FakeDiffer: Distributional Disparity Learning on Differentiated Reconstruction for Face Forgery Detection

Bo Wang[1], Zhao Zhang[1, 2*], Suiyi Zhao[1], Xianming Ye[3], Haijun Zhang[4], Meng Wang[1]

[1]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
[2]Yunnan Key Laboratory of Software Engineering, Yunan, China
[3]Department of Electronic and Computer Engineering, University of Pretoria, South Africa
[4] Department of Computer Science, Harbin Institute of Technology (Shenzhen), Shenzhen, China

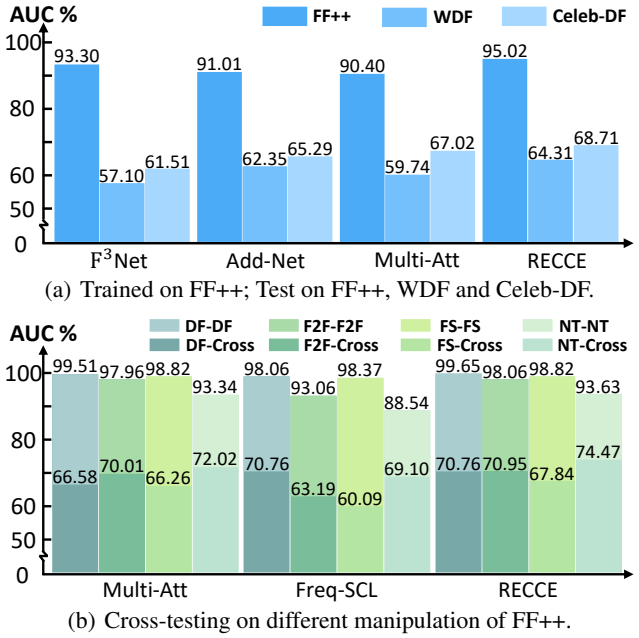
Abstract

Existing face forgery detection methods achieve promising performance when training and testing forgery data are from identical manipulation types, while they fail to generalize well to unseen samples. In this paper, we experimentally investigate and find that the poor generalization of the methods mainly arises from their overfitting on the known fake patterns. Excessively focused on seen fakes, those detectors fail to effectively learn image-intrinsic information and the distributional disparity between real and fake images. Then, to address this issue, we redefine fake learning as real-fake distributional disparity learning. We propose a novel deepfake detection framework learning distributional disparity based on the differentiated reconstruction on real and fake images for improved generalization. Specifically, distributional disparity learning on differentiated reconstruction of the real and fake images, enforces the model to learn image-invariant intrinsic representations. The reconstruction on real and fake images forces the decoders to learn the distribution of real and fake images, respectively. Moreover, to avoid the influence from the specificalization of the known fake patterns, we further propose the information interaction learning on the encoded intrinsic information and the pixel disparity between the input image and its reconstruction to distinguish face forgeries that are even unknown. Extensive experiments on large-scale benchmark datasets demonstrated the effectiveness of addressing the overfitting issue of the classification network, and verified the superior performance of our method.

Introduction

With the considerable progress in face manipulation methods (Chan et al. 2019; Gao et al. 2021; Yao et al. 2021), increasing realistic fake face images and videos are generated easily. The abuse of those face forgeries sometimes results in various pressing security concerns over fake news and impersonation (Lyu 2020). Therefore, face forgery detection, as the critical recognition task in this field, aims to develop automatic methods for identifying manipulated face images. The early-proposed face forgery detectors utilize a visual encoder (Chollet 2017) to extract the features representations from the given face image, which is followed by a classification head(Nguyen, Yamagishi, and Echizen 2019). While these feature representations are extracted in category-level, and ignore the differences between real and fake images. To address the drawbacks of category-level representations, specific forgeries are captured to distinguish fake faces from real ones (Zhao et al. 2021). Recently, inspired by the compact representations learned with real images (Ruff et al. 2020), common characteristics of genuine faces (Cao et al. 2022; Shao et al. 2023) are explored to mine the real distribution and improve the face forgery detector.

The existing face forgery detection methods, utilizing classification backbone to capture specific forgery patterns



(a) Trained on FF++; Test on FF++, WDF and Celeb-DF.

(b) Cross-testing on different manipulation of FF++.

Figure 1: Investigation of the models‘ generalization. (a) Both real images and fake types in testing forgery data are unseen. Poor generalization of the models indicates the overfitting on either real images and/or fake types. (b) Only the fake types in the testing forgery data are unseen, while the real images are seen. Compared with (a), their performance keeping matching degraded magnitude, indicates that the overfitting of models mainly focuses on the fake types.

*Corresponding author, cszzhang@gmail.com

and compact representations of real faces for binary classifying, have made great progress in recognizing fake face images (Gu et al. 2022; Huang et al. 2023). Although these methods achieve promising performance on the forgery data when training and testing forgery data are from identical manipulation types, while they fail to generalize well to unseen samples. We experimentally investigate the generalization of detection methods, and take several of them for example, which are shown in Figure 1. FaceForensics++ (FF++) (Rossler et al. 2019), WildDeepfake (WDF) (Zi et al. 2020), and Celeb-DF (Li et al. 2020b) are the widely used bentchmark datasets. FF++ contains four types of manipulation techniques, i.e., Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). Firstly, from the observation of Figure 1(a), all the models have largely declined when both real images and fake types in testing forgery data are unseen. This poor generalization indicates that **there is indeed overfitting focused on real images and/or fake types**. Secondly, we further study the models' performance, where only the fake types in the testing forgery data are unseen, while the real images are seen. Compared with the experimental results in Figure 1(a), the matching degraded magnitude of performance in Figure 1(b) indicates that the **overfitting of models mainly focuses on the fake types**. There occurs overfitting when models focus excessively on known fake patterns, thereby neglecting the underlying image-intrinsic information and the distributional disparities between real and fake images. As a result, these detectors fail to perform effectively on unseen manipulations.

In this paper, to address the above issue, we redefine fake learning as real-fake distributional disparity learning. We then propose a new deepfake detection framework (dubbed FakeDiffer) using distributional disparity learning based on the differentiated reconstruction on real and fake images to well-generalize to unseen forgery data. The existing visual representations are mined to fit the real or fake distribution of face images. Its ignoration on the image-intrinsic information makes it sensitive to the specificalization of the seen fake patterns. So we propose distributional disparity learning on the reconstruction of the real and fake image to obtain the image-invariant intrinsic representations. The reconstruction on real and fake images forces the decoders to learn the distribution of real and fake images, respectively. Furthermore, to avoid the influence from the specificalization of the known fake patterns, we further propose the information interaction learning on the encoded intrinsic information and the pixel disparity between the input image and its reconstruction to distinguish face forgery that is even unknown. Extensive experiments are conducted to prove the effectiveness of the proposed FakeDiffer. The experimental results demonstrate that our FakeDiffer can relieve the model's overfitting on seen fake patterns and remarkably improve the generalization when employed for face forgery detection.

The main contributions are summarized as follows:

- We experimentally investigate and find that the poor generalization of the existing face forgery detectors mainly arises from its overfitting on the known fake patterns. We redefine fake learning as real-fake distributional disparity learning and propose a novel detection framework using distributional disparity learning on the differentiated reconstruction for face forgery detection.
- The distributional disparity learning is proposed on the differentiated reconstruction, where the unified encoder captures the image-invariant intrinsic representation, and the decoders learn the real and fake features respectively.
- To avoid the influence of the specificalization of the known fake patterns, information interaction learning is introduced on the encoded intrinsic information and the pixel disparity between the input image and its reconstructions to comprehensively distinguish face forgery.

Related Work

Face Forgery Detection. The recent years have witnessed the considerable progress of face forgery detection methods (Li et al. 2020a; Gu et al. 2022; Yang et al. 2023). Some early approaches (Nguyen, Yamagishi, and Echizen 2019; Coccomini et al. 2022) usually utilize the image classification backbones (e.g., XceptionNet (Chollet 2017)) to extract visual features from the face images, and then feed them into a binary classifier. However, these inherent visual backbones adopted from image classifiers tend to emphasize category-level differences rather than the nuanced distinctions between real and fake face images. Recently, to address the drawbacks of category-level representations on increasingly realistic forged faces, some methods (Zhou et al. 2017; Li et al. 2021; Jeong et al. 2022; Wang et al. 2023b) are further designed to focus on specific forgery patterns, including noise statistics, local textures, and frequency information. For example, a two-stream framework (Zhou et al. 2017) is proposed to focus on visual appearance and local noise, respectively, for face forgery detection.To combine spatial information with frequency features for face forgery detection, a spatial-frequency dynamic graph method (Wang et al. 2023b) is proposed to exploit the relation-aware features in spatial and frequency domains via dynamic graph learning. These methods usually rely on the known forgery patterns appearing in the training data, which results in overfitting when models focus excessively on them, thereby neglecting the underlying image-intrinsic information and the distributional disparities between real and fake face images.

Reconstruction Learning. As a representation learning approach to feature extraction under unsupervision (Zhang et al. 2024), reconstruction learning (Han et al. 2019; Wertheimer, Tang, and Hariharan 2021; Yaman et al. 2021; Zhang et al. 2022) has been widely utilized for visual representation in various downstream tasks. The reconstruction learning method is trained to reconstruct the input from its encoded representation (Yoshihashi et al. 2019), which reinforces the model to extract meaningful and informative features that facilitate accurate reconstruction, thus obtaining robust representations. There are some prior works (Nguyen et al. 2019; Du et al. 2020; Cao et al. 2022) utilizing reconstruction learning for face forgery detection. Under sharing information settings, a reconstruction network is introduced to improve the overall performance of the multi-task learning framework, including manipulation detection and region location (Nguyen et al. 2019). After that, to explore the com-

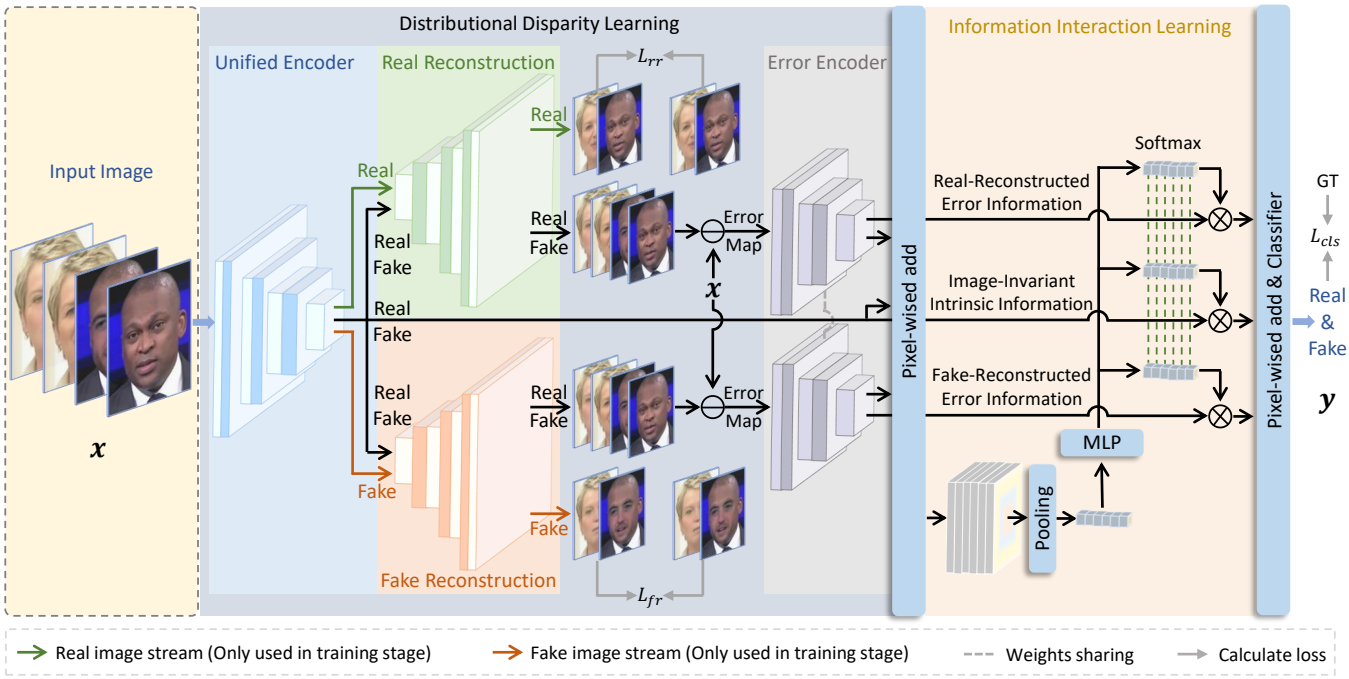


Figure 2: Overview of our proposed FakeDiffer learning distributional disparity on differentiated reconstruction for face forgery detection. Our method pipeline includes three main components, i.e., the distributional disparity learning module, information interaction learning module, and classification head. The distributional disparity learning module consists of a unified encoder, real reconstruction, fake reconstruction, and error encoder. Specifically, the distributional disparity learning module is firstly used to obtain real-reconstructed error information, image-invariant intrinsic information, and fake-reconstructed error information from the given face image, and then the above three sources of embedded features are fed into the information interaction learning module. Finally, the classification head predicts the category of input face image conditioned on the interacted features.

mon characteristics of genuine faces, RECCE (Cao et al. 2022) only reconstruct real images from their noisy versions to distinguish the forgery from the real one. These methods conduct reconstruction learning without differentiated embedding, which fails to effectively learn image-intrinsic information and the distributional disparity between real and fake images. Therefore, we explore image-intrinsic information to learn domain-invariant distribution, and integrate them with the differentiated reconstruction on real and fake images, respectively for improved generalization.

Proposed Method

To alleviate the model's overfitting focused on the known fake patterns, we redefine fake learning as real-fake distributional disparity learning and propose distributional disparity learning on differentiated reconstruction for face forgery detection. Our proposed novel framework named FakeDiffer, as shown in Figure 2, mainly consists of three components, i.e., the distributional disparity learning module, information interaction learning module, and classification head. The distributional disparity learning module includes a unified encoder, real reconstruction, fake reconstruction, and error encoder. The unified encoder is trained to encode both real and fake images to obtain intermediate representations, while the real and fake reconstruction branches decode the corresponding representation to reconstruct real and fake images, respectively. So this single encoder constrained by dual decoders in real and fake attributes, reinforces it to capture image-invariant intrinsic information, which does not bias excessively to fake patterns or real elements. Moreover, the real and fake reconstruction branches aim to correspondingly model the distributions of real and fake face images, respectively. Then, to avoid the influence of the specificalization of the known fake patterns, the above three sources of embedded features are fed into the information interaction learning module to comprehensively distinguish face forgery. Finally, the classification head predicts the category of input face image conditioned on the interacted features.

Distributional Disparity Learning

There are developing diverse types of face forgery especially in face of the increasing image generation technology. Thus it leads to overfitting when forgery detectors focus excessively on known fake patterns, thereby neglecting the underlying image-intrinsic information and the distributional disparities between real and fake images. As such, distributional disparity learning is proposed by a single unified encoder for visual representation and dual reconstruction branches for real and fake reconstruction, respectively. Moreover, the single encoder constrained by dual decoders

in real and fake attributes, reinforces it to capture image-invariant intrinsic information, which does not bias excessively to fake patterns or real elements.

Specifically, given an input image $\mathbf{X} \in \mathbb{R}^{h\times w\times 3}$, a reconstruction network $\mathcal{F}$ is trained to model the real image distribution, fake image distribution, and image-invariant intrinsic distribution. The reconstruction network $\mathcal{F}$ consists of unified encoder $\mathcal{F}_e$, real-reconstruction branch $\mathcal{F}_r$, and fake-reconstruction branch $\mathcal{F}_f$. To learn robust representations for given images, we follow the prior work settings (Zhou et al. 2021) by adding some white noises to the input images in the training stage to get $\widetilde{\mathbf{X}}$. We use XceptionNet (Chollet 2017) as the backbone, and use the bilinear interpolation to adjust the spatial size properly for mentioned operations. The input images $\widetilde{\mathbf{X}}$ can be split into real images $\widetilde{\mathbf{X}}_r$ and fake images $\widetilde{\mathbf{X}}_f$. Thus, as shown in the middle section of Figure 2, the unified encoding process can be formulated as:

$$\mathbf{M} = \mathcal{F}_e(\widetilde{\mathbf{X}}), \tag{1}$$

where $\mathbf{M} = \{\mathbf{M}_r, \mathbf{M}_f\}$ denotes the intermediate representation of given images, which can be regarded as image-invariant intrinsic representations under the constraints of dual decoders in real and fake reconstructed attributes, which does not bias excessively to real or fake distribution. $\mathbf{M}_r, \mathbf{M}_f$ denote the corresponding intermediate representation of real and fake images, respectively.

Then, in the training stage, the real and fake image reconstruction process can be respectively formulated as:

$$\widehat{\mathbf{X}}_r = \mathcal{F}_r(\mathcal{F}_e(\widetilde{\mathbf{X}}_r)) = \mathcal{F}_r(\mathbf{M}_r), \tag{2}$$

$$\widehat{\mathbf{X}}_f = \mathcal{F}_f(\mathcal{F}_e(\widetilde{\mathbf{X}}_f)) = \mathcal{F}_f(\mathbf{M}_f), \tag{3}$$

where $\widehat{\mathbf{X}}_r$ and $\widehat{\mathbf{X}}_f$ denote the real and fake image reconstruction versions, respectively. The previous work (Cao et al. 2022) points out that the reconstructed forged faces largely differ from the input forged faces in visual appearance, thus we follow it to use the reconstructed error map to indicate the probably manipulated traces. Then, we can obtain the encoded-reconstructed error maps (i.e., $\mathbf{E}_r$ and $\mathbf{E}_f$) on the real-reconstruction path and fake-reconstruction path, respectively, as follows:

$$\mathbf{E}_r = f(|\mathcal{F}_r(\mathbf{M}) - \mathbf{X}|), \tag{4}$$

$$\mathbf{E}_f = f(|\mathcal{F}_f(\mathbf{M}) - \mathbf{X}|), \tag{5}$$

where f represents the convolutional layers-based error map encoder, whose weights are shared in the above dual paths.

Thus, we achieve the differentiated reconstruction of real and fake images, where the dual reconstruction branches are constrained to learn real and fake content, respectively. The unified encoder constrained by dual decoders of real and fake attributes, captures image-invariant intrinsic information without excessive bias to the single fake patterns or real elements. The real-reconstructed error information, image-invariant intrinsic information, and fake-reconstructed error information are captured from the given face images by the distributional disparity learning module.

Information Interaction Learning

To avoid the influence of the specificalization of the known fake patterns, we further propose the information interaction learning on the encoded intrinsic information and the pixel disparity between the input image and its reconstruction to distinguish face forgery. Specifically, we obtain the fused feature maps (denoted by $\mathbf{K} \in \mathbb{R}^{H'\times W'\times C}$) of three sources of representations (i.e., $\mathbf{M}$, $\mathbf{E}_r$, and $\mathbf{E}_f$) by pixel-wised addition operation. As shown in the right section of Figure 2, we embed the global information of $\mathbf{K}$ by average pooling operation, which can be formulated as:

$$g_c = \frac{1}{H'\times W'}\sum_{i=1}^{H'}\sum_{j=1}^{W'}\mathbf{K}_c(i,j), \tag{6}$$

where g_c denotes the value in c-th channel of global feature maps $\mathbf{G} \in \mathbb{R}^{C\times 1}$. Then, the interacted weights of real-reconstructed error information, fake-reconstructed error information and image-invariant intrinsic information (i.e., $\mathbf{M}$, $\mathbf{E}_r$, and $\mathbf{E}_f$) are calculated by a multilayer perceptron followed by softmax operation in the split dimension. The interaction process can be formulated as:

$$\widehat{\mathbf{M}} = \mathbf{M} + softmax\{split[mlp(\mathbf{G})]\}\cdot\mathbf{M}, \tag{7}$$

where $\widehat{\mathbf{M}}$ denotes the weighted image-invariant intrinsic feature maps. The $mlp(\cdot)$ denotes multi-layer perceptron. Then the output values are split equally into three segments by $split(\cdot)$ function. $softmax(\cdot)$ denotes the softmax function for the corresponding weight calculation. The interacted features of $\mathbf{E}_r$ and $\mathbf{E}_f$ can be calculated in the same reasoning process, which is denoted by $\widehat{\mathbf{E}_r}$ and $\widehat{\mathbf{E}_f}$, respectively.

After that, we fuse the interacted feature maps to generate the comprehensively distinguish feature maps $\mathbf{T}$ by pixel-wised add operation as follows:

$$\mathbf{T} = \widehat{\mathbf{M}} + \widehat{\mathbf{E}_r} + \widehat{\mathbf{E}_f}, \tag{8}$$

Finally, as shown in the stern of Figure 2, the classification head predicts the category of input face image conditioned on the interacted features $\mathbf{T}$.

Objective Function

During the reconstruction process, we compute the reconstruction loss $\mathcal{L}_{rr}$ between input real images and their reconstructed versions in a mini-batch as:

$$\mathcal{L}_{rr} = \frac{1}{|R|}\sum_{i\in R}\|\widehat{\mathbf{X}}_i - \mathbf{X}_i\|_1, \tag{9}$$

where R denotes the set of real samples in a mini-batch and $|R|$ is the cardinality of R. We compute the reconstruction loss $\mathcal{L}_{fr}$ between input fake face images and their reconstructed versions in a mini-batch formally as:

$$\mathcal{L}_{fr} = \frac{1}{|F|}\sum_{i\in F}\|\widehat{\mathbf{X}}_i - \mathbf{X}_i\|_1, \tag{10}$$

where F denotes the set of fake samples in a mini-batch and $|F|$ is the cardinality of F. Besides, the model also optimized using the BCE loss for the final classification:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i\log p_i + (1-y_i)\log(1-p_i)\right), \tag{11}$$

Methods	FF++ (c23)		FF++ (c40)		Celeb-DF		WildDeepfake	
	Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC(%)
MesoNet (Afchar et al. 2018)	83.10	-	70.47	-	-	-	64.47	-
Multi-task (Nguyen et al. 2019)	85.65	85.43	81.30	75.59	-	-	-	-
Xception (Rossler et al. 2019)	95.73	96.30	86.86	89.30	97.90	99.73	77.25	86.76
Face X-ray (Li et al. 2020a)	-	87.40	-	61.60	-	-	-	-
Two-branch (Masi et al. 2020)	96.43	98.70	86.34	86.59	-	-	-	-
RFM (Wang and Deng 2021)	95.69	98.79	87.06	89.83	97.96	99.94	77.38	83.92
Freq-SCL (Li et al. 2021)	96.69	99.28	89.00	92.39	-	-	-	-
MultiAtt (Zhao et al. 2021)	97.60	99.29	88.69	90.40	97.92	99.94	82.86	90.71
Lisiam (Wang, Sun, and Tang 2022)	96.51	99.13	87.81	91.44	-	-	-	-
RECCE (Cao et al. 2022)	97.06	99.32	91.03	95.02	98.59	99.94	83.25	92.02
SIA (Sun et al. 2022)	97.64	99.35	90.23	93.45	-	-	-	-
F^2Trans-B (Miao et al. 2023)	96.60	99.24	87.20	89.91	-	-	-	-
CFM (Luo et al. 2023)	96.93	99.25	<u>93.29</u>	**96.97**	-	-	-	-
DisGRL (Shi et al. 2023)	97.69	<u>99.48</u>	91.27	95.19	98.71	99.91	84.53	<u>93.27</u>
ATSC (Liu et al. 2023)	<u>97.90</u>	**99.52**	91.96	94.54	-	-	-	-
UniAttack (Cao et al. 2024)	97.63	99.44	92.31	96.12	<u>99.24</u>	**99.96**	<u>84.63</u>	92.11
FakeDiffer (ours)	**98.04**	**99.52**	**93.37**	<u>96.54</u>	**99.35**	<u>99.95</u>	**84.76**	**93.31**

Table 1: Intra-testing comparisons. The proposed method performs favorably over current state-of-the-art methods.

where p_i is the predicted score obtained by the binary classifier. The label y_i is 0 for real faces, otherwise y_i is 1.

The total objective function $\mathcal{L}$ of our proposed FakeDiffer includes the real reconstruction loss $\mathcal{L}_{rr}$, fake reconstruction loss $\mathcal{L}_{fr}$, and the cross-entropy loss $\mathcal{L}_{cls}$ for classification:

$$\mathcal{L} = \lambda_r \mathcal{L}_{rr} + \lambda_f \mathcal{L}_{fr} + \mathcal{L}_{cls}, \tag{12}$$

where λ_r and λ_f are weight hyperparameters to balance the reconstruction loss $\mathcal{L}_{rr}$ and $\mathcal{L}_{fr}$, respectively.

Experiments

Experimental Settings

Datasets. In this paper, we evaluate our proposed FakeDiffer and existing methods on standard benchmark datasets FaceForensics++ (FF++) (Rossler et al. 2019), Celeb-DF (CDF) (Li et al. 2020b) and WildDeepfake (WDF) (Zi et al. 2020). **FF++** is the most widely used benchmark dataset, which includes 1,000 real videos and 4,000 fake videos and consists of four types of manipulation techniques, i.e., Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). Besides, each video in FF++ has three quality levels: raw, high-quality (C23), and low-quality (C40) data. In this paper, we follow the existing methods (Zhao et al. 2021; Cao et al. 2022) to consider the C23 and C40 versions for accommodating practical applications. **Celeb-DF** includes 590 real videos and 5,639 high-quality fake videos, which are crafted by the improved DeepFake algorithm (Li et al. 2020b). The fake videos in CDF have better visual quality than previous datasets, making them more challenging for detection. **WildDeepfake** is a real-world dataset that contains 3,805 real sequences and 3,509 fake sequences. All the videos are collected from the internet with diverse scenes and forgery methods. The evaluation results on WDF reflect the detector's performance in real-world scenarios.

Evaluation Metrics. For a fair comparison, the predicted results of each method are evaluated on three widely-used metrics (Huang et al. 2023), i.e., Accuracy (Acc), Area Under the Receiver Operating Characteristic Curve (AUC), and Equal Error Rate (EER). All values of those metrics are reported as percentages (%), and the larger the values of Acc and AUC the better the performance, while the smaller the value of EER the better the performance of methods.

Implementation Details. The backbone of FakeDiffer is the Xception (Chollet 2017). We train it under the total loss (consists of mean square error loss for reconstruction and binary cross-entropy loss for classification) for 30 epochs with a batch size of 32, and an Adam optimizer whose learning rate is initialized at 2e-4 with the warmup step of 10,000. The weight decay is set as 1e-5.λ_r and λ_f in Equation 12 are empirically set to 0.1. We only use random horizontal flipping for data augmentation. The default random seed is set to 42. All experiments are conducted in a single NVIDIA GeForce RTX 4090 GPU with Pytorch 1.11 platform.

Main Results

Intra-testing. As shown in Table 1, for the FF++ dataset, our proposed FakeDiffer model achieves consistent improvements under different quality settings (i.e., c23 and c40 denote high- and low-quality compresses, respectively). Especially, over-compression destroys the frequency clues that F^3-Net relies upon, while our approach yields a more robust representation through differentiated reconstruction learning

Methods	Celeb-DF		WildDeepfake	
	AUC ↑	EER ↓	AUC ↑	EER ↓
Xception	61.80	41.73	62.72	40.65
RFM	65.63	38.54	57.75	45.45
Add-Net	65.29	38.90	62.35	41.42
F^3-Net	61.51	42.03	57.10	45.12
MultiAtt	67.02	37.90	59.74	43.73
RECCE	68.71	35.73	64.31	40.53
FakeDiffer (ours)	**69.24**	**35.53**	**68.46**	**37.74**

Table 2: Cross-testing (AUC and EER) by training on FF++.

that captures effective image-invariant intrinsic information, and real- and fake-reconstruction information for forgery classification. Thus, on the challenging C40 setting, compared with F^3-Net(Miao et al. 2023), the AUC score of our method exceeds it by 7.37%. From the observation, although CFM(Luo et al. 2023) achieves the highest AUC on FF++ C40, our FakeDiffer model based on the backbone of Xception still achieves comparable results and exceeds it by a large margin on the high-quality setting. Different from the existing methods (e.g., Multi-task methods) which employ reconstruction constraints for both real and fake faces by the unified reconstruction decoder, our proposed FakeDiffer models the distribution of real and fake by two separate branches to comprehensively learn the distributional disparity. Thus, our method significantly outperforms the counterpart. Moreover, the performance gains can also be observed on Celeb-DF and especially the realistic dataset WildDeepfake, while in the latter our method reaches a state-of-the-art result both in the Acc and the AUC metrics. The above results demonstrate the effectiveness of our FakeDiffer.

Cross-testing. We further evaluate the generalization ability of our FakeDiffer on unknown forgery patterns, we conduct cross-dataset experiments by training and testing on different datasets. Specifically, we train the models on FF++ C40, and then test them on Celeb-DF, and WildeDeepfake, respectively. The results are shown in Table 2. From the table, we observe that our FakeDiffer generally outperforms all listed methods on the test data with unseen forgery patterns, often by a large margin. For instance, when testing on the WildDeepfake dataset, the AUC score of most previous methods drop to around 65%. Differently, our FakeDiffer reaches an AUC of 68.46%, which exceeds RECCE (Cao et al. 2022) by 6.45%. The performance mainly benefits from the proposed distributional disparity learning on differentiated reconstruction, which models the distribution of real and fake faces in separate branches. The image-invariant intrinsic information is integrated with real- and fake-reconstructed error information to achieve robust representations for face forgery detection. Instead of overfitting with specific forged patterns as in existing methods, our method learns image-invariant intrinsic information and the distributional disparity of real and fake images, so as to achieve better generalizability.

Methods	Train	DF	F2F	FS	NT	C-Avg.
Freq-SCL	DF	98.91	58.90	66.87	63.61	63.13
MultiAtt		99.51	66.41	67.33	66.01	66.58
RECCE		99.65	70.66	74.29	67.34	70.76
FakeDiffer (ours)		**99.73**	**71.27**	**75.75**	**68.98**	**72.00**
Freq-SCL	F2F	67.55	93.06	55.35	66.66	63.19
MultiAtt		73.04	97.96	65.10	71.88	70.01
RECCE		75.99	98.06	64.53	72.32	70.95
FakeDiffer (ours)		**76.50**	**98.96**	**67.10**	**73.96**	**72.52**
Freq-SCL	FS	75.90	54.64	98.37	49.72	60.09
MultiAtt		82.33	61.65	98.82	54.79	66.26
RECCE		82.39	64.44	98.82	56.70	67.84
FakeDiffer (ours)		**83.02**	**64.7**	**99.73**	**57.85**	**68.52**
Freq-SCL	NT	79.09	74.21	53.99	88.54	69.10
MultiAtt		74.56	80.61	60.90	93.34	72.02
RECCE		78.83	80.89	63.70	93.63	74.47
FakeDiffer (ours)		**82.06**	**81.70**	**66.23**	**97.90**	**76.66**

Table 3: Cross-testing on different manipulation techniques. The values underlined means within-dataset testing results.

To evaluate the generalization ability of our FakeDiffer more comprehensively, we further conduct fine-grained cross-testing by training on the mentioned four specific manipulation techniques and testing on the others in FF++ C40. We compare our FakeDiffer with other methods focusing on specific forgery patterns, i.e., Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT), which is shown in Table 3. We highlight the results within the dataset by being remarked with the grey region. The average cross-testing results are denoted as C-Avg for short. Our FakeDiffer generally outperforms others on unseen forgery patterns, which verifies that it is feasible to explore image-invariant intrinsic information and model rea-fake distributional disparity for distinguishing unknown types of forgeries.

Ablation Study

In this section, we conduct the ablation study on different components proposed in our framework to evaluate the effectiveness of the image-invariant intrinsic information, rea-fake distributional disparity (as shown in Table 4) and the information interaction learning mode (as shown in Table 5).

Effectiveness of the proposed Distributional Disparity Learning. In our FakeDiffer framework, the unified encoder in the reconstruction network is trained to encode both real and fake images to obtain intermediate representations, while the real and fake reconstruction branches decode the corresponding representation to reconstruct real and fake images, respectively. As such the FakeDiffer captures image-invariant intrinsic information (denoted by "inva") by the Xception encoder, real distributional information (denoted by "real-inf") by the real reconstruction branch, and fake distributional information (denoted by "fake-inf") by fake reconstruction branch. From the observation of Table 4, it has poor generalization formance using only Xception representation, while reconstruction-based real and fake dis-

Train	F2F FS NT				DF FS NT				FF++ (C23)			
Test	DF (C23)		DF (C40)		F2F (C23)		F2F (C40)		Celeb-DF		WildDeepfake	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
real-inf	93.30	94.06	90.01	93.35	94.63	95.84	92.31	94.8	71.66	74.57	70.52	75.65
fake-inf	92.78	95.03	90.46	94.14	95.14	94.46	93.60	95.42	70.34	74.38	69.09	76.07
Xception	90.45	93.35	88.03	90.11	94.75	95.60	92.60	95.09	74.95	75.56	69.11	79.74
dual recons	94.34	96.72	89.31	93.14	94.03	96.18	90.08	92.18	73.27	74.16	68.39	78.64
FakeDiffer	**96.51**	**98.47**	**95.44**	**96.85**	**97.71**	**98.36**	**95.72**	**97.55**	**76.59**	**80.46**	**74.87**	**81.67**

Table 4: Ablation of information sources under two settings: 1) Cross-test within FF++ (left); 2) From FF++ to others (right).

Train	F2F FS NT				DF FS NT				FF++ (C23)			
Test	DF (C23)		DF (C40)		F2F (C23)		F2F (C40)		Celeb-DF		WildDeepfake	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
Add (real-inf, inva)	96.13	97.69	94.87	95.91	97.02	97.54	95.10	96.67	75.89	79.32	73.91	80.07
Add (fake-inf, inva)	95.62	97.54	93.30	95.45	96.44	96.90	94.60	95.83	75.22	78.69	73.35	78.86
Add (real-inf, fake-inf)	94.76	96.14	92.27	93.17	94.93	95.44	92.74	94.89	73.92	77.06	71.47	77.36
Add (real-inf, fake-inf, inva)	96.20	97.50	94.63	95.94	97.35	97.67	94.71	96.25	75.80	79.28	73.72	79.94
inter (real-inf, inva)	96.33	97.74	95.34	96.13	97.46	97.82	95.53	96.92	76.14	79.86	74.37	80.74
inter (fake-inf, inva)	95.87	97.62	94.40	96.01	97.49	97.73	95.34	96.99	76.09	79.55	74.13	80.36
inter (real-inf, fake-inf)	94.45	96.21	92.67	94.03	95.14	95.68	92.80	95.09	74.28	77.31	71.64	77.42
FakeDiffer	**96.51**	**98.47**	**95.44**	**96.85**	**97.71**	**98.36**	**95.72**	**97.55**	**76.59**	**80.46**	**74.87**	**81.67**

Table 5: Ablation of information fusion under two settings: 1) Cross-test within FF++ (left); 2) From FF++ to others (right).

tributional information achieve similar improvements when the model is generalized to other patterns of forgery. To evaluate the effectiveness of differentiated reconstruction learning, we also reconstruct the face images without splitting the real and fake categories, which is denoted by the "dual recons" in Table 4. The outperformance of our FakeDiffer indicates the improvements benefiting from distributional disparity learning on differentiated reconstruction.

Effectiveness of the proposed Information Interaction Learning. To evaluate the effectiveness of the proposed information interaction learning, we integrate the captured image-invariant intrinsic information with real- and fake distributional information in two modes, i.e., adding operation and our interaction learning (denoted by "inter"). As shown in Table 5, it can be clearly observed that: 1) image-invariant intrinsic information has a more robust representation; 2) the interaction learning provides a more effective information fusing method for face forgery detection.

Conclusion

In this paper, we experimentally investigate and find that the poor generalization of the existing face forgery detectors mainly arises from its overfitting on known fake patterns. To address this issue, we redefine fake learning as real-fake distributional disparity learning and propose a novel deepfake detection framework using distributional disparity learning based on the differentiated reconstruction on real and fake images for improved generalization to unseen forgery data. The single unified encoder constrained by the differentiated reconstructors in real and fake attributes, can reinforce it to capture image-invariant intrinsic information, which does not bias excessively to fake patterns or real elements. Moreover, the real and fake reconstruction branches can correspondingly model the distributions of real and fake face images, respectively. Finally, information interaction learning is proposed to effectively combine the above three sources of information for comprehensively distinguishing face forgery even unseen. Extensive cross-domain experiments demonstrate the better generalization of our method. In the future, we will explore more robust and efficient domain-invariant representation methods (e.g., using vision and language (Wang et al. 2023a, 2024) for robust semantic detection) for face forgery detection in open-world applications.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62472137, 62072151), Anhui Provincial Natural Science Fund for the Distinguished Young Scholars (2008085J30), Open Foundation of Yunnan Key Laboratory of Software Engineering (2023SE103), CCF-Baidu Open Fund (CCF-BAIDU202321) and CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2022-057A). Zhao Zhang is the corresponding author of this paper.

References

Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. Mesonet: a compact facial video forgery detection network. In *Proceedings of the IEEE International Workshop on Information Forensics and Security*.

Cao, J.; Ma, C.; Yao, T.; Chen, S.; Ding, S.; and Yang, X. 2022. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Cao, J.; Zhang, K.-Y.; Yao, T.; Ding, S.; Yang, X.; and Ma, C. 2024. Towards Unified Defense for Face Forgery and Spoofing Attacks via Dual Space Reconstruction Learning. *International Journal of Computer Vision*.

Chan, C.; Ginosar, S.; Zhou, T.; and Efros, A. A. 2019. Everybody dance now. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Coccomini, D. A.; Messina, N.; Gennaro, C.; and Falchi, F. 2022. Combining efficientnet and vision transformers for video deepfake detection. In *Proceedings of the International Conference on Image Analysis and Processing*.

Du, M.; Pentyala, S.; Li, Y.; and Hu, X. 2020. Towards generalizable deepfake detection with locality-aware autoencoder. In *Proceedings of the ACM International Conference on Information & Knowledge Management*.

Gao, Y.; Wei, F.; Bao, J.; Gu, S.; Chen, D.; Wen, F.; and Lian, Z. 2021. High-fidelity and arbitrary face editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Gu, Q.; Chen, S.; Yao, T.; Chen, Y.; Ding, S.; and Yi, R. 2022. Exploiting fine-grained face forgery clues via progressive enhancement learning. In *Proceedings of the Association for the Advancement of Artificial Intelligence*.

Han, Z.; Wang, X.; Liu, Y.-S.; and Zwicker, M. 2019. Multi-angle point cloud-VAE: Unsupervised feature learning for 3D point clouds from multiple angles by joint self-reconstruction and half-to-half prediction. In *Proceedings of the IEEE International Conference on Computer Vision*.

Huang, B.; Wang, Z.; Yang, J.; Ai, J.; Zou, Q.; Wang, Q.; and Ye, D. 2023. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Jeong, Y.; Kim, D.; Ro, Y.; and Choi, J. 2022. Frepgan: robust deepfake detection using frequency-level perturbations. In *Proceedings of the Association for the Advancement of Artificial Intelligence*.

Li, J.; Xie, H.; Li, J.; Wang, Z.; and Zhang, Y. 2021. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; and Guo, B. 2020a. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020b. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, J.; Xie, J.; Wang, Y.; and Zha, Z.-J. 2023. Adaptive Texture and Spectrum Clue Mining for Generalizable Face Forgery Detection. *IEEE Transactions on Information Forensics and Security*.

Luo, A.; Kong, C.; Huang, J.; Hu, Y.; Kang, X.; and Kot, A. C. 2023. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security*.

Lyu, S. 2020. Deepfake detection: Current challenges and next steps. In *Proceedings of the IEEE international conference on multimedia & expo workshops*.

Masi, I.; Killekar, A.; Mascarenhas, R. M.; Gurudatt, S. P.; and AbdAlmageed, W. 2020. Two-branch recurrent network for isolating deepfakes in videos. In *Proceedings of the European Conference on Computer Vision*.

Miao, C.; Tan, Z.; Chu, Q.; Liu, H.; Hu, H.; and Yu, N. 2023. F 2 trans: High-frequency fine-grained transformer for face forgery detection. *IEEE Transactions on Information Forensics and Security*.

Nguyen, H. H.; Fang, F.; Yamagishi, J.; and Echizen, I. 2019. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *Proceedings of the IEEE International Conference on Biometrics Theory, Applications and Systems*.

Nguyen, H. H.; Yamagishi, J.; and Echizen, I. 2019. Capsule-forensics: Using capsule networks to detect forged images and videos. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*.

Ruff, L.; Vandermeulen, R. A.; Görnitz, N.; Binder, A.; Müller, E.; Müller, K.-R.; and Kloft, M. 2020. Deep semi-supervised anomaly detection. In *Proceedings of the International Conference on Learning Representations*.

Shao, R.; Wu, T.; Nie, L.; and Liu, Z. 2023. Deepfake-adapter: Dual-level adapter for deepfake detection. *arXiv preprint arXiv:2306.00863*.

Shi, Z.; Chen, H.; Chen, L.; and Zhang, D. 2023. Discrepancy-guided reconstruction learning for image forgery detection. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Sun, K.; Liu, H.; Yao, T.; Sun, X.; Chen, S.; Ding, S.; and Ji, R. 2022. An information theoretic approach for attention-driven face forgery detection. In *Proceedings of the European Conference on Computer Vision*.

Wang, B.; Zhang, Z.; Zhao, M.; Jin, X.; Xu, M.; and Wang, M. 2024. OSIC: A new one-stage image captioner coined. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Wang, B.; Zhang, Z.; Zhao, S.; Zhang, H.; Hong, R.; and Wang, M. 2023a. CropCap: Embedding Visual Cross-Partition Dependency for Image Captioning. In *Proceedings of the ACM International Conference on Multimedia*.

Wang, C.; and Deng, W. 2021. Representative forgery mining for fake face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, J.; Sun, Y.; and Tang, J. 2022. LiSiam: Localization invariance Siamese network for deepfake detection. *IEEE Transactions on Information Forensics and Security*.

Wang, Y.; Yu, K.; Chen, C.; Hu, X.; and Peng, S. 2023b. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wertheimer, D.; Tang, L.; and Hariharan, B. 2021. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yaman, B.; Shenoy, C.; Deng, Z.; Moeller, S.; El-Rewaidy, H.; Nezafat, R.; and Akçakaya, M. 2021. Self-supervised physics-guided deep learning reconstruction for high-resolution 3D LGE CMR. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*.

Yang, Z.; Liang, J.; Xu, Y.; Zhang, X.-Y.; and He, R. 2023. Masked relation learning for deepfake detection. *IEEE Transactions on Information Forensics and Security*.

Yao, G.; Yuan, Y.; Shao, T.; Li, S.; Liu, S.; Liu, Y.; Wang, M.; and Zhou, K. 2021. One-shot face reenactment using appearance adaptive normalization. In *Proceedings of the Association for the Advancement of Artificial Intelligence*.

Yoshihashi, R.; Shao, W.; Kawakami, R.; You, S.; Iida, M.; and Naemura, T. 2019. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, Z.; Zhao, S.; Jin, X.; Xu, M.; Yang, Y.; Yan, S.; and Wang, M. 2024. Noise self-regression: A new learning paradigm to enhance low-light images without task-related data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, Z.; Zheng, H.; Hong, R.; Xu, M.; Yan, S.; and Wang, M. 2022. Deep color consistent network for low-light image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; and Yu, N. 2021. Multi-attentional deepfake detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhou, H.-Y.; Lu, C.; Yang, S.; Han, X.; and Yu, Y. 2021. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *Proceedings of the IEEE International Conference on Computer Vision*.

Zhou, P.; Han, X.; Morariu, V. I.; and Davis, L. S. 2017. Two-stream neural networks for tampered face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zi, B.; Chang, M.; Chen, J.; Ma, X.; and Jiang, Y.-G. 2020. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the ACM International Conference on Multimedia*.