Critical Forgetting-Based Multi-Scale Disentanglement for Deepfake Detection

Kai Li[1], Wenqi Ren[2*], Jianshu Li[3], Wei Wang[2*], Xiaochun Cao[2],

[1]School of Computer Science and Engineering, Sun Yat-sen University
[2]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
[3]National University of Singapore
likai63@mail2.sysu.edu.cn, {renwq3, wangwei29, caoxiaochun}@mail.sysu.edu.cn, jianshu.l@antgroup.com

Abstract

Recent face forgery detection methods based on disentangled representation learning utilize paired images for cross-reconstruction, aiming to extract forgery-relevant attributes and forgery-irrelevant content. However, there still exist the following issues that may comprise the detector performance: i) using information-dense images as the decoupling targets increases the decoupling difficulty; ii) the extracted attribute features are reconstruction-irrelevant rather than forgery-relevant, and single-scale forgery representation decoupling cannot capture sufficient discriminative information; iii) the generalization performance of decoupled attribute features is poor as the detector focuses on learning specific artifact types in the training set. To address these issues, we propose a novel disentangled representation learning framework for deepfake detection. First, we extract features by partitioning the dense information within the image, focusing independently on texture, color, or edges. These features are then used as the decoupling targets rather than the images themselves, which could mitigate the decoupling difficulty. Second, we extend reconstruction loss from image-level to feature-level, thus extending the forgery representation decoupling from single-scale to multi-scale. Third, we propose a critical forgetting mechanism that forces the detector to forget the most salient features during training, which correspond to specific forgery artifact types in the training set. Extensive experimental results validate the efficacy of the proposed method.

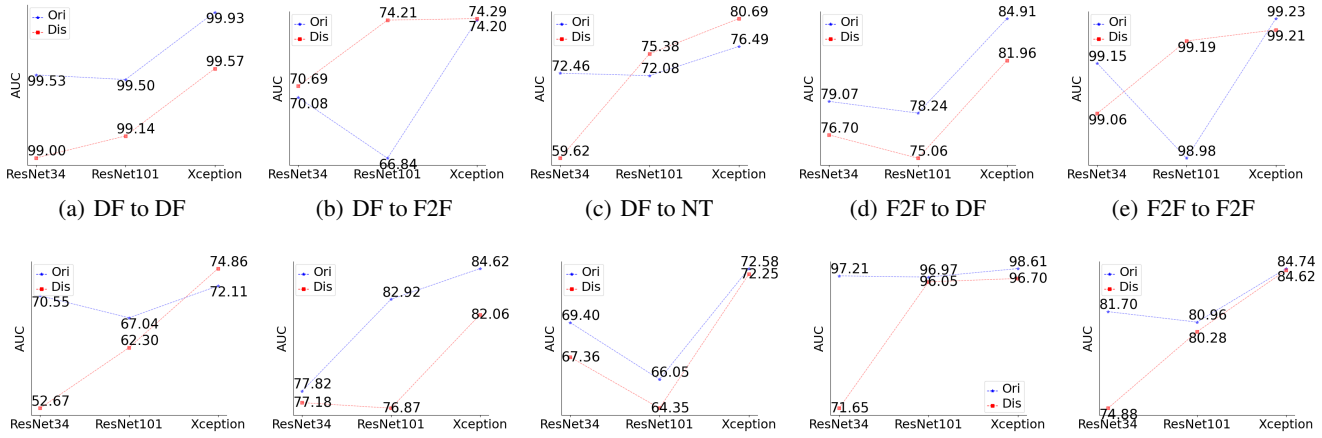
Introduction

Recent efforts (Liang, Shi, and Deng 2022; Yan et al. 2023) have been made to tackle the generalization challenge, which is caused by the discrepancy between training and test datasets, by employing disentangled representation learning. They first extract attribute and content features for a pair of real and fake images, respectively. Then, an image is cross-reconstructed by using its content feature and the attribute feature of another image, ensuring that the attribute features are forgery-relevant while content features are forgery-irrelevant. However, some issues still exist that need to be addressed. First, utilizing information-dense images as the decoupling targets increases the decoupling difficulty. Second, these methods are built on a strong assumption that the attribute features are forgery-relevant as long as they do not impact the cross-reconstruction results. However, DCI (Liang, Shi, and Deng 2022) indicates that this assumption does not hold in the intra-dataset scenario, and the attribute features extracted by cross-reconstruction degrade the performance of the detector. We argue that this assumption also does not hold in the cross-dataset scenario. As shown in Figure 1, we compare the Area Under Curve (AUC) metric of the original backbone (Ori) and that of the cross-reconstruction decoupling framework (Dis). AUC measures the model performance of binary classification, with higher values indicating better performance. It can be observed that it is difficult to guarantee that "Dis" is better than "Ori", whether in the intra-dataset (a, e, f) or cross-dataset (b, c, d, f, g, h) scenarios. Moreover, we conduct comparison experiments about the Equal Error Rate metric in the supplementary material. Third, the attribute features tend to overly capture specific artifacts in the training set, resulting in suboptimal generalization.

To address the above issues, we propose a novel deepfake detection method with enhanced disentanglement capability. Our method is based on two key observations: i) It is difficult to isolate attribute features that do not contain forgery-irrelevant information at all, but it is achievable to avoid forgery-irrelevant information dominating attribute features. ii) The specific artifacts in the training set are more readily captured, thus exerting a more pronounced effect during the training phase than general forgery features. Specifically, to alleviate the decoupling difficulty, we employ an encoder to transform the image into feature maps. Each map represents a specific facial aspect, like texture, color, or edges. Then, based on our key observation i), we introduce the Pearson coefficient in the decoupling stage, considering both direction and magnitude to further separate forgery-relevant attributes from forgery-irrelevant content. This makes it easier for the decoupler to isolate the expected information. The classification loss and Pearson loss force the forgery-relevant information to dominate the attribute features, and the forgery-irrelevant information gradually shifts to the content features.

Further, we extend the single-scale forgery representation decoupling to a multi-scale one. Deep InfoMax (Hjelm et al. 2019) indicates that a complete representation of an image is not necessary for classification tasks, e.g., we only focus on

*Corresponding authors: Wenqi Ren, Wei Wang.



(a) DF to DF (b) DF to F2F (c) DF to NT (d) F2F to DF (e) F2F to F2F

(f) F2F to NT (g) NT to DF (h) NT to F2F (i) NT to NT (j) Average

Figure 1: Comparison of the AUC metric between the cross-reconstruction decoupling framework (red) and the backbone (blue). The subtitle of "A to B" represents that the model is trained on "A" and tested on "B".

the most salient features to detect counterfeit money without learning to reconstruct the entire money. Inspired by this, we impose different constraints on decoupled features at different scales rather than only image-level reconstruction loss.

To prevent overfitting of the decoupled forgery-relevant attributes to specific artifacts in the training set, we propose a critical forgetting mechanism based on our key observation ii). Inspired by Representative Forgery Mining (Wang and Deng 2021), which forces the detector to mine previously neglected regions by generating image-level masks to track and mask the Top-N sensitive regions, the critical forgetting mechanism selectively forgets the channel information that has the greatest impact on decision-making at the channel level before the attribute features enter the classifier. These channels correspond to specific types of artifacts. In particular, we quantify the impact of different channels with their gradients during the backpropagation process: the larger the gradient, the greater the impact. By masking the channels whose gradients exceed a specific threshold, the detector is forced to devote more attention to previously overlooked channels, thereby enhancing its generalization ability. Notably, to alleviate the possible intra-dataset performance degradation due to excessive forgetting, we concatenate the attribute features with that of masked version, and regard them as the inputs of classifier. As a result, each training sample produces two types of attribute features with different detection difficulties. Thus, this strategy can also be viewed as a feature-level data augmentation technique.

In summary, our main contributions are three-fold:

- We introduce a novel multi-scale disentanglement framework that treats extracted features instead of images as decoupling targets, simplifying the decoupling process and preventing the forgery-irrelevant information from dominating the attribute features.
- We propose a critical forgetting mechanism that forces the detector to uncover forgery traces that were previously overlooked rather than specific types of artifacts in training data, enhancing the generalization capability.
- Extensive experiments on forgery datasets demonstrate that our proposed method outperforms the state-of-the-art methods in terms of generalization and robustness.

Related Works

In recent years, various research efforts (Amerini et al. 2019; Yang, Rongrong, and Yao 2020; Bonettini et al. 2021; Nguyen, Yamagishi, and Echizen 2019) have been made to improve the performance of deepfake detection, achieving acceptable results in intra-dataset detection. FWA (Li and Lyu 2019) simulates the affine distortion in deepfake generation by using resolution differences between source and target images. Xuan *et al*. (Xuan et al. 2019) enhances CNN generalization by applying image preprocessing methods during training. Face X-Ray (Li et al. 2020a) uses the inconsistency of the underlying image statistics on the boundary to find the mixed border and then perform forgery detection. MAT (Zhao et al. 2021a) formulates deepfake detection as a fine-grained classification problem. CMAD (Zhao et al. 2023) adaptively extracts fine-grained features and achieves multi-modal feature fusion. DMGT (Buyun et al. 2023) addresses depth information loss in forgeries through depth map prediction. Frequency domain information (Qian et al. 2020; Jeong et al. 2022; Luo et al. 2021) is also used to enhance detector generalization. Other studies (Hao et al. 2023; Coccomini et al. 2022; Hanqing et al. 2022; Sun et al. 2022; Dong et al. 2022; Zhuang et al. 2022) employ transformer (Vaswani et al. 2017) to detect the artifact from a global perspective, overcoming CNN's limited receptive field.

Recent works (Liang, Shi, and Deng 2022; Yan et al. 2023; Li Lin and Yan Ju 2024) introduce decoupled representation learning to enhance detector generalization by separating forgery-relevant and forgery-irrelevant features. These methods assume that features unaffected by cross-

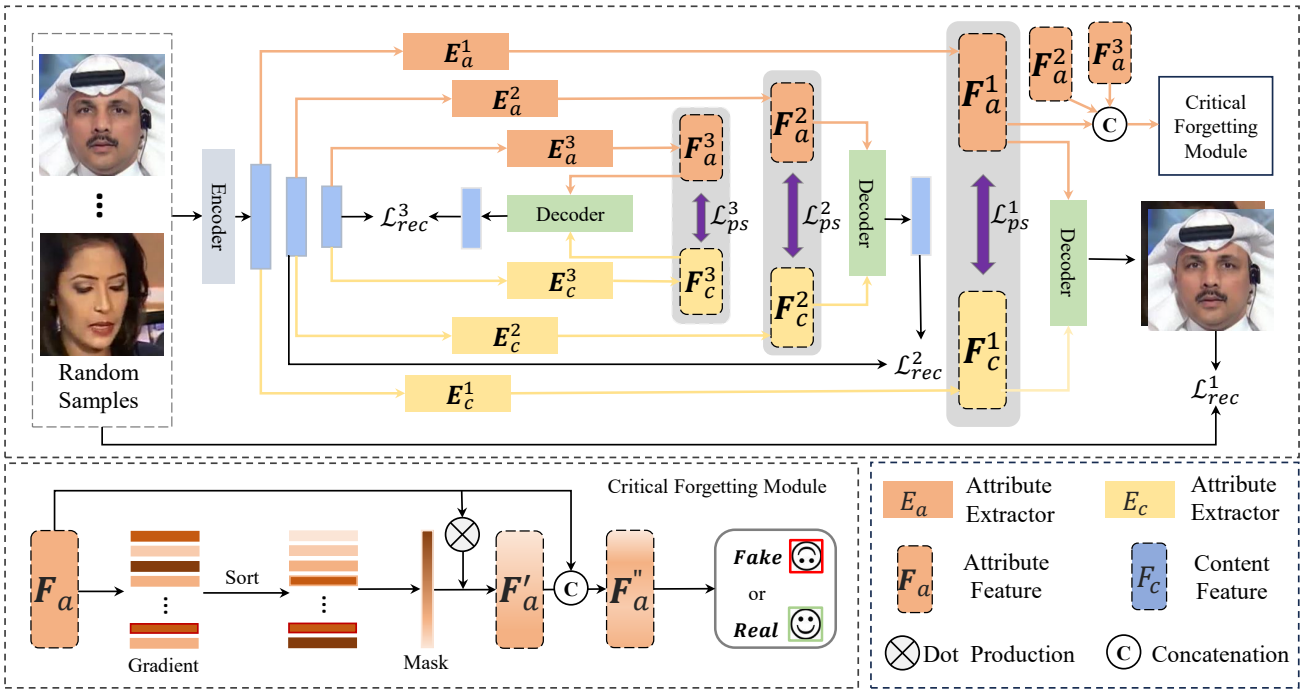


Figure 2: Architecture of the proposed framework. We use features as decoupling targets to ease the decoupling difficulty. The Pearson loss $\mathcal{L}_{ps}^{i}$ is introduced to ensure the attribute is forgery-relevant. The feature-level reconstruction loss $\mathcal{L}_{recon}^{i}$ is proposed to constrain multi-scale decoupling. The critical forgetting module forces the detector to forget specific types of artifacts in training data and uncover the common forgery traces.

reconstruction contain forgery-related information. However, our research indicates that this assumption does not always hold in cross-domain scenarios. To address this, we extend single-scale representation decoupling to a multi-scale approach and propose a critical forgetting mechanism to focus the detector on general discriminative representations.

Method

Overview

Three primary challenges impede the broader application of disentangled representation learning in face forgery detection. First, previous works (Yan et al. 2023; Li Lin and Yan Ju 2024) have overlooked the excessive density of information within an image, thus leading to the difficulty of decoupling. Second, cross-reconstruction can only ensure that attribute features are reconstruction-irrelevant rather than forgery-relevant, and single-scale forgery representation decoupling cannot capture sufficient discriminative information. Third, there is a lack of further exploitation of decoupled features, and the detector tends to prioritize learning specific artifacts within the training set to accomplish the classification task.

To address the above issues, we propose an enhanced disentanglement representation learning framework, as depicted in Figure 2. First, an encoder is employed to transform the image to mitigate the issue of excessive information density within an image and ease decoupling difficulty. Then, the extracted features are decoupled in different scales to isolate attribute features and content features, respectively. The Pearson coefficient is introduced to ensure that the attribute features and content features are uncorrelated. Meanwhile, the decoder aims to ensure the decoupled features are the complete representation of the input image or feature. Subsequently, the forgery classifier forces forgery-relevant information to converge mainly on the attribute features. Finally, the critical forgetting mechanism compels the detector to reveal the common forgery traces by forgetting specific types of artifacts in the training set, thereby improving the generalization ability of the detector.

Enhanced Disentanglement Framework

Encoder. Different from existing disentanglement-based face forgery detection frameworks that require the input to be pairs of images, the encoder $\boldsymbol{E}_n$ adopts a single image $\boldsymbol{x}$ as input and transforms it into feature maps, and each map represents information about one particular aspect of the face, as shown in Figure 3.

$$\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n = \boldsymbol{E}_n(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{f}_i, i \in [1, \ldots, n]$ represents the extracted features.

Attribute Extractor and Content Extractor. We extend the forgery representation decoupling from single-scale to multi-scale. The decoupler at the i_{th} scale contains an attribute extractor $\boldsymbol{E}_a^i$ and a content extractor $\boldsymbol{E}_c^i$, which have

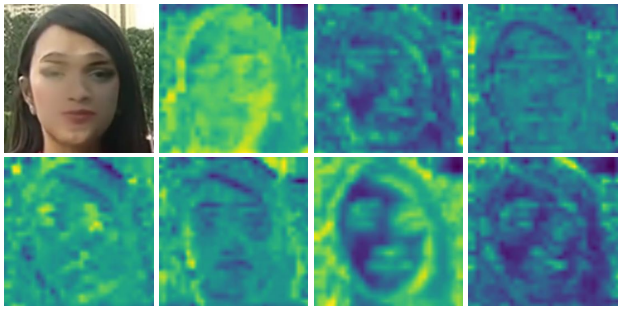


Figure 3: Visualization of the intermediate feature extracted by the encoder. Face contour, texture, and low-frequency is depicted in the fourth, sixth ,and seventh image, respectively.

the same structure but do not share parameters. The content extractor is used to extract forgery-irrelevant content $\boldsymbol{F}_c^i$, and the attribute extractor isolates forgery-relevant attributes $\boldsymbol{F}_a^i$ used for forgery classification,

$$\boldsymbol{F}_a^i,\ \boldsymbol{F}_c^i = \boldsymbol{E}_a^i(\boldsymbol{f}_i),\ \boldsymbol{E}_c^i(\boldsymbol{f}_i). \tag{2}$$

The Pearson Coefficient serves as a separation constraint, with smaller values indicating greater direction and magnitude discrepancy between two features. A sufficiently large discrepancy makes it possible to avoid possible correlation between attribute and content features. The Pearson loss can be formulated as below,

$$\mathcal{L}_{ps} = \frac{Cov(\boldsymbol{F}_a^i,\ \boldsymbol{F}_c^i)}{\sigma(\boldsymbol{F}_a^i)\sigma(\boldsymbol{F}_c^i)}, \tag{3}$$

where Cov represents the covariance between the attribute and content features, and σ is variance. Each attribute feature F_a^i is followed by a corresponding classifier to ensure it is forgery-relevant, which is illustrated below,

$$\mathcal{L}_{cls}^a = BCE(\boldsymbol{F}_a^i, \boldsymbol{y}), \tag{4}$$

where BCE denotes the cross-entropy loss, and $\boldsymbol{y} \in \{\text{fake}, \text{real}\}$ is the binary classification label corresponding to the input image $\boldsymbol{x}$. Combined with the effect of the Pearson Coefficient, the forgery-relevant information dominates the attribute features, and forgery-irrelevant information is dominant in content features accordingly.

Decoder. Previous single-scale decoupling methods only use image reconstruction loss to ensure that the extracted attribute and content features are a complete representation of an image. However, this method is not applicable in multi-scale decoupling. If the i_{th} scale and the $i + 1_{th}$ scale use the same completeness constraint, it means that the information in the features of different scales is exactly the same. In other words, we repeatedly decouple the same information using consistent decoupling constraints.

Deep InfoMax (Hjelm et al. 2019) indicates the complete representation of an image is not necessary for the classification tasks. Inspired by this, we propose the feature-level reconstruction loss to tackle this issue. Each scale of decoupled features has a corresponding decoder $\boldsymbol{D}_e^i$. The decoder $\boldsymbol{D}_e^i$ aims to guarantee that the information within attribute $\boldsymbol{F}_a^i$ and content features $\boldsymbol{F}_c^i$ is a complete representation of the input features instead of the input image, expect for $i = 1$. The feature reconstruction loss can be written as:

$$\mathcal{L}_{rec} = Mse(\boldsymbol{f}_i, \boldsymbol{D}_e^i(\boldsymbol{F}_a^i, \boldsymbol{F}_c^i)), \tag{5}$$

where Mse denotes the Mean Square Error. The 1_{th} scale reconstruction target is the input image.

Critical Forgetting Mechanism. To utilize the multi-scale attribute features $\boldsymbol{F}_a$ more effectively and avoid overfitting to specific artifacts in training data, we propose a critical forgetting mechanism. The key idea is to force the detector to forget the specific artifacts, and mine the common subtle forgery traces that were previously overlooked. Since specific types of artifacts in the training set are easy to capture, they often manifest as feature maps that have the most significant impact on gradient backpropagation. This effect can be embodied using the magnitude of the gradient. Thus, we can quantify the impact of attribute features by sorting the gradients of different feature maps during training, with larger gradients representing greater impact. The acquisition of gradient G can be formulated as:

$$G = \partial(\frac{1}{N}\sum_{i=1}^{N} -[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)])/\partial \boldsymbol{F}_a, \tag{6}$$

where N is the number of samples in a batch, y_i represents the label of the i_{th} sample, and p_i is the prediction of the detector. The attribute features are sorted identically according to the sorting results of the gradient. Then, by designing the gradient threshold, the most influential part of the sorted attribute feature maps is forced to forget, that is, the channels whose gradient are greater than the given threshold. This forgetting is achieved by masking feature maps, and the process of masking can be expressed as follows:

$$\boldsymbol{F}_a^{'} = Sort(\boldsymbol{F}_a|G) \otimes \boldsymbol{M}_\sigma, \tag{7}$$

$$\tilde{M}_{ij} = \begin{cases} 0 & \text{if } \boldsymbol{F}_{a_{ij}} > \theta\ , \\ 1 & \text{otherwise}\ , \end{cases} \tag{8}$$

where $\boldsymbol{M}_\sigma$ is a 0-1 matrix, and σ is the hyperparameter used to control the proportion of forgetting. θ is the forgetting threshold selected from the ranked feature map according to hyperparameter σ. For example, $\boldsymbol{M}_{\sigma=0.3}$ indicates that the value of the channel in the top 30% of importance in the attribute feature map is set to 0, and the values in the rest of the feature maps are unchanged.

To prevent forgetting from rendering the detector insensitive to specific types of artifacts within the training set, we concatenate attribute features $\boldsymbol{F}_a$ with forgotten features $\boldsymbol{F}_a^{'}$ as input to the classifier rather than employing the forgotten features alone for forgery classification. Therefore, N input images actually generate 2N classification samples, and the critical forgetting mechanism can be considered as a feature-level augmentation strategy. The corresponding classification loss can be written as :

$$\mathcal{L}_{cls} = BCE(Con(\boldsymbol{F}_a, \boldsymbol{F}_a^{'}), Con(\boldsymbol{y}, \boldsymbol{y})), \tag{9}$$

where Con represents concatenation operation and BCE denotes cross-entropy loss. $\boldsymbol{y} \in \{\text{fake}, \text{real}\}$ is the binary classification label corresponding to the input $\boldsymbol{x}$.

Methods	CDF		WDF		DFDC		Avg	
	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
Xception (Rossler et al. 2019)	61.80	41.73	62.72	40.65	63.61	40.58	62.71	40.99
F^3Net (Qian et al. 2020)	61.51	42.03	57.10	45.12	64.60	39.84	61.07	42.33
Add-Net (Zi et al. 2020)	65.29	38.90	62.35	41.42	64.78	40.23	64.14	40.18
RFM (Wang and Deng 2021)	65.63	38.54	57.75	45.45	66.01	39.05	63.13	41.01
MAT (Zhao et al. 2021a)	67.02	37.90	59.74	43.73	68.01	37.17	64.92	39.60
RECCE (Cao et al. 2022)	68.71	35.73	64.31	40.53	69.06	36.08	67.36	37.45
UCF (Yan et al. 2023)	71.73	—	—	—	60.03	—	—	—
DisGRL (Shi et al. 2023)	70.03	<u>34.23</u>	<u>66.73</u>	<u>39.24</u>	**70.89**	**34.27**	69.22	35.91
PFG (Li Lin and Yan Ju 2024)	<u>74.42</u>	—	—	—	61.47	—	—	—
Ours	**76.30**	**31.25**	**69.29**	**36.52**	<u>70.27</u>	<u>34.94</u>	**71.95**	**34.28**

Table 1: Cross-dataset evaluation on FF++ (c40). The training data comes from a higher compression version of FF++, and the frame-level AUC (%) results on three unseen datasets (CDF, DFDC, and WDF) are recorded. All results are cited from DisGRL (Shi et al. 2023) and PFG (Li Lin and Yan Ju 2024).

Contrastive Learning. We introduce contrastive learning (Sun et al. 2020) to further improve the generalization performance. Contrastive learning maximizes the intra-class similarity s_p and minimizes the inter-class similarity s_n, where similarity can be measured as the distance between features. Specifically, given a training batch with K same class sample pairs and L different class sample pairs, the contrastive learning loss can be formulated as:

$$\mathcal{L}_{ct} = log[1 + \sum_{i=1}^{K}\sum_{j=1}^{L} exp(\gamma(s_n^j - s_p^i + m))], \quad (10)$$

where γ is a scale factor and m is a threshold used for similarity.

Overall Loss. The overall loss function $\mathcal{L}$ used for the training process is a weighted sum of the four loss functions listed above:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1\mathcal{L}_{ct} + \lambda_2\mathcal{L}_{rec} + \lambda_3\mathcal{L}_{ps} + \lambda_4\mathcal{L}_{cls}^a, \quad (11)$$

where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are hyper-parameters measuring the impact of each loss on the overall loss.

Experiment

Settings

Dataset. Following (Chen et al. 2022; Yan et al. 2023; Zhuang et al. 2022), we use FaceForensics++ (FF++) (Rossler et al. 2019) as the training dataset, which consists of 1,000 original and 4,000 fake videos. Fake videos are manipulated by four forgery algorithms using the same set of pristine videos: DeepFakes (DF), Face2Face (F2F) (Thies et al. 2016), NeuralTexture (NT) (Thies, Zollhöfer, and Nießner 2019), FaceSwap (FS). The split setting for training and validation is the same as the initial dataset setting. To evaluate the generalization ability of the proposed method, we conduct extensive experiments on four large-scale benchmark databases: Celeb-DF V2 (CDF) (Li et al. 2020b), DeepFake Detection Challenge (DFDC) (Dolhansky et al. 2020) and WildDeepFake (WDF) (Zi et al. 2020). CDF contains 590 real videos and 5,639 synthesized videos. DFDC is a more challenging dataset comprising 28,154 fake videos manipulated by improved forgery methods. The 7,314 face sequences in WDF are collected from the Internet and are more in line with practical forgery detection scenarios.

Implemenation details. In the data pre-processing, RetinaFace (Deng et al. 2020) is employed to extract faces, and the aligned faces are resized to 256×256 for both the training and testing. In all experiments, we adopt EfficientNet (Tan and Le 2019) pre-trained on ImageNet (Deng et al. 2009) as the backbone when not otherwise specified. For training, we use a batch size of 256 and employ the Adamw optimizer with beats 0.9 and 0.999. The learning rate is set as 2e-4. The hyper-parameters λ_1, λ_2, and λ_3 in the loss function (11) are set as 0.1, 0.15, 0.1, and 0.1, respectively.

Evaluations

Cross-Dataset Evaluation on FF++ (c40). We first evaluate the proposed method against state-of-the-art deepfake detection methods under the most challenging training settings: the detector is trained on the highly compressed version (c40) of FF++ dataset and tested on CDF, WDF, and DFDC. This setting represents the most realistic approximation to real-world forgery detection scenarios. To conceal the forgery traces, malicious attackers may intentionally degrade the image quality through compression after completing the manipulation. The evaluation metrics include Area Under the Curve (AUC) and Equal Error Rate (EER). The results are tabulated in Table 1 and calculated at the frame level. As depicted, the AUC metric of our proposed method is higher than that of suboptimal work on CDF (76.30% vs. 74.42%) and WDF (69.29% vs. 66.73%). The same trend is observed for the EER metric as well (31.25% vs.34.23% and 36.52% vs. 39.24%). On DFDC, our method also achieved competitive performance, with both metrics close to the optimal method. Compared with UCF and PFG, which also employ disentanglement representation learning, our pro-

Methods	FF++	CDF
Xception (Rossler et al. 2019)	96.3	65.5
Face X-ray (Li et al. 2020a)	97.4	74.2
F^3Net (Qian et al. 2020)	98.8	71.2
RFM (Wang and Deng 2021)	98.8	67.6
LTW (Sun et al. 2021)	99.2	77.1
PCL (Zhao et al. 2021b)	99.1	81.8
Local-Relation (Chen et al. 2021)	99.5	78.3
RECCE (Cao et al. 2022)	99.3	73.2
SLADD (Chen et al. 2022)	98.4	79.7
UIA-ViT (Zhuang et al. 2022)	99.3	82.4
UCF (Yan et al. 2023)	**99.6**	82.4
CADDM (Dong et al. 2023)	99.3	80.7
IID (Huang et al. 2023)	99.3	83.8
Ours	98.7	**84.9**

Table 2: Cross-dataset evaluation on FF++ (c23).

posed method improves the AUC on CDF from 74.42% to 76.30%, and DFDC from 61.47% to 70.27%. This indicates that the proposed disentanglement-enhancing method and critical forgetting mechanism mitigate the discrepancies between training and test data as expected.

Cross-dataset Evaluation on FF++ (c23). This setting considers both intra-dataset and cross-dataset scenarios, which are the most prevalent evaluation settings in previous works. The detector is trained on the slightly compressed version (c23) of FF++ and tested on FF++ and WDF. The negative impact of compression can be disregarded due to its low degree of compression. The Area Under the Curve (AUC) is employed as the evaluation metric. Table 2 shows the frame-level results. We can observe that our method achieves the best performance on CDF (84.9% vs. 83.8%). On intra-dataset scenarios, where both training and test data come from FF++ (c23), the AUC of our method is 0.9% lower than that of the optimal UCF. Also based on disentanglement representation learning, the slightly inferior performance on intra-dataset scenarios compared to UCF can be attributed to the fact that the proposed method forces the detector to pay more attention to more generalized traces rather than specific types of artifacts in the training set. Although we employ a concentration operation to alleviate this decline, the issue persists. Given the performance improvement of the detector on the cross-dataset scenario, the slight decline in performance within the intra-dataset scenario is justified.

Cross-Manipulation Evaluation. On the cross-dataset scenarios, the discrepancies between training and test data arise from the combination of different source videos and distinct forgery techniques employed. To eliminate the interference of the different source videos used, following SLADD (Chen et al. 2022), we conduct the cross-manipulation evaluation to verify the robustness of our method. The model is trained on the NT (c23) and tested on different compressed versions (c23 & c40) of DF and FS, as demonstrated in Table 3. We can observe that the proposed method consistently performs better than other methods on DF. On the highly compressed version (c40) of DF, our method outperforms the second-best SLADD by 4.7% in terms of the AUC metric. Although our method is not the best with the highly compressed version (c40) of FS, it is second only to SLADD in the cross-compression level and achieves an AUC of 55.6%. This indicates that the proposed method is robust to the effect of compression operation.

Methods	Test set			
	c40		c23	
	DF	FS	DF	FS
Xception (Rossler et al. 2019)	58.7	51.7	77.0	71.8
Face X-ray (Li et al. 2020a)	57.1	51.0	58.5	77.9
F^3Net (Qian et al. 2020)	58.3	51.9	80.5	61.2
RFM (Wang and Deng 2021)	55.8	51.6	79.8	63.9
SRM (Luo et al. 2021)	55.5	52.9	83.8	**79.5**
SLADD (Chen et al. 2022)	62.8	**56.8**	84.6	72.1
Ours	**67.5**	55.6	**85.4**	71.1

Table 3: Generalizability comparisons across different compression levels in the term of AUC. All results are cited from SLADD (Chen et al. 2022).

Ablation Study

Impact of Different Components. To evaluate the contribution of each component in the proposed framework, we design several ablation experiments with different components, all based on baseline model Xception. The components under evaluation include the decoupling objects, multi-scale forgery representation decoupling, contrastive learning, and critical forgetting mechanism, and results are shown in Table 4. We can observe that using the image as a decoupled target improves the generalization performance of the detector. When the single-scale decoupling is extended to multi-scale decoupling, the performance of the detector is improved rapidly. On this basis, the critical forgetting mechanism effectively helps the detector learn a more general discriminative representation. Contrastive learning also achieves the expected performance. The introduction of the critical forgetting mechanism is the first exploration of further utilizing decoupled features in face forgery detection. As illustrated in Table 4, forgetting the most salient features and paying more attention to overlooked features during training improves the detection performance. The forgetting ratio is chosen manually based on experience.

Impact of Different Backbones. The primary performance gains of our method arise from the multi-scale forgery representation decoupling and the introduction of the critical forgetting mechanism, with the approach demonstrating the potential for widespread applicability due to its minimal network modifications. To validate this assumption, we conduct experiments to evaluate the performance of the proposed method using different backbone architectures. As depicted in Table 5, both frameworks built upon Xception

$\mathcal{C}_{pd}$	$\mathcal{C}_{md}$	$\mathcal{C}_{cf}$	$\mathcal{C}_{ct}$	CDF		WDF		DFDC	
				AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
✗	✗	✗	✗	70.80	34.08	64.53	40.71	66.31	38.54
✓	✗	✗	✗	72.45	33.25	66.51	39.40	67.64	37.61
✓	✓	✗	✗	74.94	32.36	67.04	38.24	68.18	37.23
✓	✓	✓	✗	75.91	31.48	68.44	38.06	68.67	36.94
✓	✓	✓	✓	**76.30**	**31.25**	**69.29**	**36.52**	**70.27**	**34.94**

Table 4: Ablation study with different components. (a) $\mathcal{C}_{pd}$: Decouple the features rather than images. (b) $\mathcal{C}_{md}$: multi-scale forgery representation decoupling. (c) $\mathcal{C}_{cf}$: the critical forgetting mechanism. (d) $\mathcal{C}_{ct}$: the contrastive learning.

Methods	CDF		WDF		DFDC	
	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
Xception	69.21	36.47	66.88	39.95	64.52	40.58
Xceprion+Ours	76.64	30.58	69.02	37.22	68.99	36.43
Efficient	70.80	34.08	64.53	40.71	66.31	38.54
Efficient+Ours	76.30	31.25	69.29	36.52	70.27	34.94

Table 5: Ablation study with different backbones. The training data comes from FF++ (c40) with a higher compression level, and the frame-level AUC (%) results on three unseen datasets (CDF, DFDC, and WDF) are recorded.

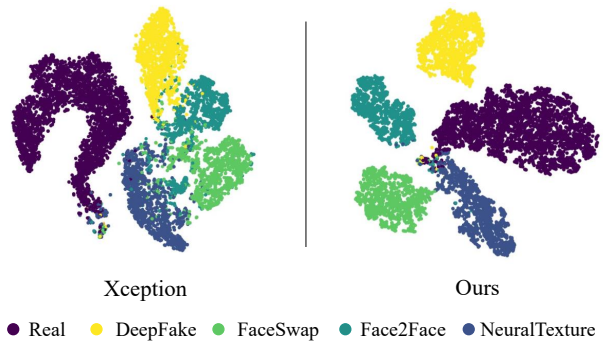


Figure 4: The t-SNE visualization on intra-dataset scenario. The detector is train on FF++ (c23) and different colors represent different subsets.

and Efficient achieve comparable performance on the cross-dataset scenarios under the same experimental settings.

Visualization

The t-SNE Visualization. We further explore the discriminative ability of the proposed decoupled framework using the t-SNE (Laurens and Hinton 2008) visualization technique. As illustrated in Figure 4, xecption roughly separates fake images from real ones and ignores the differences between different forgery methods. This signifies that the baseline model accomplishes the forgery detection task relying on the given hard label (0 and 1), and lacks the proficiency to extensively investigate forgery traces. Conversely, our method not only differentiates forged samples from real ones, but also compels samples from the identical forgery method to cluster together within the intra-dataset scenario. This accurate discrimination represents that the further utilization of the isolated attribute features is helpful to improve the generalization of the detector.

Conclusion

This paper proposes a disentangled representation learning method to enhance the generalization capability of the detector for unseen manipulations. First, we employ features that focus on an aspect of the face information as decoupled targets rather than images to ease the decoupling difficulty. Then, we extend the forgery representation decoupling to a multi-scale one, and introduce the Pearson coefficient to separate the attribute features from the content features and ensure that the attribute features are as forgery-relevant as possible. Subsequently, we propose a critical forgetting mechanism to force the detector to pay more attention to common subtle forgery traces that were previously overlooked, therefore enhancing the generalization capability. Extensive experimental results indicate that our method has the potential to be applied in real-world scenarios.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.U24B20175), the Key Lab of Information Network Security of Ministry of Public Security (The Third Research Institute of Ministry of Public Security), and the Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grants No. 23xkjc010.

References

Amerini, I.; Galteri, L.; Caldelli, R.; and Del Bimbo, A. 2019. Deepfake Video Detection through Optical Flow Based CNN. In *ICCV*, 1205–1207.

Bonettini, N.; Cannas, E. D.; Mandelli, S.; Bondi, L.; Bestagini, P.; and Tubaro, S. 2021. Video Face Manipulation Detection Through Ensemble of CNNs. In *ICPR*, 5012–5019.

Buyun, L.; Zhongyuan, W.; Baojin, H.; Qin, Z.; Qian, W.; and Jingjing, L. 2023. Depth map guided triplet network for deepfake face detection. *Neural Networks*, 159: 34–42.

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 *CVPR*, 4103–4112.

Chen, L.; Zhang, Y.; Song, Y.; Liu, L.; and Wang, J. 2022. Self-supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection. In *CVPR*, 18689–18698.

Chen, S.; Yao, T.; Chen, Y.; Ding, S.; Li, J.; and Ji, R. 2021. Local Relation Learning for Face Forgery Detection. *AAAI*, 35(2): 1081–1088.

Coccomini, D. A.; Messina, N.; Gennaro, C.; and Falchi, F. 2022. Combining EfficientNet and Vision Transformers for Video Deepfake Detection. In *Image Analysis and Processing*, 219–229.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *CVPR*, 5202–5211.

Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; and Ferrer, C. C. 2020. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, abs/2006.07397.

Dong, S.; Wang, J.; Ji, R.; Liang, J.; Fan, H.; and Ge, Z. 2023. Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization. In *CVPR*, 3994–4004.

Dong, X.; Bao, J.; Chen, D.; Zhang, T.; Zhang, W.; Yu, N.; Chen, D.; Wen, F.; and Guo, B. 2022. Protecting Celebrities from DeepFake with Identity Consistency Transformer. In *CVPR*, 9458–9468.

Hanqing, Z.; Wenbo, Z.; Dongdong, C.; Weiming, Z.; and Nenghai, Y. 2022. Self-supervised Transformer for Deepfake Detection. *ArXiv*, abs/2203.01265.

Hao, L.; Wenmin, H.; Weiq, i. L.; and Wei, L. 2023. DeepFake detection with multi-scale convolution and vision transformer. *Digital Signal Processing*, 134: 103895.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *ICLR*.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4490–4499.

Jeong, Y.; Kim, D.; Ro, Y.; and Choi, J. 2022. FrePGAN: Robust Deepfake Detection Using Frequency-Level Perturbations. In *AAAI*, volume 36, 1060–1068.

Laurens, V. D. M.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(2605): 2579–2605.

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 *CVPR*, 5000–5009.

Li, Y.; and Lyu, S. 2019. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *CVPR*.

Li, Y.; Sun, P.; Qi, H.; and Lyu, S. 2020b. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *CVPR*.

Li Lin, X. H.; and Yan Ju, e. a. 2024. Preserving Fairness Generalization in Deepfake Detection. In *CVPR*.

Liang, J.; Shi, H.; and Deng, W. 2022. Exploring Disentangled Content Information for Face Forgery Detection. In *ECCV*, volume 13674, 128–145.

Luo, Y.; Zhang, Y.; Yan, J.; and Liu, W. 2021. Generalizing Face Forgery Detection with High-frequency Features. In *CVPR*, 16312–16321. Computer Vision Foundation / IEEE.

Nguyen, H. H.; Yamagishi, J.; and Echizen, I. 2019. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos. In *ICASSP*, 2307–2311.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In *ECCV*, 86–103.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; and Thies, J. 2019. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, 1–11.

Shi, Z.; Chen, H.; Chen, L.; and Zhang, D. 2023. Discrepancy-Guided Reconstruction Learning for Image Forgery Detection. In *IJCAI*, 1387–1395.

Sun, K.; Liu, H.; Ye, Q.; Gao, Y.; Liu, J.; Shao, L.; and Ji, R. 2021. In *Domain General Face Forgery Detection by Learning to Weight*, volume 35, 2638–2646. AAAI Press.

Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; and Wei, Y. 2020. Circle Loss: A Unified Perspective of Pair Similarity Optimization. In *CVPR*, 6397–6406.

Sun, Y.; Zhang, Z.; Qiu, C.; Wang, L.; Sun, L.; and Wang, Z. 2022. FakeTransformer: Exposing Face Forgery From Spatial-Temporal Representation Modeled By Facial Pixel Variations. In *ICSP*, 705–713.

Tan, M.; and Le, Q. V. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *ICML*, volume 97, 6105–6114.

Thies, J.; Zollhöfer, M.; and Nießner, M. 2019. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4): 1–12.

Thies, J.; Zollhöfer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2016. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *CVPR*, 2387–2395.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, u.; and Polosukhin, I. 2017. Attention is All You Need. In *NIPS*, 6000–6010.

Wang, C.; and Deng, W. 2021. Representative Forgery Mining for Fake Face Detection. In *CVPR*, 14918–14927.

Xuan, X.; Peng, B.; Wang, W.; and Dong, J. 2019. On the Generalization of GAN Image Forensics. In *CCBR*, 134–141.

Yan, Z.; Zhang, Y.; Fan, Y.; and Wu, B. 2023. UCF: Uncovering Common Features for Generalizable Deepfake Detection. In *ICCV*, 22412–22423.

Yang, Y.; Rongrong, N.; and Yao, Z. 2020. Mining Generalized Features for Detecting AI-Manipulated Fake Faces. *arXiv preprint arXiv:2010.14129*.

Zhao, H.; Wei, T.; Zhou, W.; Zhang, W.; Chen, D.; and Yu, N. 2021a. Multi-attentional Deepfake Detection. In *CVPR*, 2185–2194.

Zhao, L.; Zhang, M.; Ding, H.; and Cui, X. 2023. Fine-grained deepfake detection based on cross-modality attention. *Neural Computing and Applications*.

Zhao, T.; Xu, X.; Xu, M.; Ding, H.; Xiong, Y.; and Xia, W. 2021b. Learning Self-Consistency for Deepfake Detection. In *ICCV*, 15003–15013. IEEE.

Zhuang, W.; Chu, Q.; Tan, Z.; Liu, Q.; Yuan, H.; Miao, C.; Luo, Z.; and Yu, N. 2022. UIA-ViT: Unsupervised Inconsistency-Aware Method Based On Vision Transformer For Face Forgery Detection. In *ECCV*, 391–407.

Zi, B.; Chang, M.; Chen, J.; Ma, X.; and Jiang, Y.-G. 2020. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In *ACMMM*, 2382–2390.