UCF: Uncovering Common Features for Generalizable Deepfake Detection

Zhiyuan Yan[*1]   Yong Zhang[*2]   Yanbo Fan[2]   Baoyuan Wu[†1]

[1] The School of Data Science,
The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China

[2]Tencent AI Lab

{yanzhiyuan1114, zhangyong201303, fanyanbo0124}@gmail.com, wubaoyuan@cuhk.edu.cn

Abstract

Deepfake detection remains a challenging task due to the difficulty of generalizing to new types of forgeries. This problem primarily stems from the overfitting of existing detection methods to forgery-irrelevant features and method-specific patterns. The latter has been rarely studied and not well addressed by previous works. This paper presents a novel approach to address the two types of overfitting issues by uncovering common forgery features. Specifically, we first propose a disentanglement framework that decomposes image information into three distinct components: forgery-irrelevant, method-specific forgery, and common forgery features. To ensure the decoupling of method-specific and common forgery features, a multi-task learning strategy is employed, including a multi-class classification that predicts the category of the forgery method and a binary classification that distinguishes the real from the fake. Additionally, a conditional decoder is designed to utilize forgery features as a condition along with forgery-irrelevant features to generate reconstructed images. Furthermore, a contrastive regularization technique is proposed to encourage the disentanglement of the common and specific forgery features. Ultimately, we only utilize the common forgery features for the purpose of generalizable deepfake detection. Extensive evaluations demonstrate that our framework can perform superior generalization than current state-of-the-art methods.

1. Introduction

Deepfake technology has gained significant attention in recent years due to its ability to generate highly realistic videos. While deepfake has the potential to be used for various purposes, including entertainment and marketing, it has also been misused for illegal purposes. The use of deepfake to create false content can compromise people's
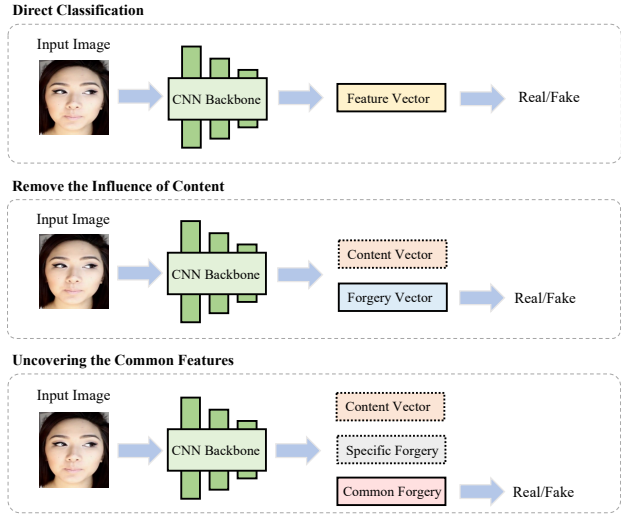

Figure 1: Comparison among different classification methods. The first is a direct classification that uses whole features. The second approach eliminates content features to prevent overfitting to forgery-irrelevant features. Our approach, the third one, not only removes the influence of content but also prevents overfitting to specific forgery patterns by uncovering common features.

privacy, spread misinformation, and erode trust in digital media, resulting in severe outcomes like reputational harm, incitement of violence, and political instability.

As a result, developing a reliable and effective deepfake detection algorithm is vitally essential. Recently, a large number of detectors [58, 25, 37, 52, 36, 55, 8] have been proposed for deepfake. Existing detectors generally perform well when the training and testing data are created using the same forgery techniques. However, in real-world applications, the testing data may be created using unknown procedures, leading to differences between the training and testing data and resulting in poor detection performance. This phenomenon, known as the generalization problem in deepfake detection, presents a significant challenge to the

---

[*]Equal contribution
[†]Corresponding Author

practical use of current detection methods.

Currently, an increasing number of studies are dedicated to tackling the issue of generalization in deepfake detection. These works typically utilize blending artifacts [23, 41] or frequency artifacts [28, 30], and some employ adversarial training to synthesize challenging forgeries [7]. However, these approaches are limited in their reliance on predefined forgery patterns. For instance, Face X-ray [23] assumes the presence of a blending region in the forged image, which could potentially curtail the effectiveness of the approach when generalized to novel and unseen forgeries. In addition, these methods consider the entire feature space when addressing the problem, which could be disrupted by irrelevant factors such as background [27] and identity [13].

To address the above challenges, we adopt the perspective of content and style [20] and formulate the problem of deepfake as an integration of two distinct components: content and fingerprint. The content in deepfake refers to elements, *e.g.,* the background, identity, and facial appearance, which are not directly related to the forgery. In contrast, the fingerprint represents the traits that are related to the forgery. The challenge, then, becomes how to effectively disentangle these two components and use only the forgery-related fingerprint for detection.

Several recent studies [51, 18, 27] attempt to address the generalization problem through disentanglement techniques. However, limited generalization capability remains a challenge in many cases. One main reason is the over-reliance on method-specific patterns in most disentanglement methods, which only aim to eliminate the influence of content. However, these methods may still learn patterns that are unique to a specific forgery method, thereby limiting their generalization performance.

To address this issue, we propose a novel disentanglement framework that differs from existing approaches (See Fig. 1). Our framework prevents overfitting to both content and specific forgery patterns. To achieve this, we employ a multi-task disentanglement framework and a conditional decoder to disentangle the input into content and fingerprint components. Moreover, we introduce a contrastive regularization technique to disentangle the fingerprint features into specific and common features. The specific features represent method-specific forgeries, while the common features are shared across different forgery methods. In our approach, only the common features are utilized for detection, which improves the generalization ability of the model. To validate our idea, we conduct a t-SNE visualization [46] in Fig. 2, demonstrating that the baseline and our specific components actually learn method-specific texture, while our common components are able to capture the common features across forgeries. Furthermore, the content does not differentiate between real and fake images, as expected.

Our contributions are summarized as follows:

- We propose a novel multi-task disentanglement framework to address two main challenges that contribute to the generalization problem in deepfake detection: overfitting to irrelevant features and overfitting to method-specific textures. By uncovering common features, our framework aims to enhance the generalization ability of the model.

- We propose a conditional decoder that helps disentangle forgery-irrelevant and forgery features, as well as a contrastive regularization technique that facilitates the disentanglement of common and specific forgery features. By utilizing these technologies, we aim to achieve improved disentanglement.

- Extensive experiments show that our framework can outperform the performance of current state-of-the-art methods in unseen testing datasets, demonstrating its effectiveness in generalization.

2. Related Work

To date, deepfake detection can be broadly categorized into two types of tasks: image forgery detection [37, 23, 7] and video forgery detection [38, 16, 57]. This paper specifically focuses on detecting image forgery.

Classical Detection Methods. Conventional deepfake detectors [3, 34, 37] typically focus on developing optimal CNN architectures. However, these methods often overlook the details present in the frequency domain of fake images, such as compression artifacts. To this end, several works [36, 14, 22, 15] utilize frequency information to improve the performance of detectors. Other notable directions are focusing on some specific representations, *i.e.,* forgery region location [33], neuron behaviors [49], optical flow [4], landmark geometric features [42], 3D decomposition [59], erasing technology [47], and attentional networks [10, 55, 48]. However, the generalization ability towards unseen forgery technologies of these conventional deepfake detectors is still limited.

Detection Methods toward Generalization. Deepfake detection poses a significant challenge in terms of generalization, where detectors perform poorly when training and testing on different data distributions. Despite this challenge, there is a limited amount of research in this area. One early method, FWA [25], leverages differences in resolution between forgery faces and backgrounds to detect deepfake. Recent works make significant progress in improving the generalization ability. Face X-ray [23] detects blending boundary artifacts, SPSL [28] proposes a frequency-based method by phase spectrum analysis, Lip-Forensics [16] leverages spatial-temporal networks to identify unnatural mouth movements, SRM [30] utilizes the
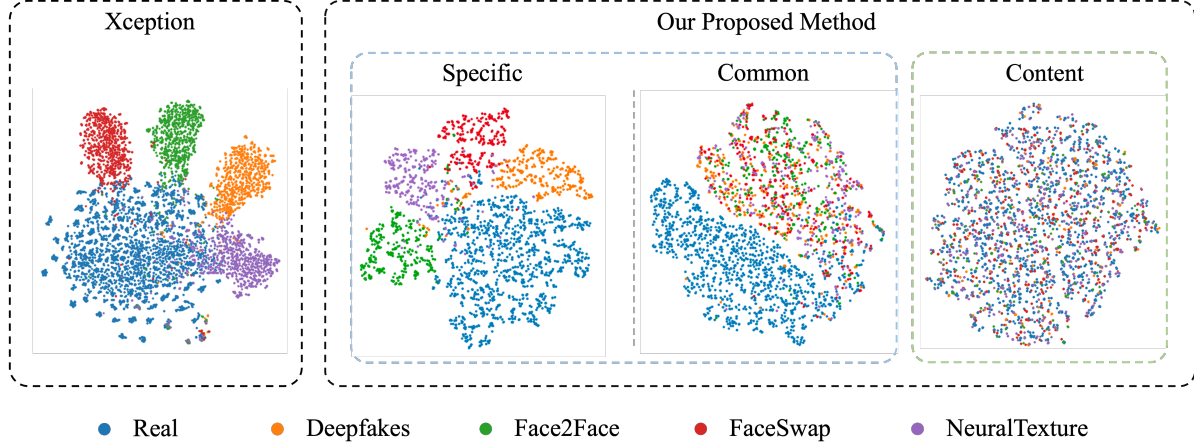
Figure 2: The t-SNE [46] visualization of features extracted from the baseline Xception [37] and our framework on FF++ [37]. In the visualization, images generated by the four methods locate separately in the latent space, which reveals that the baseline Xception actually learns method-specific features, consistent with our forgery-specific module. This observation explains that Xception can mainly recognize specific types of forgeries and thus fail to generalize well to a broader range of forgeries. Additionally, as expected, the common module of our method captures the common forgery features across different methods, while the content module captures only forgery-irrelevant features.

high-frequency noises for generalizable detection, PCL [56] measures patch-wise similarities of input images to identify deepfake, SBIs [41] and SLADD [7] improve generalization ability by combining data augmentation and blending. Although these approaches largely improve the generalization ability of classical detection methods, they are limited by the reliance on predefined forgery patterns and the consideration of the entire feature space, which can be disrupted by unrelated factors such as background [27] and identity [13].

Disentanglement Learning for Deepfake Detection. Disentanglement learning is a method that decomposes complex features into simpler, more narrowly defined variables and encodes them as separate dimensions with high discriminative power [5, 27]. In the field of deepfake detection, there are relatively few papers that are based on disentanglement learning. These works aim to separate forgery-irrelated and forgery-related features to extract forgery information from variations present in facial images. Hu *et al.* [18] propose a disentanglement framework that separates features, only using the manipulation-related features for detection. Zhang *et al.* [54] go further step by adding additional supervision to improve generalization ability. To ensure the independence of the disentangled features, Liang *et al.* [27] ensure feature independence through content consistency and global representation contrastive constraints.

Despite these efforts to tackle the generalization problem through disentanglement learning, this challenge still exists because these methods only remove the influence of content. In some cases, these methods may still fail to achieve complete disentanglement of forgery features, resulting in overfitting to method-specific textures and thereby limiting their ability to generalize to other unseen forgeries.

3. Methods

3.1. Motivation

There are two main factors that contribute to the generalization problem in deepfake detection. Firstly, many detectors are prone to focus too much on content information that is not directly related to the forgery, *i.e.,* the background, identity, and facial appearance. Secondly, different forgery techniques produce distinct forgery artifacts. These artifacts can be easily detected by a detector that is trained on a specific set of artifacts. However, detectors may be overfitted to one or more specific forgery technologies, leading to a lack of generalization to unseen forgeries. The second problem is often overlooked in previous works.

To address these issues, we propose a multi-task disentanglement learning framework to uncover common features for generalizable deepfake detection. Our framework aims to disentangle the input into the content, specific, and common forgery features. By only utilizing the common forgery features for detection, our framework can help improve the generalization ability of deepfake detectors and avoid the overfitting of both forgery-irrelated and method-specific features. Additionally, we introduce a conditional decoder and a contrastive regularization loss to further aid in disentanglement and enhance the generalization ability of the framework.
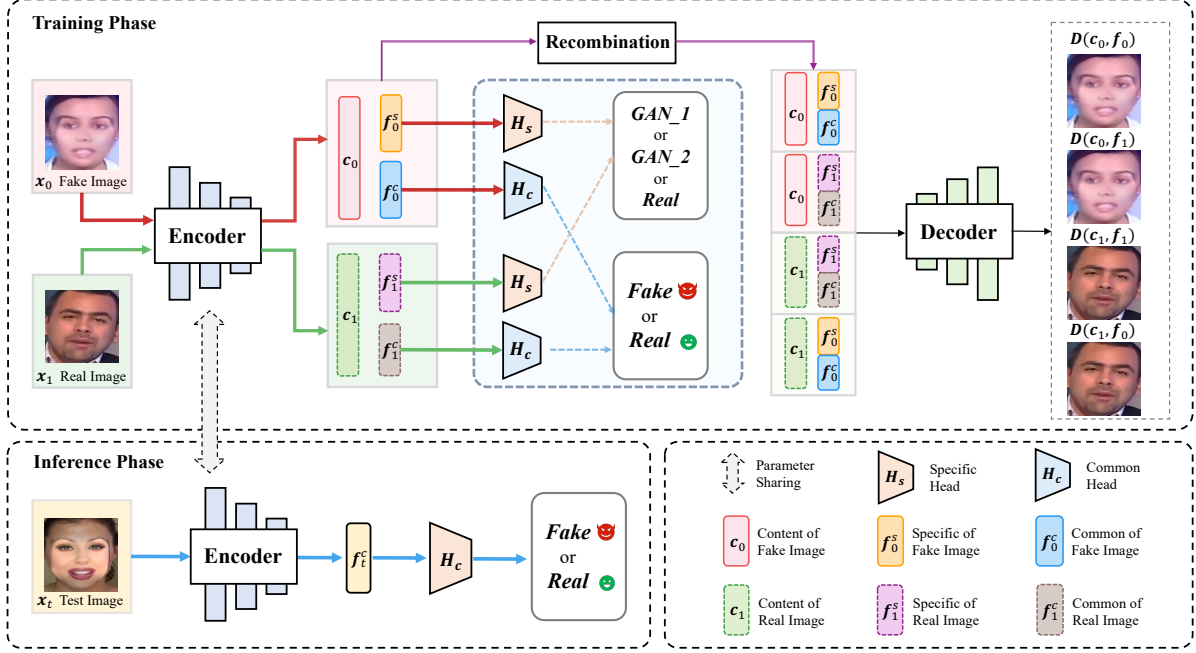
Figure 3: The overview framework of our proposed method. 1) For the encoder (E), we utilize it to obtain three distinct components: content, specific fingerprint, and common fingerprint. 2) For the recombination module, we recombine the fingerprints and contents from different input images. 3) For the decoder (D), we take the fingerprint and content as inputs to generate corresponding reconstruction images. 4) For the classification, we obtain the prediction results of specific and common fingerprints by two different heads (H_s and H_c) to classify the forgery method and determine whether the image is real or fake, respectively.
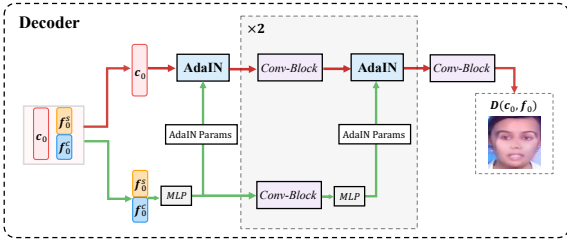

Figure 4: The architecture of our decoder D, involves combining the fingerprint and content through AdaIN layers, which are then processed through multiple convolutional blocks along with upsampling layers (indicated as "Conv-Block" in the figure). The AdaIN layers are utilized twice during this process to fuse the fingerprint as a condition along with the content. Ultimately, the output of the final "Conv-Block" layer is decoded to reconstruct the image.

3.2. Disentanglement Framework

Our disentanglement framework, depicted in Fig. 3, consists of an encoder, a decoder, and two classification heads. The encoder comprises a content encoder and a fingerprint encoder that extract content and fingerprint features, respectively. While the two encoders share the same structure, they do not share parameters. The decoder includes multi-

ple convolutional and upsampling layers that reconstruct the image by utilizing fingerprint features as a condition along with content features. The classifier consists of two different heads, one for learning method-specific textures and the other for learning generalizable features across different forgeries. More details about our encoder can be found in the supplementary material.

3.3. Architecture

Encoders. Our encoder processes a pair of images (x_0, x_1), where x_0 represents the fake image and x_1 represents the real image. The encoder E comprises a content encoder E_c and a fingerprint encoder E_f, extracting the content and forgery features, respectively. We apply the encoder to each pair to obtain the corresponding fingerprint and content features as follows:

$$f_i^s,\ f_i^c,\ c_i = E(x_i), \tag{1}$$

where $i \in \{0, 1\}$ is the index of the image. Also, f_0^s, f_0^c, c_0 and f_1^s, f_1^c, c_1 are denoted by the specific fingerprint, common fingerprint, and content corresponding to each input image pair, respectively.

Decoders. Our decoder (See Fig. 4) reconstructs an image by utilizing its content and fingerprint through a series of upsampling and convolutional layers. Unlike other disentanglement-based deepfake detection frameworks [54, 51, 27] that linearly add forgery and content features for recombination, our decoder applies Adaptive Instance Normalization (AdaIN) [19] for improved reconstruction and decoding, inspired by stylization techniques [20]. The AdaIN aligns the mean and variance of the content code to match those of the fingerprint. The formula is written as:

$$\text{AdaIN}(\boldsymbol{c}, \boldsymbol{f}) = \sigma(\boldsymbol{f}) \left(\frac{\boldsymbol{c} - \mu(\boldsymbol{c})}{\sigma(\boldsymbol{c})} \right) + \mu(\boldsymbol{f}), \quad (2)$$

where $\boldsymbol{c}$ and $\boldsymbol{f}$ are the content and the style vectors of the image pair, respectively. The functions $\mu(\cdot)$ and $\sigma(\cdot)$ compute the mean and variance of the input.

3.4. Objective Function

To attain disentangled feature representation for detection, we design three distinct loss functions: two classification losses for common and specific forgery features, a contrastive regularization loss for similar and dissimilar image embeddings, and a reconstruction loss to ensure consistency between original and reconstructed images at the pixel level. These losses are combined in a weighted sum to create the overall loss function for training the framework.

Multi-Task Classification Loss. For classification loss, we propose two different losses of classification. First, we propose a binary classification loss $\mathcal{L}_{ce}^c$ computed by the cross-entropy for supervising the model to learn the common feature of different forgery methods:

$$\mathcal{L}_{ce}^c = \mathcal{L}_{ce}(\boldsymbol{H}_c(\boldsymbol{f}_i^c), \ \boldsymbol{y}_i), \quad (3)$$

where $\mathcal{L}_{ce}$ denotes the cross-entropy loss, $\boldsymbol{H}_c$ is the head for the common forgery feature, which is implemented by several MLP layers. $\boldsymbol{y}_i \in \{\text{fake}, \text{real}\}$ is the binary classification label. In addition, $\mathcal{L}_{ce}^s$ is proposed to learn the method-specific patterns by guiding the model to identify which forgery method is applied to the fake image:

$$\mathcal{L}_{ce}^s = \mathcal{L}_{ce}(\boldsymbol{H}_s(\boldsymbol{f}_i^s), \ \boldsymbol{y}_i'), \quad (4)$$

where $\boldsymbol{H}_s$ is the head for specific forgery feature and $\boldsymbol{y}_i' \in \{\text{real}, GAN_1, GAN_2, \cdots\}$ donates the label for identifying which instance belong to the image. Note that $\boldsymbol{H}_c$ and $\boldsymbol{H}_s$ share the same architecture but not share the parameters.

The multi-task classification loss enables the model to learn both method-specific textures and common features in different forgeries, enhancing the generalization ability of the model.

Contrastive Regularization Loss. The objective of contrastive regularization loss is to optimize the similarity and dissimilarity measurements between images. The contrastive regulation loss is formulated mathematically as:

$$\mathcal{L}_{con} = \max\left(|\boldsymbol{x}_A - \boldsymbol{x}_P|_2 - |\boldsymbol{x}_A - \boldsymbol{x}_N|_2 + \alpha, \ 0\right), \quad (5)$$

where α serves as a margin hyper-parameter. This method minimizes the Euclidean distance between an anchor image ($\boldsymbol{x}_A$) and its similar counterpart ($\boldsymbol{x}_P$), while simultaneously maximizing the gap between the anchor image and its dissimilar counterpart ($\boldsymbol{x}_N$). For instance, if $\boldsymbol{x}_A$ denotes common features of a genuine image, $\boldsymbol{x}_P$ would represent common attributes of another real image, while $\boldsymbol{x}_N$ signifies the attributes of a manipulated image.

For the common features, we compute the loss between real and fake images to encourage the model to learn a generalizable representation that across different forgeries. For the specific features, we compute the loss between images of the same forgery to encourage the model to learn method-specific textures for each forgery.

Reconstruction Loss. In general, there are two types of reconstruction in our framework: self-reconstruction and cross-reconstruction. For the self-reconstruction, the decoder $\boldsymbol{D}$ is applied to the content and fingerprint encoded from the same image to reconstruct the corresponding image, and the formula is written as:

$$\mathcal{L}_{rec}^s = \|\boldsymbol{x}_0 - \boldsymbol{D}(\boldsymbol{f}_0, \boldsymbol{c}_0)\|_1 + \|\boldsymbol{x}_1 - \boldsymbol{D}(\boldsymbol{f}_1, \boldsymbol{c}_1)\|_1. \quad (6)$$

For the cross-reconstruction, we consider the different combinations of fingerprint and content features encoded from the different images. Similarly, the formula of the cross-reconstruction loss is written as:

$$\mathcal{L}_{rec}^c = \|\boldsymbol{x}_0 - \boldsymbol{D}(\boldsymbol{f}_1, \boldsymbol{c}_0)\|_1 + \|\boldsymbol{x}_1 - \boldsymbol{D}(\boldsymbol{f}_0, \boldsymbol{c}_1)\|_1. \quad (7)$$

Considering both the self-reconstruction and cross-reconstruction loss, the overall image reconstruction loss can be computed as follows:

$$\mathcal{L}_{rec} = \mathcal{L}_{rec}^s + \mathcal{L}_{rec}^c. \quad (8)$$

The image reconstruction loss ensures that the reconstructed image and the original image are consistent at the pixel level. In addition, the two reconstruction losses enhance the disentanglement of features. The self-reconstruction loss penalizes the reconstruction errors by leveraging the latent features of the input image, while the cross-reconstruction loss penalizes the reconstruction errors using the forgery feature and the swapped content features.

Overall Loss. The final loss function of the training process is the weighted sum of the above loss functions.

$$\mathcal{L} = \mathcal{L}_{ce}^c + \lambda_1 \mathcal{L}_{ce}^s + \lambda_2 \mathcal{L}_{rec} + \lambda_3 \mathcal{L}_{con}, \qquad (9)$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyper-parameters for balancing the overall loss.

4. Experiments

4.1. Settings

Datasets. To evaluate the generalization ability of the proposed framework, our experiments are conducted on four large-scale benchmark databases: FaceForensics++ (FF++) [37], DeepfakeDetection (DFD) [11], Deepfake Detection Challenge (DFDC) [12], and CelebDF [26]. FF++ [37] is a large-scale database comprising more than 1.8 million forged images from 1000 pristine videos. Forged images are generated by four face manipulation algorithms using the same set of pristine videos, *i.e.,* Deep-Fakes (DF) [1], Face2Face (F2F) [45], FaceSwap (FS) [2], and NeuralTexture (NT) [44]. To evaluate the generalization ability of our framework, we follow prior research works [23, 7] and conduct experiments on three widely used face-manipulated datasets, *i.e.,* DFDC [12], CelebDF [26], and DFD [11]. Note that there are three versions of FF++ in terms of compression level, *i.e.,* raw, lightly compressed (HQ), and heavily compressed (LQ). Since realistic forgeries often have a limited quality, the HQ and LQ versions are used in experiments. Following previous works [23, 7], **the HQ version of FF++ is adopted by default.** If any deviation from this default, it will be explicitly stated.

Implementation. We use a modified version of Xception [37] as the backbone network, with model parameters initialized by pre-training on ImageNet. Face extraction and alignment are performed using DLIB [39]. Following previous works [7], the aligned faces are resized to 256×256 for both the training and testing. We use the Adam [21] for optimization with the learning rate of 0.0002, and the batch size is fixed as 32. In the overall loss function in Eq. (9), we set λ_1 to λ_3 as 0.1, 0.3, 0.05 empirically. The margin α in Eq. (5) is set to 3. We also apply some widely used data augmentations, *i.e.,* image compression, horizontal flip, and random brightness contrast.

Evaluation Metrics. We report the Area Under Curve (AUC) metric to compare our proposed method with prior works, which is consistent with the evaluation approach adopted in many previous works [23, 36, 37, 30, 7]. **The default evaluation metric employed is the AUC.** We also report other metrics such as Accuracy (ACC), Average Precision (AP), and Equal Error Rate (EER) for a more comprehensive evaluation of our method. Please refer to our supplementary for more details.

4.2. Generalization Ability Evaluation

Comparison with competing methods. To assess the generalization capacity of our framework, **we reproduce ten competing methods under consistent conditions for a comprehensive comparison:** Xception [37], Face X-ray [23], F3Net [36], SRM [30], SPSL [28], RECCE [6], CORE [35] , SLADD [7], and Liang *et al.* [27]. We use the provided codes of Xception, RECCE, SLADD, CORE, FWA, and SRM from the authors. We reimplement Face X-ray, F3Net, SPSL, and Liang *et al.* [27] rigorously following the companion paper's instructions and train these models under the same settings.

We conduct this experiment by training the models on the FF++ [37] and then evaluate these models in DFD [11], DFDC [12], and CelebDF [26], respectively. This setting is challenging in generalization ability evaluation since the testing sets are collected from different sources and share much less similarity with the training set.

The results of the comparison between different methods are presented in Tab. 1, which shows the performance in terms of the AUC metric. It is evident that the proposed disentanglement framework and multi-task learning strategy lead to superior performance compared to other models in most cases, achieving the overall best results.

Liang *et al.* [27] proposes a disentanglement framework for content information removal, but their model is still prone to overfitting to method-specific patterns, leading to the limitation of the generalization. On the contrary, the disentanglement framework we proposed is designed to learn generalizable features across different forgeries by the multi-task learning strategy, thereby achieving improved generalization performance.

Face X-ray [23] and FWA [25] use blended artifacts in forgeries to achieve generalization. However, these two methods have limited generalization ability when the patterns in the training and testing datasets differ. This is because Face X-ray learns to identify the boundary patterns that are sensitive to the post-processing operations varying in different datasets. On the contrary, our proposed framework learns common representations that are not dependent on specific post-processing operations.

SRM [30], SPSL [28], and F3Net [36] utilize frequency components of images to distinguish between forgeries and pristine images. However, the experimental results show that their generalization performance is inferior to the proposed approach. This could be due to the fact that these frequency cues that are effective on the FF++ may not generalize to other datasets with different post-processing steps.

CORE [35], RECCE [6], SLADD [7] are recent detectors that focus on different detection algorithms: loss design, reconstruction learning, and adversarial training. However, these detectors could still be disrupted by unrelated factors such as race, gender, or identity because they

Method	FF-ALL			FF-wo-DF			FF-wo-F2F			FF-wo-FS			FF-wo-NT			Avg.
	DFDC	CelebDF	DFD	DFDC	CelebDF	DFD	DFDC	CelebDF	DFD	DFDC	CelebDF	DFD	DFDC	CelebDF	DFD	
Xception [37]	0.651	0.672	0.727	0.651	0.660	0.633	0.646	0.716	0.794	0.665	0.737	0.826	0.647	0.709	0.798	0.702
Liang *et al.* [27]	0.700	0.706	0.829	0.707	0.699	0.794	0.705	0.698	0.844	0.709	0.713	0.851	0.667	0.672	0.750	0.736
CORE [35]	0.658	0.708	0.917	0.630	0.644	0.807	0.671	0.708	0.923	0.663	0.711	0.925	0.653	0.689	0.920	0.748
RECCE [6]	0.635	0.756	<u>0.933</u>	0.636	0.604	0.821	0.661	0.724	**0.930**	0.651	0.778	<u>0.928</u>	0.659	0.754	0.932	0.760
FWA [25]	0.650	0.755	0.870	0.635	**0.771**	<u>0.885</u>	0.701	0.778	0.924	0.689	0.752	0.850	0.670	0.744	0.875	0.770
Face X-ray [23]	0.710	0.740	0.890	0.726	0.668	0.838	0.734	0.716	0.899	0.705	0.693	0.907	0.731	0.731	0.901	0.773
SLADD [7]	0.751	0.753	0.900	0.738	0.705	0.841	0.757	0.754	0.815	<u>0.713</u>	0.733	0.883	0.754	0.741	0.894	0.782
F3Net [36]	0.743	0.668	0.926	0.748	0.648	0.872	**0.765**	0.692	0.914	**0.719**	0.680	0.925	0.779	0.760	<u>0.933</u>	0.785
SRM [30]	<u>0.771</u>	0.770	0.915	0.743	0.746	0.874	0.745	0.757	0.909	0.699	0.768	0.923	0.778	0.779	0.891	0.805
SPSL [28]	0.742	<u>0.787</u>	0.927	<u>0.749</u>	<u>0.753</u>	**0.898**	<u>0.759</u>	**0.797**	<u>0.929</u>	0.664	<u>0.794</u>	0.924	<u>0.797</u>	**0.813**	0.924	<u>0.817</u>
Ours	**0.805**	**0.824**	**0.945**	**0.767**	0.749	0.870	**0.765**	<u>0.782</u>	0.908	0.711	**0.800**	**0.943**	**0.800**	<u>0.808</u>	**0.943**	**0.828**

Table 1: Comparisons of generalization ability with competing methods implemented by ourselves. We use two different data configurations: "FF-ALL", which includes all data generated by four forgeries, and "FF-wo-DF", "FF-wo-F2F", "FF-wo-FS", and "FF-wo-NT", which use the FF++ dataset but drop DF, F2F, FS, and NT, respectively. The best results are highlighted in bold font, while the second-best results are underlined.

Table 2: Comparison with state-of-the-art methods on CelebDF and DFDC. The results of other works are mainly cited from [7, 53, 50]

Model	Training Set	CelebDF	DFDC
Two-stream [58]	FF++	0.538	-
Meso4 [3]	Self-made	0.548	0.497
MesoInception4 [3]	Self-made	0.536	0.499
DSP-FWA [25]	FF++	0.646	0.646
VA-MLP [32]	FF++	0.550	-
Multi-task [33]	FF++	0.543	-
Headpose [52]	UADFV	0.546	-
Capsule [34]	FF++	0.575	0.575
SMIL [24]	FF++	0.563	0.563
Two-branch [31]	FF++	0.734	0.734
Schwarcz *et al.* [40]	FF++	0.667	0.673
PEL [15]	FF++	0.692	0.633
MADD [55]	FF++	0.674	-
Local-relaion [9]	FF++	0.783	0.765
CFFs [53]	FF++	0.742	0.721
Zhuang *et al.* [60]	FF++	0.728	-
SFDG [50]	FF++	0.758	0.736
Ours	FF++	**0.824**	**0.805**

Table 3: Comparison on FF++ with methods using disentanglement learning.

Training	Method	Testing AUC			
		FF++(LQ)	CelebDF	DFD	DFDC
FF++	Xception [37]	0.683	0.672	0.727	0.651
	Liang *et al.* [27]	0.714	0.706	0.829	0.700
	Ours	**0.833**	**0.824**	**0.945**	**0.805**

to enhance the independence of disentangled features. To ensure a fair comparison, we carefully implement their framework by following the settings of the original paper, as it is not available as an open-source resource. We train the baseline Xception, Liang *et al.* [27], and ours on the FF++ and evaluated them on FF++ (LQ), CelebDF, DFD, and DFDC. As reported in Tab. 3, we observe that Liang *et al.* [27] improve upon the baseline largely, demonstrating the essential of removing content information. Additionally, UCF outperforms Liang *et al.* [27] on all testing datasets, showing the efficacy of uncovering common features.

Comparison with state-of-the-art methods. We further evaluate our method against other state-of-the-art models. The results, as shown in Tab. 2, demonstrate the effective generalization ability of our framework as it outperforms other methods, achieving the best performance in terms of the AUC metric on both CelebDF and DFDC. The results of some methods are directly cited from [7, 53, 50]. **Following a comprehensive evaluation against 27 state-of-the-art detectors (10 implemented in this study and 17 referenced), we demonstrate the robust generalization capability of our proposed framework.** It is worth noting that both UCF and Zhuang *et al.* [60] aim to tackle the challenging issue of overfitting to method-specific artifacts. However, their technical methodologies are totally different (disentangle vs. adversarial learning). Actually, we offer a fresh and distinct solution to this problem. Moreover,

operate in the entire feature space, which inevitably includes these unrelated aspects (similarly indicated in previous work [54]). From Tab. 1, our UCF (82.8%) largely outperforms SLADD (78.2%) in terms of the average AUC.

Finally, Xception [37] serves as a CNN baseline and does not incorporate any augmentation, disentanglement, feature engineering, or frequency information. Its performance drops dramatically in the case of unseen forgeries, highlighting the importance of incorporating these techniques in face forgery detection models.

Comparison with disentanglement-based methods. For disentanglement-based detection frameworks, we identify one prior work, Liang *et al.* [27], that shares similarities with our approach. Their framework aims to remove content information and introduces two modules

Table 4: Ablation study regarding the effectiveness of our disentanglement framework, multi-task learning strategy, and contrastive regularization loss. "D" and "M" represent our basic disentanglement framework and the multi-task learning module, respectively. "C" represents the contrastive learning module. Results in gray indicate the within-dataset performance.

Training	Method	Testing AUC			
		FF++	CelebDF	DFD	DFDC
FF++	Xception [37]	0.986	0.672	0.727	0.651
	Xception + D	0.995	0.785	0.933	0.772
	Xception + D + M	0.995	0.804	0.944	0.785
	Xception + D + M + C	**0.996**	**0.824**	**0.945**	**0.805**

UCF (82.4%) significantly outperforms Zhuang *et al*. [60] (72.8%) on CelebDF in terms of AUC.

4.3. Ablation Study

Effects of our disentanglement framework and multi-task learning strategy. To evaluate the impact of the proposed disentanglement framework and multi-task learning strategy on generalization ability, we conduct an ablation study on several datasets. Specifically, we train all models on FF++ and evaluate their performance on FF++ [37], DFD [11], DFDC [12], and CelebDF [26]. The results are reported in Tab. 4 using the AUC metric. The evaluated variants include the baseline Xception, Xception with the proposed disentanglement framework (Xception + D), the proposed disentanglement framework with the multi-task learning strategy (Xception + D + M), and the multi-task disentanglement framework with the contrastive regularization (Xception + D + M + C).

Regarding the ablation study, we observed the following. Firstly, the four variants achieve relatively similar results on FF++ (within-dataset evaluation). Secondly, implementing the basic disentanglement framework leads to a significant improvement in DFD, DFDC, and CelebDF (cross-dataset evaluation), indicating the generalization ability is improved largely when applying the proposed disentanglement framework for the content information removal. Thirdly, the multi-task disentanglement outperforms the basic disentanglement, indicating that the multi-task learning strategy is effective in improving the generalization ability of the model. Finally, combining the proposed multi-task disentanglement with the contrastive regularization loss achieves the best results in both within-dataset and cross-dataset evaluations, supporting the effectiveness of each module.

Effects of our conditional decoder. In contrast to other disentanglement-based detection frameworks [54, 51, 27] that use linear addition to combine fingerprint and content features for recombination, our proposed decoder utilizes

Table 5: Ablation study regarding the effectiveness of the conditional decoder. "CD" represents that we use the conditional decoder for image reconstruction. Otherwise, we use the linearly add for combining the fingerprint and content. Results in gray indicate the within-dataset performance.

Training	Method	Testing AUC			
		FF++	CelebDF	DFD	DFDC
FF++	Ours	0.995	0.811	0.932	0.786
	Ours + CD	**0.996**	**0.824**	**0.945**	**0.805**

Table 6: Comparison with binary classification results of different forgery features.

Training	Method	Testing AUC			
		FF++	CelebDF	DFD	DFDC
FF++	Xception [37]	0.986	0.672	0.727	0.651
	Specific Forgery	0.987	0.681	0.842	0.667
	Whole Forgery	0.995	0.785	0.933	0.772
	Common Forgery	**0.996**	**0.824**	**0.945**	**0.805**

Table 7: Comparing the performance of the baseline and our proposed framework using different backbones. The best result is highlighted in bold font. "Avg." represents the average AUC for cross-datasets.

Training	Method	Testing AUC				
		FF++(LQ)	CelebDF	DFD	DFDC	Avg.
FF++	Xception [37]	0.683	0.672	0.727	0.651	0.683
	Ours (Xception)	0.833	0.824	0.945	**0.805**	0.852
	ConvNext [29]	0.779	0.788	0.912	0.753	0.808
	Ours (ConvNext)	**0.845**	**0.869**	**0.946**	0.802	**0.866**

AdaIN [19] to incorporate the fingerprint as a condition with the content for improved reconstruction and decoding. To evaluate the impact of the conditional decoder on the generalization ability, we conduct an ablation study on the proposed framework with and without the conditional decoder. Results in Tab. 5 demonstrate that our proposed conditional decoder can achieve improved performance on both within- and cross-datasets, highlighting the importance of using AdaIN layers for reconstruction and decoding.

Comparison with binary classification results of different forgery features. To evaluate the effectiveness of the proposed multi-task learning strategy, binary classification results are compared based on the common, specific, and whole forgery features. Tab. 6 shows that the common features exhibit superior generalization performance compared to the specific features. The comparison of the common and whole forgery features reveals that the whole forgery features are not as effective as the common features, mainly due to the presence of specific features, which may lead to overfitting to method-specific textures.
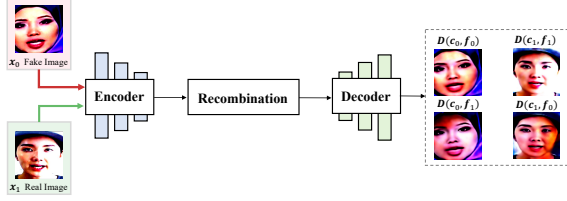
Figure 5: Visualization of the reconstruction images during the training process.

Exploring Generalization performance of different backbone choices. In this section, we investigate the choice of backbone on the generalization ability of our proposed framework. We use Xception as the backbone in our previous experiments to align with other related works, but our multi-task framework is not limited to this choice. To evaluate the effectiveness of our framework with different backbones, we adopt a recent SOTA backbone ConvNeXt [29] and conduct an ablation study. The results of the ablation study, shown in Tab. 7, demonstrate that our proposed framework can largely improve the generalization performance of both Xception and ConvNeXt backbones. This suggests that our framework is effective and applicable to different backbone choices. Additionally, to further highlight the plug-and-play nature of our proposed framework, we extend its application to ResNet [17] and EfficientNet [43] backbones. More details can be accessed in our supplementary materials.

4.4. Visualization

Visual examples of reconstructed images. Within the framework we propose, the generation of reconstruction images occurs during the recombination phase of training. These images serve a pivotal role in ensuring the effective disentanglement of content and fingerprint features, as depicted in Fig. 5.

The content features within our framework are specifically designed to capture appearance, identity, gender, and other forgery-related features. In the visual examples of the reconstructed images (see Fig. 5), it is evident that the images sharing the same content code exhibit a marked similarity. This resemblance persists even when the fingerprint features, which represent unique identifiers separate from the content, are derived from other individuals. The observed similarity in the reconstructed images with identical content codes substantiates the efficacy of our framework in accurately isolating content features from other elements of the image. Contrary to the content features, the fingerprint features do not alter the content information of the original inputs. However, a close examination of the reconstructed images with the same content code but different fingerprint codes reveals subtle differences.

5. Conclusion

In this paper, we propose a novel disentanglement framework that can generalize well in unseen deepfake datasets. Our approach is grounded in the idea that a generalizable deepfake detector should be able to capture the generalizable features across different types of forgeries. To this end, we introduce a multi-task disentanglement framework to uncover the common features. Additionally, we also introduce a conditional decoder and a contrastive regularization loss to enhance the disentanglement process. In this manner, the model can avoid overfitting to forgery-irrelevant and method-specific forgery textures, leading to a more generalizable detector. To evaluate the effectiveness of our proposed method, we conduct extensive experiments on several benchmark datasets and compare our results against existing state-of-the-art methods. Overall, our proposed framework represents a promising step toward the development of more generalizable deepfake detectors.

Ethics Statement. All facial images utilized in this work are sourced from publicly available datasets and are appropriately attributed through proper citations. Our research adheres to strict ethical guidelines throughout the experimental process. There is no compromise on personal privacy during the experiments conducted in this work.

References

[1] Deepfakes. https://github.com/iperov/DeepFaceLab. Accessed: 2020-05-10.

[2] Faceswap. https://github.com/MarekKowalski/FaceSwap. Accessed: 2020-05-10.

[3] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2018.

[4] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision*, pages 0–0, 2019.

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[6] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2022.

[7] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18710–18719, 2022.

[8] Liang Chen, Yong Zhang, Yibing Song, Jue Wang, and Lingqiao Liu. Ost: Improving generalization of deepfake detection via one-shot test-time training. In *Proceedings of the Neural Information Processing Systems*, 2022.

[9] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[10] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[11] Deepfakedetection. https://ai.googleblog.com/2019/09/contributing-data-to-deepfakedetection.html Accessed 2021-04-24.

[12] Deepfake detection challenge. https://www.kaggle.com/c/deepfake-detection-challenge Accessed 2021-04-24.

[13] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3994–4004, 2023.

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

[15] Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 735–743, 2022.

[16] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[18] Jiashang Hu, Shilin Wang, and Xiaoyong Li. Improving the generalization ability of deepfake detection via disentangled representation learning. In *Proceedings of the IEEE International Conference on Image Processing*, pages 3577–3581. IEEE, 2021.

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *CVPR*, 2017.

[20] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision*, pages 172–189, 2018.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[22] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[23] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[24] Xiaodan Li, Yining Lang, Yuefeng Chen, Xiaofeng Mao, Yuan He, Shuhui Wang, Hui Xue, and Quan Lu. Sharp multiple instance learning for deepfake video detection. In *Proceedings of the ACM International Conference on Multimedia*, 2020.

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

[26] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[27] Jiahao Liang, Huafeng Shi, and Weihong Deng. Exploring disentangled content information for face forgery detection. In *Proceedings of the European Conference on Computer Vision*, pages 128–145. Springer, 2022.

[28] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[29] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[30] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[31] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *Proceedings of the Proceedings of the European Conference on Computer Vision*, 2020.

[32] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 2019.

[33] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019.

[34] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2019.

[35] Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. Core: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, pages 12–21, 2022.

[36] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of the European Conference on Computer Vision*, 2020.

[37] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision*, 2019.

[38] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, 2019.

[39] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Journal of Image and Vision Computing*, 47:3–18, 2016.

[40] Steven Schwarcz and Rama Chellappa. Finding facial forgery artifacts with parts-based detectors. In *CVPRW*, 2021.

[41] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022.

[42] Zekun Sun, Yujie Han, Zeyu Hua, Na Ruan, and Weijia Jia. Improving the efficiency and robustness of deepfakes detection through precise geometric features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[43] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[44] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Journal of ACM Transactions on Graphics*, 38(4):1–12, 2019.

[45] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[46] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008.

[47] Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[48] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 615–623, 2022.

[49] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. *arXiv preprint arXiv:1909.06122*, 2019.

[50] Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7278–7287, 2023.

[51] Tianyun Yang, Juan Cao, Qiang Sheng, Lei Li, Jiaqi Ji, Xirong Li, and Sheng Tang. Learning to disentangle gan fingerprint for fake image attribution. *arXiv preprint arXiv:2106.08749*, 2021.

[52] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.*, 2019.

[53] Peipeng Yu, Jianwei Fei, Zhihua Xia, Zhili Zhou, and Jian Weng. Improving generalization by commonality learning in face forgery detection. *Journal of IEEE Transactions on Information Forensics and Security*, 2022.

[54] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Ying Tai, Shouhong Ding, Jilin Li, Feiyue Huang, Haichuan Song, and

Lizhuang Ma. Face anti-spoofing via disentangled representation learning. In *Proceedings of the European Conference on Computer Vision*, pages 641–657. Springer, 2020.

[55] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[56] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision*, 2021.

[57] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision*, pages 15044–15054, 2021.

[58] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Two-stream neural networks for tampered face detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, 2017.

[59] Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and Stan Z Li. Face forgery detection by 3d decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2929–2939, 2021.

[60] Wanyi Zhuang, Qi Chu, Haojie Yuan, Changtao Miao, Bin Liu, and Nenghai Yu. Towards intrinsic common discriminative features learning for face forgery detection using adversarial learning. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1–6, 2022.