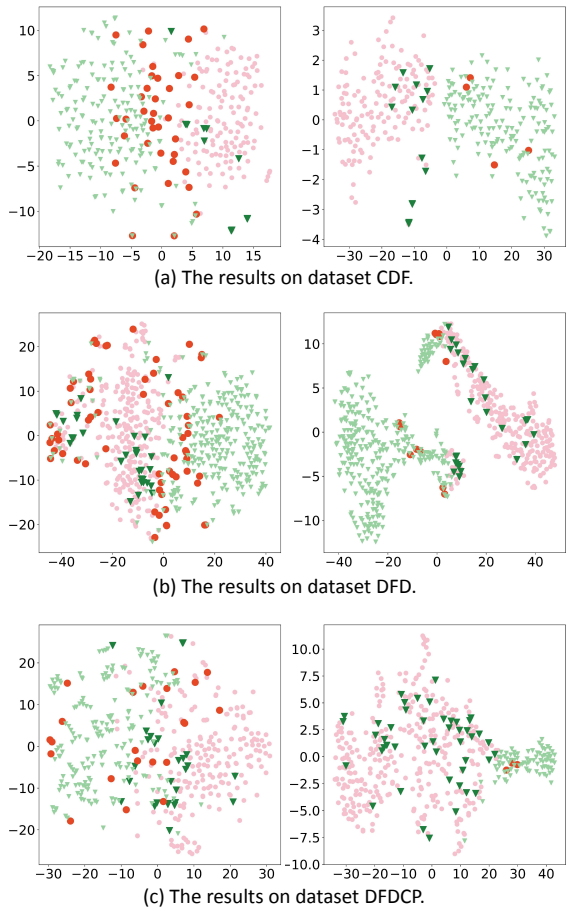

Figure 6: t-SNE visualization. **Left:** The features extracted by the baseline model. **Right:** The features extracted by our RealID model.

In Figure 6, we reported the t-SNE visualization for features in testing datasets. For each subfigure, the left side represents the distribution of test samples under SBI, while the right side shows the distribution under our RealID. We can observe that our RealID significantly reduces the misclassified real samples (marked as large red circles). For instance, a notable number of outliers are corrected on the CDF dataset.

5 Conclusion

In this work, we address the critical challenge of generalization obstacle of deepfake detection models, particularly their tendency to misclassify real instances as fake when applied to unseen datasets. We identify that this issue stems from an over-reliance on forgery artifacts and a limited understanding of 'real'. To overcome these limitations, we propose RealID, a novel approach comprising the Real Concept Capture Module and the Independent Decision Classification Module. Extensive experiments demonstrated that RealID significantly outperforms state-of-the-art baselines in cross-dataset detection scenarios. It not only advances the field of deepfake detection but also offers a solution for future research on balanced and robust detection strategies.

Acknowledgements

This project is supported in part by National Natural Science Foundation of China (Grant No. 62172256, No. 62202278, and No. 62202272); in part by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001).

References

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

[Cao *et al.*, 2022] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *CVPR*, pages 4103–4112, 2022.

[Chen *et al.*, 2022] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *CVPR*, pages 18689–18698, 2022.

[Cheng *et al.*, 2023] Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. Voice-face homogeneity tells deepfake. *ACM TOMM*, pages 1–22, 2023.

[Cheng *et al.*, 2024a] Harry Cheng, Yangyang Guo, Tianyi Wang, Liqiang Nie, and Mohan Kankanhalli. Diffusion facial forgery detection. In *ACM MM*, page 5939–5948, 2024.

[Cheng *et al.*, 2024b] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? In *NeurIPS*, 2024.

[Choi *et al.*, 2024] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake video detection. In *CVPR*, pages 1133–1143, 2024.

[Dolhansky *et al.*, 2020] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. The deepfake detection challenge dataset. *CoRR*, pages 1–13, 2020.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, pages 1–12, 2021.

[Dufour and Gully, 2020] Nick Dufour and Andrew Gully. Deepfake detection dataset, 2020.

[Fei *et al.*, 2022] Jianwei Fei, Yunshu Dai, Peipeng Yu, Tianrun Shen, Zhihua Xia, and Jian Weng. Learning second order local anomaly for general face forgery detection. In *CVPR*, pages 20238–20248, 2022.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[Guan *et al.*, 2024] Weinan Guan, Wei Wang, Jing Dong, and Bo Peng. Improving generalization of deepfake detectors by imposing gradient regularization. *IEEE TIFS*, 19:5345–5356, 2024.

[Haliassos *et al.*, 2022] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *CVPR*, pages 14930–14942, 2022.

[Hong *et al.*, 2024] Cheng-Yao Hong, Yen-Chi Hsu, and Tyng-Luh Liu. Contrastive learning for deepfake classification and localization via multi-label ranking. In *CVPR*, pages 17627–17637, 2024.

[Jia *et al.*, 2022] Shuai Jia, Chao Ma, Taiping Yao, Bangjie Yin, Shouhong Ding, and Xiaokang Yang. Exploring frequency adversarial attacks for face forgery detection. In *CVPR*, pages 4093–4102, 2022.

[Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019.

[Khalid *et al.*, 2021] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. In *NIPS*, pages 1–15, 2021.

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, pages 1–14, 2014.

[Kwon *et al.*, 2021] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. In *ICCV*, pages 10724–10733, 2021.

[Le *et al.*, 2021] Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *ICCV*, pages 10097–10107, 2021.

[Li *et al.*, 2018] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. In *WIFS*, 2018.

[Li *et al.*, 2020a] 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 *CVPR*, pages 5000–5009, 2020.

[Li *et al.*, 2020b] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *CVPR*, pages 3204–3213, 2020.

[Li *et al.*, 2023] Shuang Li, Fan Li, Jinxing Li, Huafeng Li, Bob Zhang, Dapeng Tao, and Xinbo Gao. Logical relation inference and multiview information interaction for

domain adaptation person re-identification. *IEEE TNNLS*, 2023.

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

[Masi *et al.*, 2020] 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, 2020.

[Narayan *et al.*, 2023] Kartik Narayan, Harsh Agarwal, Kartik Thakral, Surbhi Mittal, Mayank Vatsa, and Richa Singh. Df-platter: Multi-face heterogeneous deepfake dataset. In *CVPR*, pages 9739–9748, 2023.

[Ni *et al.*, 2022] Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. CORE: consistent representation learning for face forgery detection. In *CVPRW*, pages 12–21, 2022.

[Nirkin *et al.*, 2022] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on discrepancies between faces and their context. *IEEE TPAMI*, 44(10):6111–6121, 2022.

[Park *et al.*, 2020] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *CVPR*, pages 14372–14381, 2020.

[Qiao *et al.*, 2024] Tong Qiao, Shichuang Xie, Yanli Chen, Florent Retraint, and Xiangyang Luo. Fully unsupervised deepfake video detection via enhanced contrastive learning. *IEEE TPAMI*, 46(7):4654–4668, 2024.

[Rössler *et al.*, 2019] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11, 2019.

[Shiohara and Yamasaki, 2022] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *CVPR*, pages 18699–18708, 2022.

[Sun *et al.*, 2024] Zhimin Sun, Shen Chen, Taiping Yao, Ran Yi, Shouhong Ding, and Lizhuang Ma. Rethinking open-world deepfake attribution with multi-perspective sensory learning. *IJCV*, pages 1–24, 2024.

[Suwajanakorn *et al.*, 2017] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM TOG*, 36(4):95:1–95:13, 2017.

[Tan and Le, 2019] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, volume 97, pages 6105–6114, 2019.

[Tan *et al.*, 2024] Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *CVPR*, pages 28130–28139, 2024.

[Thies *et al.*, 2016] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In *CVPR*, pages 2387–2395, 2016.

[Tian *et al.*, 2024] Jiahe Tian, Peng Chen, Cai Yu, Xiaomeng Fu, Xi Wang, Jiao Dai, and Jizhong Han. Learning to discover forgery cues for face forgery detection. *IEEE TIFS*, 19:3814–3828, 2024.

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

[Wang *et al.*, 2024] Tianyi Wang, Xin Liao, Kam Pui Chow, Xiaodong Lin, and Yinglong Wang. Deepfake detection: A comprehensive survey from the reliability perspective. *ACM CSUR*, 57(3), 2024.

[Wu *et al.*, 2020] Xi Wu, Zhen Xie, YuTao Gao, and Yu Xiao. Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features. In *ICASSP*, pages 2952–2956, 2020.

[Xia *et al.*, 2024a] Ruiyang Xia, Decheng Liu, Jie Li, Lin Yuan, Nannan Wang, and Xinbo Gao. Mmnet: multi-collaboration and multi-supervision network for sequential deepfake detection. *IEEE TIFS*, 2024.

[Xia *et al.*, 2024b] Ruiyang Xia, Dawei Zhou, Decheng Liu, Lin Yuan, Shuodi Wang, Jie Li, Nannan Wang, and Xinbo Gao. Advancing generalized deepfake detector with forgery perception guidance. In *ACM MM*, pages 6676–6685, 2024.

[Xu *et al.*, 2022a] Chao Xu, Jiangning Zhang, Miao Hua, Qian He, Zili Yi, and Yong Liu. Region-aware face swapping. In *CVPR*, pages 7622–7631, 2022.

[Xu *et al.*, 2022b] Xiaogang Xu, Ying-Cong Chen, Xin Tao, and Jiaya Jia. Text-guided human image manipulation via image-text shared space. *IEEE TPAMI*, 44(10):6486–6500, 2022.

[Yan *et al.*, 2023] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. UCF: uncovering common features for generalizable deepfake detection. In *ICCV*, pages 22355–22366, 2023.

[Yan *et al.*, 2024] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *CVPR*, pages 8984–8994, 2024.

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