Transfer Learning of Real Image Features with Soft Contrastive Loss for Fake Image Detection

Ziyou Liang[1], Weifeng Liu[1], Run Wang[1*], Mengjie Wu[1], Boheng Li[2], Yuyang Zhang[1], Lina Wang[1], Xinyi Yang[3]

[1]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, China
[2]Nanyang Technological University, Singapore
[3]NSFOCUS, China

Abstract

In the last few years, the artifact patterns in fake images synthesized by different generative models have been inconsistent, leading to the failure of previous research that relied on spotting subtle differences between real and fake. In our preliminary experiments, we find that the artifacts in fake images always change with the development of the generative model, while natural images exhibit stable statistical properties. In this paper, we employ natural traces shared only by real images as an additional target for a classifier. Specifically, we introduce a self-supervised feature mapping process for natural trace extraction and develop a transfer learning based on soft contrastive loss to bring them closer to real images and further away from fake ones. This motivates the detector to make decisions based on the proximity of images to the natural traces. To conduct a comprehensive experiment, we built a high-quality and diverse dataset that includes generative models comprising GANs and diffusion models, to evaluate the effectiveness in generalizing unknown forgery techniques and robustness in surviving different transformations. Experimental results show that our proposed method gives **96.2%** mAP significantly outperforms the baselines. Extensive experiments conducted on popular commercial platforms reveal that our proposed method achieves an accuracy exceeding **78.4%**, underscoring its practicality for real-world application deployment.

Introduction

With the rapid development and maturity of generative models, the increasing proliferation of fake images has attracted widespread attention. Compared to Generative Adversarial Networks (GANs), diffusion models (DMs), as today's SOTA generative models, exhibit better generation quality (Dhariwal 2021) and even support powerful text-to-image models such as DALL-E2 (Ramesh et al. 2022), Stable Diffusion (Rombach et al. 2022). Currently, one can use different types of generative models to create realistic faces or complex scene images and it is foreseeable that more generative models for image synthesis will emerge in the future. Therefore, it is the goal for the community to develop a more practical method to distinguish fake images that are synthesized with unknown forgery techniques as the unseen generative models will emerge inadvertently.

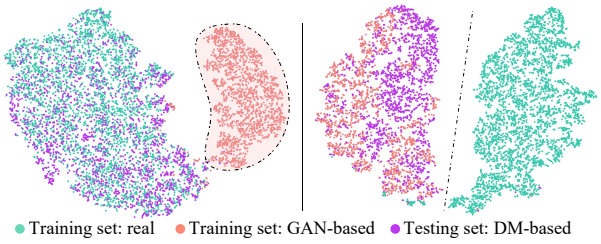

Figure 1: **TSNE visualization. Left**: The detector is easily overfitted to fake images in the training set; **Right**: Training with natural traces can generalize to unknown fake images.

The prior paradigm for fake image detection is to learn artifacts of fake images by capturing the subtle differences between real and fake images (Wang et al. 2019, 2020; Huang et al. 2020; Wang et al. 2021). Our empirical research confirms that there is a huge challenge in detecting unknown fake images: the model can easily form a classification manifold on the images of the training set, while new generative fake images and their distinctive artifacts will be randomly scattered in this space, as shown in Figure 1 left. Unfortunately, the existing studies are trapped in the endless efforts to spot the artifacts by investigating the subtle differences between real and fake (Corvi et al. 2023). Since the classifier can easily overfit the fake image artifacts of the training set, unknown fake images will be scattered outside the manifold like real images.

In this paper, we propose a transfer learning method based on natural trace features for fake image detection. We propose a novel perspective on fake image detection, emphasizing *the inherent similarities among real images rather than the differences between real and fake*. Our insight is grounded in the premise that images from the real world possess a stable intrinsic naturalness. We posit that real images share common characteristics, akin to those found in fake images, which we refer to as "natural traces." We introduce natural trace forensics (NTF) and adopt a substitution strategy to replace shared features with homogeneous

*Corresponding author. E-mail: wangrun@whu.edu.cn

features. Specifically, we exploit a self-supervised feature mapping process to decouple heterogeneous features of real images, allowing for the capture of stable homogeneous features. We then develop a transfer learning for homogeneous features based on a soft contrastive loss, which simultaneously solved the problem that similar clusters formed by self-supervised contrastive learning are still scattered with large distances and outliers under the constraints of supervised contrastive learning do not converge. We jointly optimized transfer training and binary classification to force the classifier to aggregate homogeneous features close to real images and constrain them away from fake images. In Figure 1 right, the classifier reduces its dependence on specific artifact patterns and gains the ability to detect images synthesized by unknown generative models with natural traces.

To better evaluate whether our proposed method can be well generalized to other unknown generative models, we build a generated image dataset consisting of 12 SOTA generative models, including 6 GANs and 6 DMs. Our dataset covers a variety of categories, such as faces and scenes, to evaluate detectors' capability across various types of fake images. In addition, we also evaluate the performance of our method in identifying images generated by Midjourney[1] and Kolors[2], which is currently a popular commercial tool for text-to-image generation. Experimental results show that our proposed method could discriminate GAN-based, DM-based, and Multi-step fake images (synthesized with at least two different generated models) in high confidence with an average accuracy of more than **96.1%** and is sufficiently robust to various image perturbation transformations.

Our main contributions are summarized as follows:

- We are the first to propose soft contrastive transfer learning, which utilizes the disentangled shared features of real images for fake image detection. Our results also demonstrate its effectiveness in detecting previously unseen generative fake images, highlighting its potential as a versatile solution in this domain.
- To conduct a comprehensive evaluation, we build a dataset including GAN-based, DM-based, and Multi-step manipulation for generating fake images. For the first time, we generate Multi-step fake images by employing multiple synthesis methods.
- Experimental results show the effectiveness of our proposed method in tackling fake images generated by SOTA GANs and diffusion models, giving an average precision of more than 96.18%, significantly outperforming the baselines.

Related Work

Transfer Learning in Fake Detection

In recent years, the most common method of transfer learning is to fine-tune the pre-trained model on the target dataset. (Suratkar 2023) detect fake videos through the utilization of transfer learning in autoencoders and a hybrid model of CNN and RNN. (Lee and Kim 2022; Ghayoomi 2022) used pre-trained BiLSTM(Zhou et al. 2016) and RoBerta(Liu et al. 2019) for transfer learning on fake news about COVID-19 in Korean and Persian, respectively. Although the finetuning method is highly flexible for downstream tasks, it is also prone to catastrophic forgetting. Freezing layers of the pre-trained backbone network and finetuning only lateral fully connected layers is one of the most effective transfer learning methods. (Ranjan 2020) explored the effectiveness and interpretability of freezing convolutional layers and fine-tuning fully connected layers in DeepFake detection, and (Elhassan et al. 2022) studied the transfer of lip-motion-based DeepFake detection methods on 11 models. However, the pre-training of these frozen transfer learning methods is obviously isolated from the target, and the parameter transfer effect is limited. As far as we know, there is still a gap in the exploration of transfer learning in fake image detection. Our method balances the feature transfer and classification by additional soft contrastive loss constraints.

Fake Images Detection

Existing research focuses on exploring subtle differences between real and fake images, and these can be categorized into explicit and implicit artifact-based methods.

Explicit-based. Some researchers noticed that they often contained specific artifacts or unnatural patterns(Zhang 2019; Liu et al. 2023). Additionally, several studies focused on exploring GAN-based artifacts in the frequency domain(Frank et al. 2020) and the failure to accurately re-enact certain biological features when generating fake faces(Hu 2021; Tan et al. 2023b). These findings have motivated researchers to use explicit artifacts to detect fake images through simple classifiers. However, as generative models are continuously updated and iterated, these artifacts become imperceptible or even disappear, making detection methods reliant on explicit artifacts less effective.

Implicit-based. Wang *et al.*(Wang et al. 2020) demonstrated that with appropriate training data and data augmentation, neural networks could detect other GAN-based images, while (Tan et al. 2023a) used CNNs to transform images into gradient form to present a broader range of artifacts. These studies indicate that images generated by both GAN and diffusion models possess distinct "fingerprints" different from the real images(Yu, Davis, and Fritz 2019; Sha et al. 2023). Nevertheless, there is still insufficient evidence to prove that generative models from different families have universal fingerprints for detection.

A Diverse Generated Image Dataset

Due to the lack of DM-based generated images in current datasets, we create a dataset with a wide range of generative models. This dataset aims to enhance the evaluation of fake detection methods' capability. It includes fake images generated by various models, alongside an equal number of real images from corresponding training sets for each method.

Our dataset covers two major families: GANs and diffusion models, with each fake image synthesized using one generative model. Particularly, we have developed a novel multi-step fake image generation method, involving collaboration between two or more generative models, to achieve

[1]https://www.midjourney.com/

[2]https://klingai.kuaishou.com/

Family	Type	Method	Year	Image Source	# Images
GAN-based	Unconditional	ProGAN	2017	CelebA-HQ	4.0k
		StyleGAN2	2019	CelebA-HQ/FFHQ/LSUN	12.0k
		ProjGAN	2021	FFHQ/LSUN/Landscape	12.0k
		VQGAN	2020	CelebA-HQ	4.0k
		Diff-StyleGAN2	2022	FFHQ/LSUN	8.0k
	Image-to-Image	SimSwap	2020	CelebA-HQ	4.0k
DM-based	Unconditional	DDPM	2020	CelebA-HQ/LSUN	8.0k
		DDIM	2021	CelebA-HQ	4.0k
		PNDM	2022	CelebA-HQ/LSUN	12.0k
	Image-to-Image	DiffFace	2022	CelebA-HQ	2.0k
	Prompt-guided	LDM	2022	CelebA-HQ/LAION	8.0k
		SDM	2022	LAION	4.0k
Multi-step	GAN-GAN	SimSwap_Style2	2024	CelebA-HQ	2.0k
		SimSwap_VQ	2024	CelebA-HQ	2.0k
	GAN-DM	SimSwap_LDM	2024	CelebA-HQ	2.0k
		DiffFace_Style2	2024	CelebA-HQ	2.0k
		DiffFace_Proj	2024	FFHQ	2.0k
	DM-DM	DiffFace_LDM	2024	CelebA-HQ	2.0k

Table 1: Statistics of the self-built dataset, including GAN-based, DM-based, and Multi-step synthesis.

identity swapping or attribute editing between real and fake faces. To ensure diversity, the dataset comprises various categories of generation methods: unconditional generation, image-to-image generation, and prompt-guided generation, as shown in Table 1. Moreover, our carefully selected generative models exhibit fundamental differences in their generations, ensuring extensive representation of the dataset.

GAN-based forgery: We select six representative GANs to generate forgery images including unconditional and image-to-image. The innovation of ProGAN (Karras et al. 2017) in introducing progressive training has a significant impact on subsequent research, while StyleGAN2 (Karras et al. 2020) refines style control through decoupling image features and generates more realistic images. In terms of architectural improvements, ProjGAN (Sauer et al. 2021) enhances generator feedback using pre-trained weights, while VQGAN (Esser 2021) and Diff-StyleGAN2 (Wang et al. 2022) innovatively replace the backbone network with transformer and diffusion processes, respectively, resulting in higher image quality and more stable training processes. SimSwap's (ss)(Chen et al. 2020) ID injection module achieves breakthroughs in arbitrary face swapping.

DM-based forgery: DM-based surpasses GAN-based in terms of image quality and diversity. Here, we select six different DMs for creating fake images including unconditional, image-to-image, and prompt-guided. DDPM(Ho 2020), as the initial diffusion model, lays the groundwork, with DDIM(Song 2020) and PNDM(Liu et al. 2022) enhancing execution speed and quality. DiffFace (df)(Kim et al. 2022) is the first identity-conditioned DDPM that uses diffusion models for face swapping. And LDM(Rombach et al. 2022), which employs pre-trained self-encoders to map pixels to latent space, combined with context learning in cross-attention layers for prompt-guided image generation, along with Stable Diffusion (SDM), a popular LDM-based prompt-guided model, represent further advancements.

Multi-step: In the real scenario, the creator tends to employ multiple forgery techniques to achieve better forgery. We create a Multi-step face synthesis dataset where two face-swapping methods (SimSwap and DiffFace) based on GAN or diffusion models swap real faces onto synthetic ones, including three hybrid modes: GAN-GAN, GAN-DM, and DM-DM. This generation method provides simulated threats of artifact disappearance or blending in real-world scenarios.

Our Method

In this work, we propose a novel method, named Natural Trace Forensics (NTF), which involves training the classifier using the natural traces shared merely by real images as an additional predictive target. Figure 2 overviews the pipeline of NTF. We start by learning natural trace representations from real datasets. Then, under a soft contrastive learning (SCL) framework, the network is trained to align natural traces closer to real images and further from fake ones. With such constraints, the network is motivated to detect fakes based on the distance between images and natural traces. Next, we elaborate on how to learn the natural traces and apply the extracted traces for detection.

Natural Trace Representation Learning

We first explore the natural traces in real images to provide learnable features for the next fake image identification. However, it is impossible to analyze every real image in existence to identify shared features. To address this, we employ an innovative strategy using the same intrinsic features found in real images as a substitute for shared features. These intrinsic features, known as homogeneous features, are derived from the inherent properties and statistical regularities of images, which are commonly present in real images. We develop a self-supervised feature mapping mechanism to extract the homogeneous features. This mechanism decouples the natural image features into homogeneous features and heterogeneous features, where the latter are associated with specific images. As opposed to direct embedding features, feature decoupling not only enhances generalization by ensuring homogeneous features more accurately capture commonalities across images, but also improves feature quality by eliminating noise and redundant information. Additionally, the network should access a large variety of natural images. This exposure enables the network to learn the patterns of feature coupling in various types of real images so that it can decouple features for unseen images.

Formulation. We assume access to a large-scale dataset of real images D_r. Sample $x \in D_r$ is an augmented sample from an image $x_r \in \mathbb{R}^{H\times W\times 3}$. As shown in Figure 2(a), our architecture consists of a feature encoder followed by two projection heads that map real image embeddings into homogeneous and heterogeneous features.

Specifically, feature encoder, E, producing feature embedding $e = E(x_r)$ from the inputs, then decouples the e into homogeneous and heterogeneous features, $z^{hom} = f^{hom}(e) \in \mathbb{R}^C$ and $z^{het} = f^{het}(e) \in \mathbb{R}^C$ through two projection heads, f^{hom} and f^{het}, respectively, where C is the dimensionality of the features. Let $i \in I \equiv \{1...2N\}$

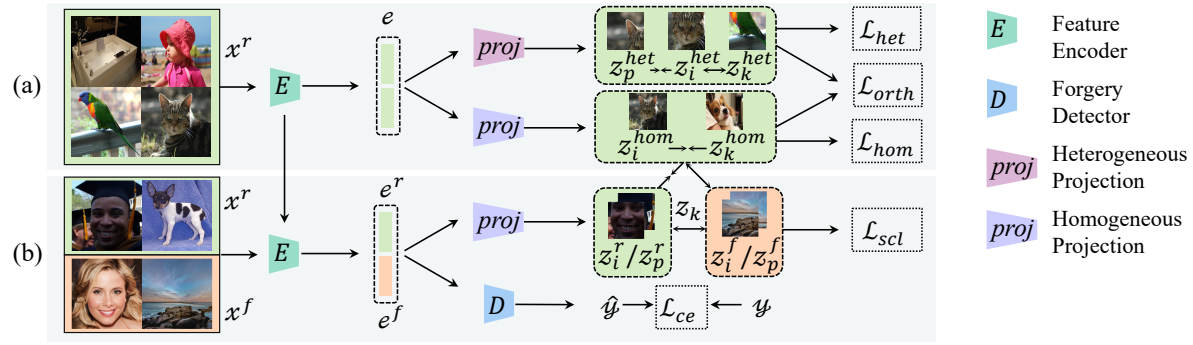

Figure 2: **NTF architecture**. We first decouple the feature representation of real images into homogeneous and heterogeneous features. Next, the homogeneous features z^{hom} participate in SCL with the real and fake image features. Detector classifies real and fake images guided by the target of intra-class aggregation and inter-class separation. Better view in color.

Category	Method	ProGAN				StyleGAN2				VQGAN				ProjGAN				Diff-StyleGAN2				SimSwap				Average			
		AP ↑	ACC ↑	FPR ↓	FNR ↓	AP	ACC	FPR	FNR	AP	ACC	FPR	FNR	AP	ACC	FPR	FNR	AP	ACC	FPR	FNR	AP	ACC	FPR	FNR	AP	ACC	FPR	FNR
Explicit-based	SBIs*	68.6	66.0	39.8	28.2	65.5	60.2	41.8	37.9	56.9	46.2	76.9	30.8	59.9	50.0	53.9	46.2	58.5	54.9	58.3	32.0	55.3	53.9	**33.1**	59.1	60.8	55.2	41.8	41.9
	CADDM*	51.5	53.4	40.6	52.6	52.3	54.4	46.8	44.3	52.6	54.5	50.7	40.3	54.3	57.6	43.0	41.7	51.3	52.9	44.2	50.0	46.6	36.6	82.9	43.6	52.1	53.3	48.4	45.0
Implicit-based	Xception	99.3	84.5	**0**	31.1	**100**	99.9	0.1	**0**	**100**	**99.9**	0.2	**0**	**100**	**99.1**	1.8	**0**	97.9	84.0	0.2	31.8	**99.9**	90.1	19.7	**0**	99.5	**92.9**	3.7	10.5
	Wang2020	97.9	85.3	27.1	2.0	86.2	56.3	87.3	0.1	61.2	50.5	98.9	0.1	62.2	50.4	99.2	0.1	87.8	68.5	61.3	1.8	98.4	65.0	69.8	0.1	82.3	62.7	73.9	0.7
	Grag2021	**100**	**99.9**	**0**	**0.1**	**100**	**100**	**0**	**0**	99.9	91.6	16.8	0.1	99.7	94.6	10.8	**0**	99.8	**94.5**	11.1	**0**	94.1	50.5	98.9	**0**	98.9	88.5	22.9	**0**
	Ojha2023	98.8	99.1	1.5	0.5	83.1	87.0	15.9	10.1	83.6	87.3	16.1	9.4	75.2	78.7	33.2	9.4	64.6	64.9	69.3	0.8	80.3	83.8	23.5	8.9	80.9	83.5	26.6	6.5
	NTF(Ours)	**100**	92.5	**0**	14.9	**100**	92.5	**0**	15.0	**100**	92.6	**0**	14.9	99.8	92.2	**0.7**	15.1	**99.9**	92.7	**0.2**	14.6	99.8	**92.2**	**0.6**	15.1	**99.9**	92.4	**0.2**	14.9

Table 2: **Intra-family generalization on GANs**. Performance of NFT and baselines in spotting 6 GAN-based generated images. These baselines include the classic approach, XcetionNet, Wang2020, and Grag2021, as well as the latest methods, Ojha2023, and CADDM. The method with $*$ means that the baseline is only evaluated on non-face images. The *Average* column represents the weighted average of the corresponding metrics. Among those, the best and second-best performances are highlighted in **bold** and underlined, respectively.

be the index of an arbitrary augmented sample, where there are a total of N samples in a batch, each with two random augmentations ($2N$). The self-supervised contrastive loss of heterogeneous (het.) feature representations in real images can be formulated as follows:

$$\mathcal{L}_{het} = -\sum_{i \in I} \log \frac{exp(z_i^{het} \cdot z_p^{het}/\tau)}{\sum_{k \in K(i)} exp(z_i^{het} \cdot z_k^{het}/\tau)}, \quad (1)$$

where the index i is called the anchor, the index p is called the *positive*, τ is a temperature hyperparameter and $K(i) \equiv I \backslash \{i\}$, which contains all the augmented samples except i in a batch. Furthermore, the loss of homogeneous (hom.) feature representations in real images can be formulated as:

$$\mathcal{L}_{hom} = \arg \max_{i \in I, k \in K'(i)} \parallel z_k^{hom} - z_i^{hom} \parallel_F^2, \quad (2)$$

where $\| \cdot \|_F$ denotes the Frobenius norm, K' contains $2N-2$ augmented samples, *i.e.*, target i is compared only with samples from different sources.

Considering the potential issues of high feature coupling in the embedding space, we further use soft orthogonality (orth.) to reduce information redundancy and dependencies between these homogeneous and heterogeneous features:

$$\mathcal{L}_{orth} = \sum_{i \in I} \cos(z_i^{hom}, z_i^{het}), \quad (3)$$

Finally, we combine these constraints to form the natural trace representation learning loss:

$$\mathcal{L}_{tra} = \mathcal{L}_{hom} + \mathcal{L}_{het} + \lambda \mathcal{L}_{orth}, \quad (4)$$

where λ is a scaling factor.

Transfer Learning for Fake Image Detection

To capture the natural traces as an additional target, we develop the SCL to simultaneously encode real and fake images further for fake image detection. Specifically, we incorporate these natural traces into SCL and constrain their distance from positive and negative samples. This will motivate the detector to identify fake images based on distance. Note that the feature encoder is frozen during this stage.

Formulation. We now assume access a full dataset, $D = D_r \bigcup D_f$, where D_r is used in the previous stage, D_f is a dataset of fake images. Our architecture consists of a feature encoder, an auxiliary projection head for SCL, and another classification head for supervised classification(shown in Figure 2(b). With the auxiliary projection head, the real and fake feature embeddings are mapped to z_i^r and z_i^f, respectively. To motivate the network to focus on the intra-class aggregation of real images more than the inter-class differences between real and fake images, we adopt the homogeneous features z^{hom} from the previous stage as extra positive instances for real anchors and additional negative instances for fake anchors. SCL mitigates the negative im-

Category	Method	DDPM				DDIM				PNDM				LDM				SDM				DiffFace				Average			
		AP ↑	ACC ↑	FPR ↓	FNR ↓	AP	ACC	FPR	FNR	AP	ACC	FPR	FNR	AP	ACC	FPR	FNR	AP	ACC	FPR	FNR	AP	ACC	FPR	FNR	AP	ACC	FPR	FNR
Explicit-based	SBIs*	63.1	64.0	37.5	34.5	71.1	69.5	15.5	47.6	64.2	59.2	48.5	33.0	83.8	75.7	14.6	34.0	57.0	55.8	59.2	29.1	70.6	56.9	28.0	28.4	68.3	63.5	33.9	34.4
	CADDM*	53.0	55.2	46.5	43.2	55.8	59.8	35.8	44.6	51.8	52.5	51.9	43.1	52.3	53.9	48.3	43.9	51.6	54.5	38.3	52.5	53.0	52.6	43.9	44.2	52.9	54.7	44.1	45.2
Implicit-based	Xception	74.7	**72.3**	**5.1**	50.2	75.8	73.9	**12.4**	39.9	**89.6**	49.8	99.9	0.6	65.8	65.0	8.8	61.2	41.1	40.3	28.5	91.0	59.2	40.8	99.5	19.0	66.2	57.0	**14.0**	59.0
	Wang2020	58.2	50.2	99.7	0.1	60.9	50.2	99.6	0.1	66.4	50.0	99.6	0.1	67.9	50.4	99.2	**0**	39.8	49.5	99.6	1.5	34.8	49.8	99.8	2.4	54.6	50.0	99.6	0.7
	Grag2021	**79.9**	50.1	99.8	**0**	83.7	50.3	99.5	**0.1**	86.5	50.1	99.3	**0**	**99.9**	**98.8**	2.5	**0**	98.2	63.6	72.9	**0**	63.2	54.4	91.3	**0**	85.2	61.2	77.5	**0**
	Ojha2023	64.2	67.0	57.4	8.6	71.8	75.3	39.7	9.8	61.7	64.4	62.2	9.0	80.2	83.9	22.9	9.4	55.6	56.0	87.3	0.8	88.4	**88.7**	22.0	0.6	70.3	72.6	48.6	6.3
	NTF(Ours)	76.9	68.8	47.2	15.3	**91.8**	**84.1**	16.4	**15.5**	79.4	**71.0**	**42.7**	15.1	**99.9**	92.7	**0.1**	14.7	**99.5**	**92.1**	**1.1**	14.8	**99.3**	87.4	**0.6**	15.5	**91.2**	**82.7**	18.0	15.1

Table 3: **Cross-family generalization on diffusion models**. Performance of NFT and baselines in spotting 6 DM-based generated images.

Category	Method	SimSwap_Style2				SimSwap_VQ				SimSwap_LDM				DiffFace_Style2				DiffFace_Proj				DiffFace_LDM				Average			
		AP ↑	ACC ↑	FPR ↓	FNR ↓	AP	ACC	FPR	FNR	AP	ACC	FPR	FNR	AP	ACC	FPR	FNR	AP	ACC	FPR	FNR	AP	ACC	FPR	FNR	AP	ACC	FPR	FNR
Explicit-based	SBIs*	62.7	69.2	38.5	23.1	74.5	65.5	31.1	38.1	47.4	48.2	74.3	28.2	74.6	69.9	24.8	35.4	65.2	58.3	44.0	45.2	63.0	58.2	54.1	28.8	64.6	61.6	44.4	33.1
	CADDM*	46.8	37.3	83.6	41.5	44.6	50.3	60.3	41.1	48.1	44.7	65.8	44.7	18.4	51.1	81.2	40.7	23.5	50.6	67.2	42.7	53.3	55.6	53.7	41.6	39.1	48.2	68.6	42.1
Implicit-based	Xception	44.6	50.0	100	**0**	46.6	50.0	100	**0**	48.5	50.00	100	**0**	**100**	**100**	**0**	**0**	**100**	**100**	**0**	**0**	38.0	50.0	100	**0**	62.9	66.7	66.7	**0**
	Wang2020	89.3	55.7	88.6	**0**	76.1	50.8	98.2	0.2	79.4	50.0	98.4	**0**	69.9	78.9	90.9	0.1	65.2	50.1	99.4	0.4	54.7	50.0	100	**0**	72.4	55.9	95.9	0.1
	Grag2021	99.8	85.7	28.6	**0**	97.1	61.9	76.0	**0**	95.2	55.4	89.2	**0**	**100**	99.9	0.7	**0**	80.9	79.4	41.2	**0**	53.0	50.0	100	**0**	87.7	72.0	55.9	**0**
	Ojha2023	88.9	91.6	6.6	10.2	90.3	**94.1**	**2.6**	9.2	89.3	**93.4**	3.0	10.2	88.2	92.1	6.1	9.6	87.2	92.0	3.6	12.4	81.0	**83.6**	**26.6**	6.3	87.5	**91.1**	8.1	9.7
	NTF(Ours)	**99.8**	**91.8**	**0.4**	16.0	**99.9**	92.2	3.3	15.4	**99.9**	92.0	**0.2**	15.8	99.9	87.8	**0**	15.7	99.7	91.8	0.8	15.6	**84.6**	74.3	35.6	15.8	**97.3**	88.3	**6.7**	15.7

Table 4: **Cross-family generalization on multi-step methods.** Performance of NTF and baselines in spotting 6 multi-step generated images.

Method	Wang2020	Grag2021	Ojha2023	NTF
Midjourney/Acc(%) ↑	63.02	57.88	56.82	**78.41**
Kolors/Acc(%) ↑	48.26	60.24	57.69	**78.12**

Table 5: **Evaluation on commercial generative models**.

pact of fake image features on the classifier by assigning weights to the homogeneous features. Formally, the soft contrastive loss is as follows:

$$\mathcal{L}_{scl} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \sum_{j \in J} \frac{exp[z_i(z_p + y \cdot \eta \cdot z_j^{hom})/\tau]}{\sum_{k \in K(i)} exp(z_i \cdot z_k/\tau)}, \tag{5}$$

where $P(i)$ is the positive samples set for the anchor i, $|P(i)|$ is its cardinality, $K(i)$ is the negative set and J is homogeneous set in a batch. Note that for real anchor, $y = 1$, otherwise $y = -1$. η is a balance factor. To achieve fake image detection, for a given sample and label, the classifier D is optimized on the binary cross entropy loss:

$$\mathcal{L}_{ce} = -\frac{1}{N} \sum_{i}^{N} y_i \cdot \log(\hat{y_i}) + (1 - y_i) \cdot \log(1 - \hat{y_i}), \tag{6}$$

Finally, the discriminative loss is given by:

$$\mathcal{L}_d = \mathcal{L}_{scl} + \gamma \mathcal{L}_{ce}, \tag{7}$$

where γ is a balance factor. Please refer to the technical appendix for more implementation details.

Experiments

Experiments Setup

Dataset. We use the dataset provided by (Wang et al. 2020), which consists of 720K images for training and 4K images for validation. The fake images were generated by ProGAN(Karras et al. 2017), while an equal number of real images were sourced from LSUN(Yu et al. 2015). We conduct evaluation experiments on the self-built dataset, which covered GAN-based, DM-based, and multi-step fake images.

Evaluation Metrics. In evaluating the performance in spotting fake images synthesized with diverse generative models, we adopt four popular metrics to get a comprehensive result of our proposed method. Specifically, we report ACC (accuracy), AP (average precision), FPR (false positive rate), and FNR (false negative rate), respectively.

Baselines. We compare with the following six baselines, including fake detectors based on explicit and implicit artifacts: 1) Xception (Rossler et al. 2019) is widely employed as the baseline in the studies of DeepFake forensics; 2) Wang2020(Wang et al. 2020) focuses on the artifacts exposed by CNN-generated images; 3) Grag2021(Gragnaniello et al. 2021) uses spectral super-resolution to reconstruct visual cues for detection. 4) Ojha2023(Ojha 2023) uses a feature space not explicitly trained to distinguish real from fake images. 5) SBIs(Shiohara 2022) mixes image pairs with various masks to generate training data. 6) CADDM(Dong et al. 2023) focuses on local information so that the network ignores identity information leakage caused by irregular face changes. Except that SBIs and CADDM are trained on FaceForensics++ (Rossler et al. 2019), the training set for the

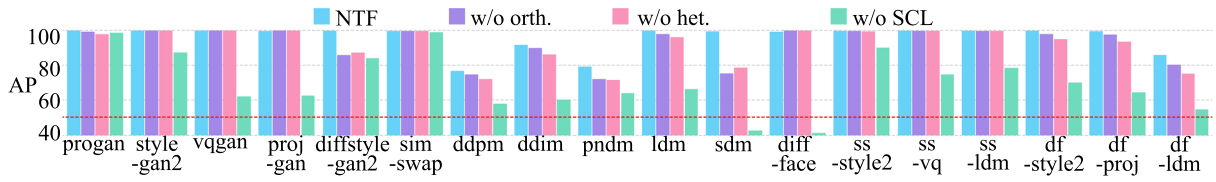

Figure 3: **Ablation study on NTF architecture.** All detectors were trained using the ProGAN and tested on other generative models. The designs of NTF architecture improve generalization ability. The red dotted line depicts chance performance.

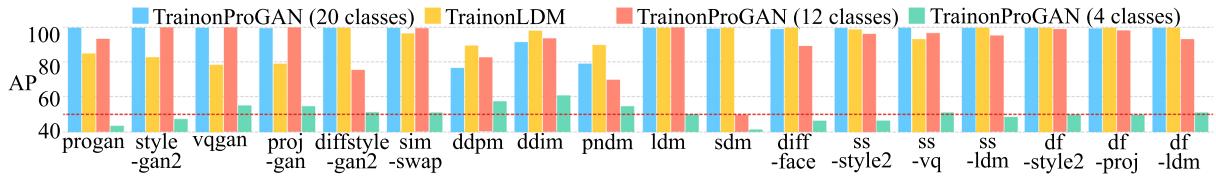

Figure 4: **Ablation study on training data.** All detectors trained on different data sources (ProGAN or LDM) or different numbers of classes of the ProGAN data source (20 classes, 12 classes, and 4 classes).

other baselines is consistent with ours. Note that SBIs and CADDM are limited to DeepFake datasets, and for the sake of fairness, they are not included in the test of non-face data.

Implementation Details. We use ResNet50 pre-trained on ImageNet as the feature encoder. All the projection heads contain two layers of MLPs with an output dimension of 128. For all datasets, we use a 224×224 crop for both training and testing (random crop for training and center crop for testing). For soft contrastive learning, we perform a random crop of the input image to 32 px. The optimizer is an SGD with a momentum of 0.9, an initial learning rate of 0.1, and an attenuation of 0.001. In the first stage, training is conducted for 200 epochs, followed by 10 epochs in the second stage. The empirical setting for λ, η, γ is set at 0.1, 0.5, 0.5.

Effectiveness Evaluation

We explore the capability of NTF to detect fake images from different generative models in the self-built dataset.

Performance on GAN-based fake images. The results in Table 2 demonstrate the capability of NTF in tackling unknown GAN-based fake images with the highest AP **99.9%** and the lowest FPR **0.2%** on average. In particular, NTF attains optimal performance on GAN-based fake images, which could be attributed to the similarity in the image generation principle of these models to that of ProGAN. That is precisely why certain baselines, such as Xception and Grag2021, have achieved high-performance generalization on GAN-based generative models. This indicates that existing fake image detections are adept at handling generalization scenarios within the same model family.

Performance on DM-based fake images. As illustrated in Table 3, NTF exhibits superior cross-family generalization capability on the DMs with the highest AP **91.2%** and ACC **82.7%** on average. Specifically, NTF improves the AP and ACC by nearly **6%** and **19.2%** compared to the best baseline. The detection accuracy of NTF surpasses all baselines across four DMs. However, it remains comparable to Grag2021 on LDM and to XeceptionNet on DDPM. Overall, NTF detects different DM-base generation models in a more balanced way, although it is slightly inferior to some baselines on certain models.

Performance on multi-step fake images. As shown in Table 4, NTF detects fake images synthesized by six multi-step methods with the highest mAP of **97.3%**. It is noteworthy that the method from Ojha2023 demonstrated an average ACC of 91.1%, slightly outperforming NTF. This can be attributed to the use of ViT-L/14(Dosovitskiy et al. 2020), a vision transformer variant pre-trained on CLIP(Radford et al. 2021) for the backbone, aiding in effectively modeling details for real and fake image classification.

Performance on commercial generative models. To better evaluate the performance in tackling commercial generative models, we assess the detection capability of NTF on fake images generated by *Midjourney* and *Kolors* in Table 5. Experimental results show that our proposed method NTF gives an accuracy more than **78.4%** which significantly outperforms the baselines.

In summary, NTF effectively detects GAN-based unknown generative models with a mAP **99.9%**. It also demonstrates the capability to detect DM-based and multi-step generation methods with mAPs of **91.2%** and **97.3%** respectively, achieving an overall mAP of **96.2%** across all datasets. Additionally, NTF shows a slightly higher FNR, which can be attributed to the inadequate coverage of real-world images in the training dataset.

Ablation Studies

We evaluated the effect of architecture and training data on the generalization ability of NTF. It initially employed homogeneous, heterogeneous, and soft orthogonality losses with Eq (4) for learning natural trace representations and it trains with ProGAN/LSUN dataset consisting of 20 classes.

Effect of network architecture. We conducted experiments with different variants of NTF, exploring the following configurations: 1) without *ort.* loss, 2) without *het.* and *orth.* loss and 3) without SCL. For each, We maintain ProGAN real/fake image data as training data. Figure 3 shows the performance of these variants on the same models. We find

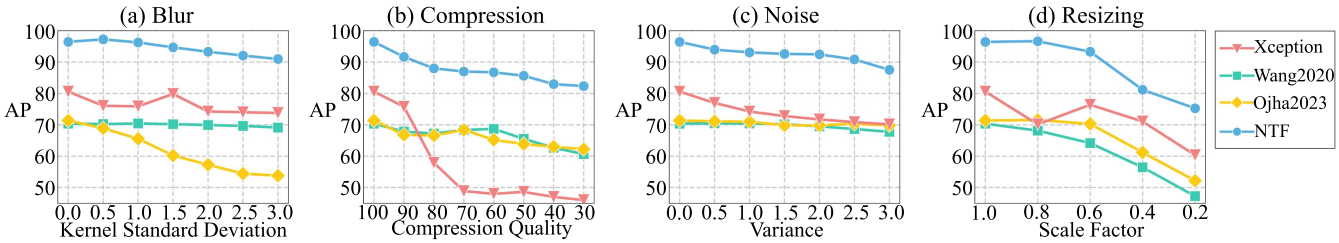

Figure 5: Robustness to four image processing operations, *i.e.*, Gaussian blur (a), JPEG compression (b), Gaussian Noise (c), Scaling (d).

Figure 6: The Grad-CAM++ visualization of the self-built dataset.

that the architectural design of NTF has an important role in generalization, improving performance on conditional generative models such as LDM, SDM, and multi-step synthesis methods. This improvement is likely attributed to the implementation of soft orthogonality and heterogeneous loss constraints on upper bounds, which enables the NTF to learn homogeneous features capable of effectively separating traces of conditional generative models from natural traces. Moreover, this design also improves performance on unconditional generative models, particularly more significant in models like Diff-StyleGAN and PNDM.

Effect of training data Next, we investigate how the source and diversity of the training data influence a detector's generalization ability. Our approach involves altering either the data source or the number of classes in the training set. Specifically, we train multiple detectors by 1) employing a pre-trained LDM, substituting the ProGAN, and 2) utilizing a subset of the full ProGAN dataset, excluding real and fake images from certain LSUN classes, as shown in Figure 4. With access only to the LDM dataset, the model displays impressive generalization capabilities, even though the dataset consists of 200k reals from LAION(Schuhmann et al. 2022) and 200k fakes generated by LDM. As expected, diversifying the training set does enhance generalization to some extent, but the benefits diminish with increasing diversity. This suggests the potential existence of a real image dataset capable of extracting universally present natural traces.

Robustness Evaluation

In this section, we mainly explore the robustness of our proposed method in surviving diverse input transformations, such as blur, compression, noise, and scaling.

As shown in Figure 5, NTF is typically robust to blur, compression, and noise operations. The performance of Wang2020 is more consistent under different levels of blur/-compression/noise. This is reasonable, as it employs random compression and blur for data augmentation during training. Although Ojha2023 also employs data augmentation, it also performs in addition to being less robust in blur operations. In addition, both NTF and other baselines show significant performance degradation in image scaling operations. When the image scaling factor is 0.2, the image content becomes imperceptible, making robustness in such extreme scenarios less critical. Please refer to the technical appendix for more details.

Qualitative Analysis

To understand how the network generalizes to different synthesis methods, we visualize the model saliency map. We apply GradCAM++ (Chattopadhay et al. 2018) to NTF on the self-built dataset to visualize where models are paying their attention to images, as shown in Figure 6. This demonstrates how the network captures artifacts from different generative models; for example, with the dog generated by StyleGAN2, the network focuses on the inconsistency in shadow.

Conclusion

In this paper, we escape the trap of exploring subtle differences between real and fake for fake image detection. Motivated by the presence of common shared features in real images, We propose a novel framework named NTF, pre-trained by natural trace representation learning and soft contrastive learning, to significantly improve the generalization ability of fake image detection. Extensive experiments on the self-built dataset demonstrate that our method exhibits state-of-the-art generalization capability for unknown generative models. Our research also offers a fresh perspective on fake image detection, focusing on exploring stable detectable features rather than those that continuously change, thereby paving the way for future studies in this field.

Acknowledgements

This research was supported in part by the National Key Research and Development Program of China under No.2021YFB3100700, the National Natural Science Foundation of China (NSFC) under Grants No. 62202340, 62372334, the CCF-NSFOCUS 'Kunpeng' Research Fund under No. CCF-NSFOCUS 2023005, the Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness under No. HNTS2022004, the Fundamental Research Funds for the Central Universities under No. 2042023kf0121.

References

Chattopadhay, A.; Sarkar, A.; Howlader, P.; and Balasubramanian, V. N. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 839–847.

Chen, R.; Chen, X.; Ni, B.; and Ge, Y. 2020. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2003–2011.

Corvi, R.; Cozzolino, D.; Zingarini, G.; Poggi, G.; Nagano, K.; and Verdoliva, L. 2023. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Dhariwal, P. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3994–4004.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Elhassan, A.; Al-Fawa'reh, M.; Jafar, M. T.; Ababneh, M.; and Jafar, S. T. 2022. DFT-MF: Enhanced deepfake detection using mouth movement and transfer learning. *SoftwareX*, 19: 101115.

Esser, P. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, 3247–3258. PMLR.

Ghayoomi, M. 2022. Deep transfer learning for COVID-19 fake news detection in Persian. *Expert Systems*, 39(8): e13008.

Gragnaniello, D.; Cozzolino, D.; Marra, F.; Poggi, G.; and Verdoliva, L. 2021. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In *2021 IEEE international conference on multimedia and expo (ICME)*, 1–6. IEEE.

Ho, J. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Hu, S. 2021. Exposing GAN-generated faces using inconsistent corneal specular highlights. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2500–2504. IEEE.

Huang, Y.; Juefei-Xu, F.; Wang, R.; Guo, Q.; Ma, L.; Xie, X.; Li, J.; Miao, W.; Liu, Y.; and Pu, G. 2020. Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction. In *Proceedings of the 28th ACM international conference on multimedia*, 1217–1226.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Kim, K.; Kim, Y.; Cho, S.; Seo, J.; Nam, J.; Lee, K.; Kim, S.; and Lee, K. 2022. Diffface: Diffusion-based face swapping with facial guidance. *arXiv preprint arXiv:2212.13344*.

Lee, J.-W.; and Kim, J.-H. 2022. Fake sentence detection based on transfer learning: applying to Korean COVID-19 fake news. *Applied Sciences*, 12(13): 6402.

Liu, C.; Zhu, T.; Shen, S.; and Zhou, W. 2023. Towards Robust GAN-generated Image Detection: a Multi-view Completion Representation. *arXiv preprint arXiv:2306.01364*.

Liu, L.; Ren, Y.; Lin, Z.; and Zhao, Z. 2022. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ojha, U. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24480–24489.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ranjan, P. 2020. Improved Generalizability of Deep-Fakes Detection using Transfer Learning Based CNN Framework. In *2020 3rd International Conference on Information and Computer Technologies (ICICT)*, 86–90.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent

diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1–11.

Sauer, A.; Chitta, K.; Müller, J.; and Geiger, A. 2021. Projected GANs Converge Faster. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Sha, Z.; Li, Z.; Yu, N.; and Zhang, Y. 2023. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 3418–3432.

Shiohara, K. 2022. Detecting Deepfakes with Self-Blended Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18720–18729.

Song, J. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Suratkar, S. 2023. Deep fake video detection using transfer learning approach. *Arabian Journal for Science and Engineering*, 48(8): 9727–9737.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; and Wei, Y. 2023a. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12105–12114.

Tan, L.; Wang, Y.; Wang, J.; Yang, L.; Chen, X.; and Guo, Y. 2023b. Deepfake video detection via facial action dependencies estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 5276–5284.

Wang, R.; Juefei-Xu, F.; Luo, M.; Liu, Y.; and Wang, L. 2021. Faketagger: Robust safeguards against deepfake dissemination via provenance tracking. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3546–3555.

Wang, R.; Juefei-Xu, F.; Ma, L.; Xie, X.; Huang, Y.; Wang, J.; and Liu, Y. 2019. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. *arXiv preprint arXiv:1909.06122*.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. CNN-generated images are surprisingly easy to spot...for now. In *CVPR*.

Wang, Z.; Zheng, H.; He, P.; Chen, W.; and Zhou, M. 2022. Diffusion-GAN: Training GANs with Diffusion. *arXiv preprint arXiv:2206.02262*.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Yu, N.; Davis, L. S.; and Fritz, M. 2019. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhang, X. 2019. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, 1–6. IEEE.

Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; and Xu, B. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, 207–212.