

Robustness in AI-Generated Detection: Enhancing Resistance to Adversarial Attacks

Haoxuan Sun

Jiahui Zhan

guwangtu@sjtu.edu.cn

Shanghai Jiao Tong University

Shanghai, China

Yan Hong

Haoxing Chen

Jun Lan

Huijia Zhu

Weiqiang Wang

Alibaba Group

Shanghai, China

Liqing Zhang

Jianfu Zhang

Shanghai Jiao Tong University

Shanghai, China

c.sis@sjtu.edu.cn

ABSTRACT

The rapid advancement of generative image technology has introduced significant security concerns, particularly in the domain of face generation detection. This paper investigates the vulnerabilities of current AI-generated face detection systems. Our study reveals that while existing detection methods often achieve high accuracy under standard conditions, they exhibit limited robustness against adversarial attacks. To address these challenges, we propose an approach that integrates adversarial training to mitigate the impact of adversarial examples. Furthermore, we utilize diffusion inversion and reconstruction to further enhance detection robustness. Experimental results demonstrate that minor adversarial perturbations can easily bypass existing detection systems, but our method significantly improves the robustness of these systems. Additionally, we provide an in-depth analysis of adversarial and benign examples, offering insights into the intrinsic characteristics of AI-generated content. All associated code will be made publicly available in a dedicated repository to facilitate further research and verification.

1 INTRODUCTION

The rapid progress of generative models [11, 20, 44], particularly diffusion models [7, 16, 28, 30, 32, 36], has significantly improved the authenticity of face generation technology. These methods have enabled the creation of highly realistic images that can often bypass detection, even by human observers. While this progress demonstrates the remarkable potential of artificial intelligence in image generation, it also introduces a major concern: the misuse of such high-fidelity synthetic images could contribute to the spread of misinformation and fake content. The ability to generate such high-quality images raises ethical issues and underscores the urgent need for robust detection systems to safeguard the integrity of digital media.

Recent strides in AIGC (Artificial Intelligence Generated Content) detection [1, 4, 38, 40] have achieved remarkable accuracy under standard conditions, with most state-of-the-art generated images classified from real images with near-perfect accuracy. However, a critical vulnerability persists: these models often fail when subjected to adversarial attacks [6, 12, 23, 24]. Our research reveals that similar to classification tasks, even minimal and imperceptible perturbations can lead to misclassifications, causing detection systems to incorrectly identify synthetic images as real and vice versa as shown in Fig. 1. Although there have been many studies

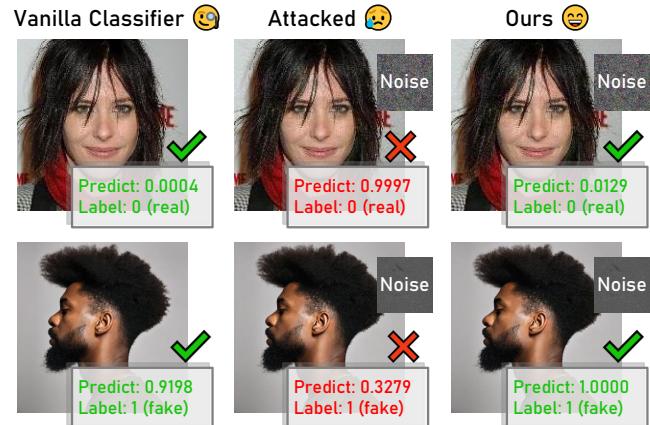


Figure 1: Although SOTA detectors achieve near-perfect accuracy in discriminating fake images, even minimal and imperceptible perturbations can cause misclassifications, leading to synthetic images being identified as real, and vice versa. In contrast, our proposed method maintains robust classification accuracy even under adversarial attacks.

on adversarial robustness in current research, there is a lack of research on the robustness of AIGC detection systems, underscoring a critical area for improvement in detection system defenses. Current research lacks advanced techniques capable of adapting to and mitigating the detrimental effects of adversarial attacks. This remains a pressing issue, as adversarial perturbations are imperceptible to humans but can almost completely disrupt detection models.

To support the practical deployment of more stable and secure detection systems in real-world scenarios, we analyze two settings of AI-generated image detection: in-domain settings and cross-domain settings. We evaluate both detection performance and robustness under adversarial attacks. Our findings indicate that adversarial training effectively mitigates the impact of adversarial perturbations. Furthermore, DIffusion Reconstruction Error (DIRE) [38], the inverse reconstruction residue of the diffusion models, further enhances robustness. As shown in Fig. 1, our new strategy significantly improves the detector's resistance. In summary, this paper makes the following contributions:

- Experiments reveal that even the most advanced AI-generated face detection models are easily deceived by minor adversarial perturbations. We are the first to focus on this critical issue in AIGC detection.
- We demonstrate that adversarial training and diffusion reconstruction error enhance the robustness of detection models when the test samples and training samples follow the same distribution.
- Through a detailed performance analysis across different datasets, we show that adversarial training alone struggles to generalize to new, unseen data distributions. Incorporating the diffusion reconstruction error approach significantly improves adversarial generalizability.

2 RELATED WORKS

AIGC (*i.e.*, Deepfakes) refers to images, videos, or audio that are edited or generated using artificial intelligence. AI-generated content detection aims to differentiate deepfakes from natural images, videos, or audio. Early detection methods focused on identifying artifacts and inconsistencies caused by the limitations of GANs [11, 19, 44], such as anomalies in lighting [9, 10], shadows [3], and reflections [3]. However, as models such as diffusion models [7, 16, 25, 30, 32], DiT [28] and VAR [36] improved realism and reduced artifacts, these techniques became less effective.

Recent research emphasizes unique traces left by image generation processes. Deep Image Fingerprint (DIF) [33] uses convolutional neural networks to extract unique fingerprints for identifying images from specific models. DIRE [38] and SeDID [22] utilize reverse diffusion to detect subtle differences between real and synthetic images. DIRE focuses on reconstruction accuracy at the initial timestep, while SeDID leverages errors from intermediate diffusion steps for richer analysis. Diffusion Reconstruction Contrastive Learning (DRCT) [4] generates challenging samples through high-quality diffusion reconstruction and employs contrastive training to improve the detector’s generalizability. Addressing dataset imbalances, Xu *et al.* [41] highlighted how attribute inconsistencies negatively affect detection results and proposed the creation of annotated datasets to remedy this issue. Furthermore, the Contrastive Deepfake Embeddings (CoDE) [1] framework introduces a new embedding space trained through contrastive learning that emphasizes global-local similarities to enhance detection performance. To combat generalization issues stemming from overfitting to specific artifacts, Latent Space Data Augmentation (LSDA) [42] expands forgery representation diversity, thereby enabling models to adopt more flexible decision boundaries. Moreover, Tan *et al.* [35] focus on generator architectures, specifically rethinking CNN-based structures to show how upsampling operators can produce generalized forgery artifacts and introducing Neighboring Pixel Relationships (NPR) for effective characterization of these artifacts, leading to notable performance gains across various datasets. Additionally, while some studies have concentrated on video deepfake detection [5, 27, 43], this paper focuses on still image detection and will not elaborate on those contributions. Incorporating multimodal information has also shown promise. Lasted [40] uses language-guided contrastive learning with text labels like “real/synthetic

Algorithm 1 Adversarial Training with DIRE

```

1: Input: Detect Model  $f$ , dataloader  $D$ , number of epochs  $E$ , step size  $\alpha$ , perturbation bound  $\epsilon$ , number of PGD iterations  $T$ .
2: for epoch = 1 to  $E$  do
3:   for  $(x, y)$  in  $D$  do
4:      $x' \leftarrow x + \delta_0$ 
5:     for  $t = 1$  to  $T$  do
6:        $x' \leftarrow x' + \alpha \cdot \text{sign}(\nabla_{x'} \mathcal{L}(f(x'), y))$ 
7:        $x' \leftarrow B(x', x - \epsilon, x + \epsilon)$ 
8:     end for
9:      $x_0 \leftarrow \text{DDIM Reconstruction}(\text{DDIM Inversion}(x))$ 
10:     $x'_0 \leftarrow \text{DDIM Reconstruction}(\text{DDIM Inversion}(x'))$ 
11:     $DIRE(x_0) \leftarrow |x_0 - x|$ 
12:     $DIRE(x'_0) \leftarrow |x'_0 - x'|$ 
13:     $x_{\text{combined}} \leftarrow \text{Concat}(DIRE(x_0), DIRE(x'_0))$ 
14:     $y_{\text{combined}} \leftarrow \text{Concat}(y, y)$ 
15:     $y_{\text{predicted}} \leftarrow f(x_{\text{combined}})$ 
16:     $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}(y_{\text{predicted}}, y_{\text{combined}})$ 
17:     $f \leftarrow \text{Update}(\mathcal{L}_{\text{total}}, f)$ 
18:  end for
19: end for
20: Output: Trained model  $f$ 

```

photo” and “real/synthetic painting” to uncover forensic features and improve detection across diverse generative methods.

Despite these advancements, existing detection methods remain vulnerable to adversarial attacks. Although adversarial perturbations are often imperceptible to humans, they can significantly degrade the performance of detection models. Techniques like Fast Gradient Sign Method (FGSM) [12], Fast Gradient Method (FGM) [24], and Projected Gradient Descent (PGD) [23] generate adversarial examples by altering inputs based on loss gradients. Among these, PGD, with its iterative updates and constraint projections, is particularly effective at producing adversarial samples. In the field of facial recognition, there have been many studies on improving model robustness [13, 14, 31], but there is still a gap in research on AIGC detection. Furthermore, some works utilize generative models to defend against attacks [26, 39]. However, for the task of distinguishing whether an image is generated by AI, we should avoid using generative models in the judgment process to prevent artifacts from the generative models from interfering with the assessment.

Our study shows that even subtle perturbations can severely impact the performance of current frameworks, emphasizing the need for robust detection systems that can withstand adversarial conditions and reliably identify AI-generated images.

3 METHODOLOGIES

Given an input image x , we define $f(\cdot)$ as the detection model. The function $f(x)$ outputs a binary classification label, determining whether the input image is real or synthetic. Each image is assigned a label $y \in \{0, 1\}$ where 1 indicates a synthetic image and 0 indicates a real image. The goal of $f(\cdot)$ is to map $f(x)$ to the corresponding y . Recent studies [1, 4] have demonstrated that simple classification models achieve near-perfect accuracy under standard conditions,

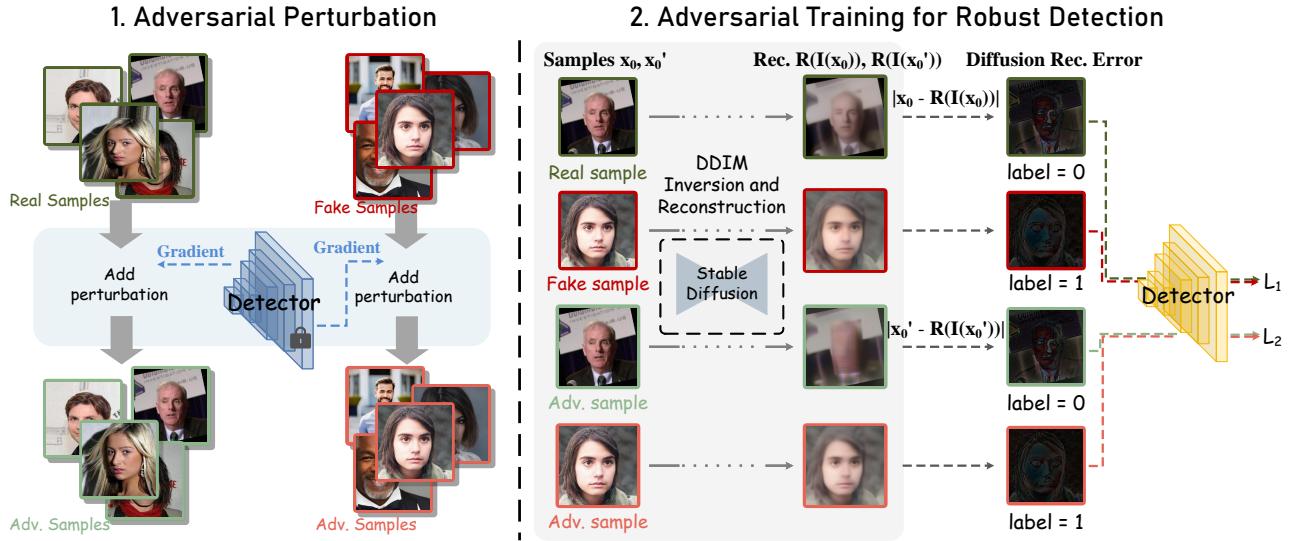


Figure 2: Pipeline of our robustness-centric method, which integrates adversarial training with diffusion inversion reconstruction to improve detection robustness. Given a set of real and fake images, we first apply adversarial perturbations and then use the resulting adversarial samples along with the original samples for residual reconstruction. Finally, we train the detector using the reconstruction residual maps for more robust performance.

successfully distinguishing most of the state-of-the-art generated images from real ones. However, we will show that these models remain highly susceptible to adversarial perturbations. Our complete pipeline for enhancing the robustness of the detection model is illustrated in Fig.2 and detailed in Algorithm1.

3.1 Adversarial Perturbation

This section outlines the process of generating adversarial examples against AI-generated image detection models using the Projected Gradient Descent (PGD) [23] method, a widely adopted approach to craft adversarial perturbations capable of deceiving machine learning models. The PGD method begins with a clean input image, denoted as x . The objective is to generate an adversarial example x' by adding a small perturbation δ to x . This perturbation is designed to be imperceptible to the human eye while effectively misleading the model into making incorrect predictions.

Initialization. The process starts by initializing the perturbation δ , where δ is uniformly distributed within the ϵ neighborhood, where ϵ defines the maximum allowable perturbation, resulting in the first adversarial example as $x' = x + \delta_0$.

Gradient Calculation. The next step involves calculating the gradient of the loss function with respect to the input image. This is performed by conducting a forward pass through the model to obtain the predicted output $f(x)$ and then computing the gradient:

$$g = \Delta_x \mathcal{L}(f(x), y), \quad (1)$$

where \mathcal{L} is the loss function, $f(x)$ is the output of model given x as input, and y is the true label.

Update the Perturbation. The perturbation is then updated in the direction of the gradient to maximize the loss:

$$\delta_{i+1} = \delta_i + \alpha \cdot \text{sign}(g), \quad (2)$$

where α is a small step size that controls the magnitude of the perturbation.

Bound. To ensure that the adversarial perturbation remains imperceptible, a bound is applied to constrain its magnitude. Specifically, the perturbation is restricted within a predefined norm L_p and is projected as:

$$\delta = B(\delta, -\epsilon, \epsilon), \quad (3)$$

Iteration. The steps *gradient calculation* through the *bound* are repeated for a predefined number of iterations, refining the perturbation each time to produce the final adversarial example $x' = x + \delta$.

Finally, the adversarial example x'_n after n iterations can be expressed as:

$$x'_0 = x + \delta_0, x'_n = B_{x', \epsilon}(x'_{n-1} + \alpha \cdot \text{sign}(\Delta_x \mathcal{L}(x'_{n-1}, y))), \quad (4)$$

The effectiveness of these adversarial perturbations lies in their ability to exploit the inherent vulnerabilities of machine learning models. Despite being nearly invisible to the human eye, these perturbations can induce significant changes in model predictions. For example, a facial recognition system might incorrectly classify the adversarial example as a different individual or fail to detect a face entirely. Similarly, adding such perturbations to real images results in their misclassification as AI-generated images, while perturbing generated images leads to their misclassification as real images. Due to the intrinsic characteristics of the generated images, this type of noise is even more difficult to detect, as it is challenging to discern whether the noise originates from the adversarial attack or from the generative model itself. This phenomenon highlights the

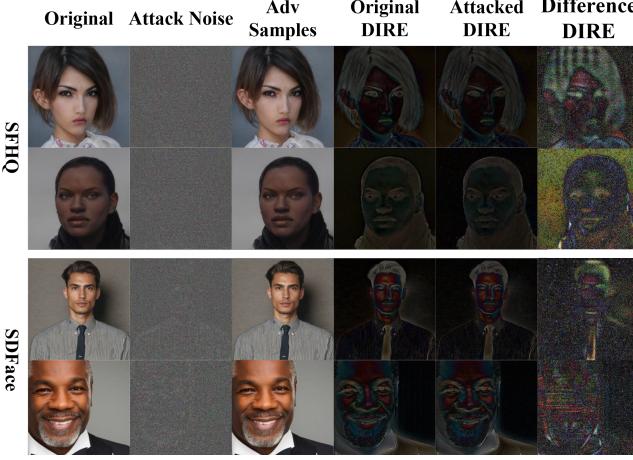


Figure 3: Visualization of attack noise, adversarial samples, and DIRE residual maps of fake images. “Difference DIRE” represents the disparity between DIRE maps of the original samples and those of the adversarially attacked samples. For enhanced clarity, the attack noise was amplified by a factor of 20, and the Difference DIRE maps by a factor of 10.

critical security risks of deploying deep learning models in sensitive applications, where even minor, undetectable modifications can severely compromise their accuracy and reliability.

3.2 Adversarial Training

One effective approach to defending against adversarial perturbations is adversarial training. Adversarial training [23] is a widely used defense mechanism designed to improve the robustness of machine learning models against adversarial attacks. By incorporating adversarial examples into the training process, the model learns to adapt to perturbations that would otherwise lead to misclassifications, thereby enhancing its resistance to adversarial manipulation.

The training process uses a combined loss function that accounts for both the standard loss in clean examples and the loss in adversarial examples. This can be formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_1(f(x), y) + \lambda \cdot \mathcal{L}_2(f(x'), y), \quad (5)$$

where, \mathcal{L}_1 and \mathcal{L}_2 are cross-entropy loss, $f(x)$ represents the model’s prediction, λ is a hyperparameter balancing the contributions of clean and adversarial losses, x' is the adversarial example generated using the method described earlier.

The training process uses mixed data, each batch containing clean images and their corresponding adversarial examples. This ensures that the model learns not only the standard features of the data but also the features susceptible to adversarial exploitation. By exposing the model to these challenging examples, its decision boundaries become more robust against perturbations. Incorporating adversarial examples into the training pipeline significantly enhances the robustness of the model. Our subsequent experiments also confirm that adversarial training effectively improves robustness in AI-generated content detection. By learning

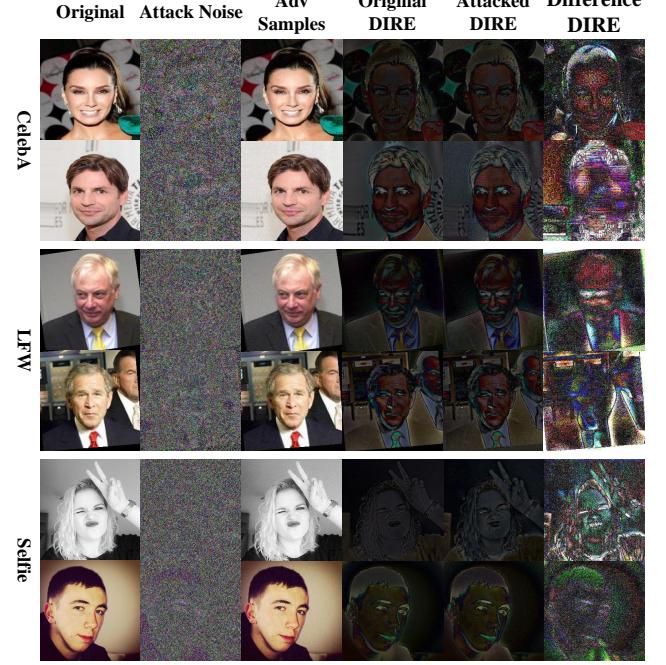


Figure 4: Visualization of attack noise, adversarial samples, and DIRE maps of real images.

from adversarial perturbations, the model develops a deeper understanding of the input space, enabling it to maintain accurate predictions even under adversarial manipulations.

3.3 Diffusion Reconstruction for Robustness Enhancement

Although adversarial training significantly mitigates the impact of adversarial attacks, we found that when training data is insufficient, detection models remain vulnerable to attacks on data distributions different from the training set, even when adversarial examples are incorporated. To address this limitation, we leverage the robustness of the diffusion reconstruction against white-box attacks and propose enhancements to further improve its resilience.

In diffusion models [16], the forward process transforms sample x_0 to noise latent x_T by progressively adding Gaussian noise, while the reverse process denoises x_T back to x_0 . T is the number of steps. Diffusion Reconstruction Error (DIRE) [38] employs the DDIM [34] inversion process to gradually add noise to x_0 , mapping it into a noise latent space. The reverse process of diffusion models in DDIM is defined as:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1 - \alpha_t} \epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(x_t, t) + \sigma_t \epsilon_t, \quad (6)$$

where $\epsilon_\theta(x_t, t)$ represents the noise calculated by the noise prediction model given x_t and t , and its model parameter is θ . σ_t represents the standard deviation parameter in the time step t , which decreases as T increases. If $\sigma_t = 0$ (T is large enough) the

Table 1: Performance obtained through training on all five datasets using different methods.

Method	Dataset	Clean Images		Adversarial Images		Robustness Score	
		w/o AT	w/ AT	w/o AT	w/ AT	w/o AT	w/ AT
ResNet	CelebA	99.99%	99.99%	0.00%	99.02%	0.00	0.99
	LFW	99.97%	99.92%	0.00%	89.02%	0.00	0.89
	Selfie	99.99%	99.72%	0.00%	97.18%	0.00	0.97
	SFHQ	99.99%	99.99%	0.00%	99.98%	0.00	0.99
	SDFace	99.98%	82.53%	0.00%	70.77%	0.00	0.85
ViT	CelebA	99.99%	99.96%	0.00%	99.84%	0.00	0.99
	LFW	99.99%	99.50%	0.00%	99.01%	0.00	0.99
	Selfie	99.99%	98.89%	0.00%	97.77%	0.00	0.99
	SFHQ	99.99%	99.98%	0.00%	99.98%	0.00	1.00
	SDFace	99.85%	85.02%	0.00%	75.78%	0.00	0.89
DIRE	CelebA	100.0%	100.0%	81.24%	99.99%	0.81	1.00
	LFW	100.0%	100.0%	97.66%	100.0%	0.97	1.00
	Selfie	99.89%	99.77%	98.07%	99.91%	0.98	1.00
	SFHQ	100.0%	100.0%	99.73%	99.99%	0.99	1.00
	SDFace	99.89%	99.89%	38.94%	99.89%	0.39	1.00

reverse process becomes deterministic (*reconstruction process*), in which one noise latent x_T determines one generated sample x_0 .

The DDIM inversion process deterministically maps x_0 to x_T , which can be treated as the reversion of the reconstruction process in Eq. 6:

$$\frac{x_{t+1}}{\sqrt{\alpha_{t+1}}} = \frac{x_t}{\sqrt{\alpha_t}} + \left(\sqrt{\frac{1 - \alpha_{t+1}}{\alpha_{t+1}}} - \sqrt{\frac{1 - \alpha_t}{\alpha_t}} \right) \epsilon_\theta(x_t, t), \quad (7)$$

After T steps, x_0 becomes a point x_T in the isotropic Gaussian noise distribution. The inversion process identifies the corresponding point in the noisy space, and the reconstruction process is then used to reconstruct the input image, producing a reconstructed version x'_0 .

The differences between x_0 and x'_0 help to distinguish real or generated. The DIRE map is then defined as:

$$DIRE(x_0) = |x_0 - R(I(x_0))|, \quad (8)$$

where $|\cdot|$ denotes the computation of the absolute value, and $I(\cdot)$ is a series of the inversion process. $R(\cdot)$ is a series of the reconstruction process. Images generated by diffusion models are sampled from the distribution of the diffusion generation space, whereas real images originate from a distinct distribution. As a result, samples from the diffusion generation space are more likely to be faithfully reconstructed by a pre-trained diffusion model, while real images are not. Therefore, DIRE naturally benefits AI-generated content detection. Additionally, the multi-step inversion and reconstruction process can mitigate the impact of adversarial perturbations. Experimental results demonstrate that DIRE exhibits a certain level of resistance to adversarial attacks, even when directly confronted with them. More importantly, while adversarial training struggles to handle adversarial examples from different distributions (cross-dataset scenarios) when training data is limited, incorporating DIRE significantly enhances adversarial robustness in such cases. We will

further validate and analyze this improvement through subsequent experiments.

4 EXPERIMENTS

4.1 Datasets

In this study, we employ two well-established classification models, ResNet50 [15] and Vision Transformer (ViT) [8], to detect AI-generated images. Each model is independently utilized in our framework to support detection tasks in facial analysis. For training and evaluation, we leverage multiple datasets that encompass diverse facial characteristics and scenarios. The real face datasets used include CelebA [21], LFW [17], and Selfie [18]. For AI-generated faces, we select SDFace [37] and SFHQ [2]. Please refer to *Appendix* for details of these datasets.

For images in different datasets, the clarity may vary; therefore, the images were converted to JPEG format with a quality setting of 95, and during pre-processing for the model, they were consistently cropped to a size of 224x224 pixels.

4.2 Implementation Details

We deployed our method on an NVIDIA A100 GPU. The ResNet50 and ViT-B/16 are pre-trained on ImageNet and were employed as our backbone detectors. For each setting, Adam optimizer is used with a learning rate of 5e-6 and $\beta_1 = 0.9$, $\beta_2 = 0.999$. During training, we conducted 10 epochs per process with a batch size of 128. The trade-off parameter $\lambda = 1$ of Eqn. 5.

The attack method used in this study is Projected Gradient Descent (PGD), applied under the L_∞ norm constraint with a step size of 10. Training with the ResNet50 backbone took approximately 3.3 hours without adversarial training and 15 hours with adversarial training, while a single evaluation required about 9.5 ms. Training with DIRE required approximately 9 hours to extract DIRE images, followed by 3.5 hours to train the classifier. When applying

Table 2: Performance obtained through training on LFW and SFHQ using different methods.

Method	Dataset	Clean Images		Adversarial Images		Robustness Score	
		w/o AT	w/ AT	w/o AT	w/ AT	w/o AT	w/ AT
ResNet	CelebA	94.35%	1.58%	0.00%	0.55%	0.00	0.35
	LFW	100.0%	92.67%	0.00%	64.07%	0.00	0.69
	Selfie	41.36%	11.05%	0.00%	6.37%	0.00	0.57
	SFHQ	100.0%	100.0%	0.00%	99.73%	0.00	1.00
	SDFace	86.89%	99.39%	0.00%	64.50%	0.00	0.65
ViT	CelebA	79.76%	23.61%	0.00%	2.24%	0.00	0.09
	LFW	99.89%	98.34%	0.00%	90.37%	0.00	0.92
	Selfie	10.11%	49.85%	0.00%	12.77%	0.00	0.25
	SFHQ	100.0%	100.0%	0.00%	99.78%	0.00	1.00
	SDFace	97.61%	53.83%	0.00%	9.33%	0.00	0.17
DIRE	CelebA	91.82%	72.05%	0.83%	28.00%	0.01	0.39
	LFW	99.96%	99.96%	65.51%	99.96%	0.65	1.00
	Selfie	26.11%	25.60%	7.28%	40.47%	0.28	1.58
	SFHQ	100.0%	100.0%	100.0%	100.0%	1.00	1.00
	SDFace	92.83%	81.28%	78.72%	41.06%	0.85	0.50

adversarial training with DIRE, the total computational cost was approximately doubled.

4.3 Detection Performance Metrics

Since testing on a single dataset may only produce binary outcomes (true or false), Precision and Recall metrics may lack significance. Therefore, we do not report Precision and Recall. Furthermore, to systematically evaluate the resistance of models, we propose a robustness score, as follows:

$$\text{Robustness Score} = \frac{\text{Acc}_{\text{adv}}}{\text{Acc}_{\text{clean}}}, \quad (9)$$

where Acc_{adv} and $\text{Acc}_{\text{clean}}$ denotes accuracy under adversarial (attacked) and clean (unattacked) contexts respectively. We will report both Accuracy and Robustness Score in the following subsections.

We adopt two evaluation settings in our experiments:

- **All-set:** In this setting, detectors are trained on the combined datasets and tested on their respective test sets. This approach evaluates the overall performance of the detectors across all datasets. The results are in Table 1;
- **Cross-domain:** In this setting, detectors are trained on a selected subset of datasets (one real and one fake dataset) and tested on the remaining test sets from the other datasets. This approach assesses the model’s generalization ability to unseen data distributions. The results of training on LFW and SFHQ while testing on other datasets are shown in Table 2.

In these tables, we report results for both clean images and images with adversarial perturbations. Additionally, we compare models with adversarial training (“w/ AT”) and without adversarial training (“w/o AT”). For the rest of the combinations, we will report in the *Appendix*.

4.4 All-Set Detection Performance

4.4.1 Impact of Adversarial Perturbations. As shown in Table 1, comparing “Clean Images / w/o AT” and “Adversarial Images / w/o AT” highlights the impact of adversarial perturbations. Under the standard setup without adversarial perturbations or adversarial training, test accuracy for both ResNet and ViT architectures approaches 100%, demonstrating near-perfect performance. This holds true for both the all-set and cross-dataset settings, indicating minimal challenges under ideal conditions. However, when subtle white-box adversarial perturbations are introduced, test accuracy and robustness scores drop sharply to near 0, rendering the detector ineffective. Even DIRE, a method explicitly designed for detecting fake images generated by diffusion models, experiences significant performance degradation across all subsets. That said, DIRE shows a certain level of resilience on specific datasets, such as SDFace, Selfie and CelebA, where its performance against adversarial attacks is comparatively better. These results illustrate the high vulnerability of detection models to adversarial attack techniques while highlighting DIRE’s potential to mitigate such vulnerabilities under certain conditions.

4.4.2 Effectiveness of Adversarial Training. As shown in Tables 1, comparing “Adversarial Images / w/o AT” and “Adversarial Images / w/ AT” highlights the significant improvement in accuracy achieved through adversarial training. This improvement is attributed to the inclusion of hard samples during training, which helps the detector further reduce the error space beyond the standard setup, particularly under adversarial attack conditions.

When comparing “Clean Images / w/ AT” and “Adversarial Images / w/ AT”, the performance remains comparable, demonstrating that adversarial training effectively preserves accuracy even under attack conditions. Experimental results highlight the importance of tailored adversarial training in enhancing the robustness of the detector, as reflected in the improved robustness scores.

Table 3: Performance obtained through training on CelebA and SDFace using different methods.

Method	Dataset	Clean Images		Adversarial Images		Robustness Score	
		w/o AT	w/ AT	w/o AT	w/ AT	w/o AT	w/ AT
ResNet	CelebA	100.0%	99.98%	0.00%	99.56%	0.00	0.99
	LFW	100.0%	95.96%	0.00%	21.23%	0.00	0.22
	Selfie	28.77%	92.03%	0.00%	48.13%	0.00	0.52
	SFHQ	100.0%	1.50%	0.00%	0.00%	0.00	0.00
	SDFace	100.0%	89.17%	0.00%	47.72%	0.00	0.53
ViT	CelebA	100.0%	100.0%	0.00%	99.88%	0.00	1.00
	LFW	100.0%	91.80%	0.00%	34.95%	0.00	0.38
	Selfie	23.84%	67.74%	0.00%	26.65%	0.00	0.39
	SFHQ	98.44%	28.35%	0.00%	2.13%	0.00	0.07
	SDFace	100.0%	97.33%	0.00%	85.44%	0.00	0.87
DIRE	CelebA	100.0%	100.0%	58.93%	100.0%	0.59	1.00
	LFW	98.07%	97.81%	64.11%	90.03%	0.65	0.92
	Selfie	19.27%	23.85%	8.76%	27.58%	0.45	1.15
	SFHQ	97.52%	85.20%	95.45%	66.42%	0.98	0.78
	SDFace	100.0%	99.80%	99.56%	99.78%	0.99	1.00

4.4.3 Effectiveness of DIRE. We evaluate the performance of DIRE under both standard and perturbed conditions. A comparison between DIRE, ResNet and ViT in the “Adversarial Images / w/o AT” columns highlights a significant improvement in accuracy achieved through DIRE. This enhancement can be attributed to DIRE’s ability to mitigate adversarial noise using its diffusion inversion and reconstruction process. As shown in the “Robustness Score” column of Table 1, compared to ResNet and ViT, which directly operate in pixel space, DIRE exhibits a certain level of robustness against white-box attacks. This robustness stems from the inherent denoising effect of DIRE during image reconstruction using DDIM, which helps suppress adversarial noise and prevents a drastic decline in accuracy. However, this denoising effect is not absolute, as subtle perturbations can still degrade the performance of the detector.

4.4.4 Visualization. To analyze the anti-attack properties of DIRE, we visualized the DIRE maps of real and synthetic images before and after adversarial attacks, as shown in Fig. 3 and Fig. 4. For clarity, we computed the absolute point-wise difference between the DIRE maps before and after the attacks, amplified it by a factor of 10, and presented it in the last column.

DIRE maps of synthetic images appear lighter than those of real images, which aligns with expectations due to the DDIM inversion/reconstruction process. From the figures, we observe that the adversarial noise in synthetic images is smoother than in real images, but this effect is imperceptible in the original images. Notably, compared to synthetic images, the difference maps of real images exhibit sharper, white-highlighted regions, primarily concentrated at the edges. This observation indicates that the discrepancy between the DIRE maps of real images after an attack and their original DIRE maps is more provoked than that of synthetic images. In other words, DIRE amplifies the differences between real and synthetic images under adversarial noise, revealing patterns that remain indistinguishable in the original image space.

With this observation, along with the results presented in the “DIRE” rows under the “w/ AT” column of Table 1, we confirm that this configuration achieves optimal performance under white-box attack conditions. The results highlight two key points: first, training detection models in the DIRE space enhances the adversarial robustness; second, training based on DIRE, when combined with adversarial training, enables more precise discrimination by leveraging the amplified differences between real and synthetic images.

4.5 Cross-Domain Generalization Performance

Rather than restricting training and testing to the same distribution, this section evaluates the generalization capability of our method across different domains. Specifically, we selected training sets from one real and one synthetic dataset and then tested on all five datasets. The evaluation results for training on LFW and SFHQ are presented in Table 2, training on CelebA and SDFace in Table 3. The results for other randomly selected training sets are provided in the *Appendix*. The performance of nearly all configurations declines significantly under cross-dataset testing scenarios, and adversarial attacks further exacerbate this issue, reducing the detector’s accuracy to 0%.

Incorporating adversarial training enhances cross-domain robustness. For instance, as shown in Table 2, even when SDFace is not included in the training set, the post-attack accuracy improves from 0% to 64.5%. This improvement can be attributed to the detector learning not only the inherent characteristics of real and synthetic images but also generalized representations of adversarial noise to new datasets. Furthermore, leveraging DIRE further stabilizes the detector. A comparison between DIRE, ResNet and ViT in columns “Adversarial Images / w/o AT” highlights a significant improvement in accuracy achieved through DIRE, regardless of whether the dataset is trained or not. DIRE, with pre-trained

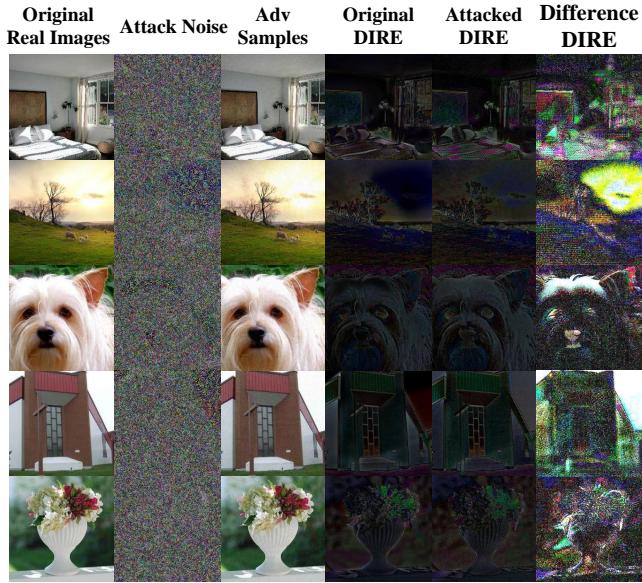


Figure 5: Visualization of attack noise, adversarial samples, and DIRE maps of non-face real images.

Table 4: Performance on non-face datasets.

Method	Dataset	Clean Images		Adversarial Images	
		w/o AT	w/ AT	w/o AT	w/ AT
ResNet	DF-imagenet	99.28%	67.62%	0.00%	67.62%
	Artifact	100.0%	95.96%	0.00%	21.23%
DIRE	DF-imagenet	99.30%	98.65%	74.86%	98.05%
	Artifact	56.61%	57.04%	54.23%	99.40%

diffusion models, maintains robustness under such conditions, allowing our method to achieve the highest resistance across most test sets.

A similar trend is observed in Table 3. Undoubtedly, the attack noise triggers the classifier’s complete failure across all datasets for ResNet and ViT. However, incorporating adversarial samples into training significantly improves performance in both seen and unseen domains, although it does not fully restore accuracy to pre-attack levels. Further integrating the diffusion reconstruction error further improves the accuracy, recovering performance to near pre-attack levels. It is noteworthy that the accuracy of Selfie remains relatively low in both the standard setup (28.77%) and the final setup with adversarial training and reconstruction error (27.58%). This may be due to the heavy stylization and post-processing applied to Selfie images, which likely introduced misclassifications.

Across all adversarial image results in Table 2 and Table 3, we observe that DIRE improves model performance in the “w/o AT” setting, demonstrating its ability to mitigate some adversarial perturbations. Additionally, models in the “w/ AT” setting always outperform those in the “w/o AT” setting, further validating the effectiveness of adversarial training in improving robustness.

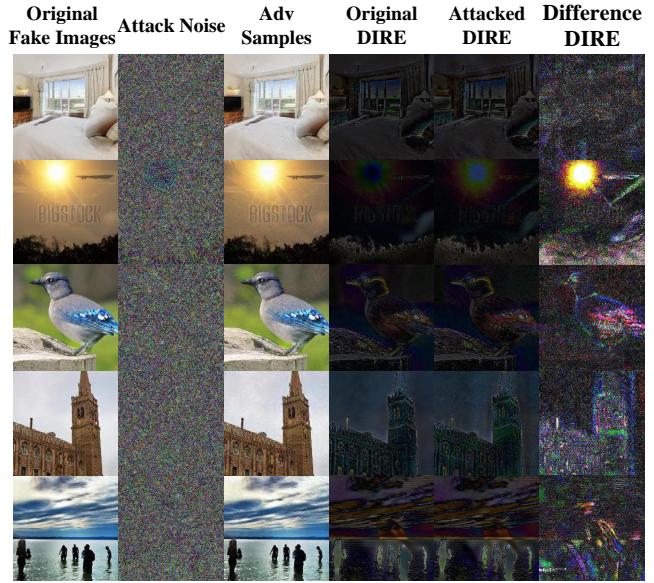


Figure 6: Visualization of attack noise, adversarial samples, and DIRE maps of non-face fake images.

5 DISCUSSIONS

In this paper, we primarily focus on face images. However, we also conduct experiments on natural images by employing Diffusion Forensics [38] (DF-Imagenet) and the Artifact [29]. We observed that adversarial training caused the detection model to fail when tested on fake images. This failure manifested as fixed output predictions (either always true or always false), as shown in Table 4. To further analyze this issue, we visualize adversarial noise, DIRE, and their differences in Figs. 5 and 6. Unlike facial images, adversarial noise in non-face images is less smooth. Additionally, the DIRE differences for both real and fake samples are sharp, diminishing the amplification effect of DIRE. Our analysis suggests that this issue arises due to fundamental differences in image patterns and complexity. In face datasets, the relatively fixed patterns make it easier to identify inconsistencies in AI-generated images. However, ImageNet encompasses a diverse range of images with complex backgrounds, varied object types, lighting conditions, and pose variations, making detection models more susceptible to these factors. When adversarial samples are introduced during training on such diverse datasets, the model struggles to learn sufficient universal features, leading to poor generalization and degraded performance.

6 CONCLUSIONS

Our research demonstrates that current generative face image detection models are highly susceptible to malicious perturbations that are imperceptible to humans yet significantly degrade model accuracy. By incorporating adversarial training, we can substantially enhance the model’s resilience to adversarial challenges. Diffusion reconstruction further provides a promising approach to improving robustness. We evaluate our method in both in-domain and cross-domain settings to comprehensively assess its effectiveness in enhancing detection robustness.

REFERENCES

- [1] Lorenzo Baraldi, Federico Cocchi, Marcella Cornia, Alessandro Nicolosi, and Rita Cucchiara. 2025. Contrasting deepfakes diffusion via contrastive learning and global-local similarities. In *European Conference on Computer Vision*. Springer, 199–216.
- [2] David Benigague. 2022. Synthetic Faces High Quality (SFHQ) dataset. doi:10.34740/kaggle/dsv/4737549
- [3] Ali Borji. 2023. Qualitative failures of image generation models and their application in detecting deepfakes. *Image and Vision Computing* 137 (2023), 104771.
- [4] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. 2024. DRCT: Diffusion Reconstruction Contrastive Training towards Universal Detection of Diffusion Generated Images. In *Forty-first International Conference on Machine Learning*.
- [5] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. 2024. Exploiting Style Latent Flows for Generalizing Deepfake Video Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1133–1143.
- [6] Francesco Croce and Matthias Hein. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*. PMLR, 2206–2216.
- [7] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.
- [8] Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [9] Hany Farid. 2022. Lighting (in) consistency of paint by text. *arXiv preprint arXiv:2207.13744* (2022).
- [10] Hany Farid. 2022. Perspective (in) consistency of paint by text. *arXiv preprint arXiv:2206.14617* (2022).
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).
- [12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).
- [13] Gaurav Goswami, Akshay Agarwal, Nalini Ratha, Richa Singh, and Mayank Vatsa. 2019. Detecting and mitigating adversarial perturbations for robust face recognition. *International Journal of Computer Vision* 127 (2019), 719–742.
- [14] Gaurav Goswami, Nalini Ratha, Akshay Agarwal, Richa Singh, and Mayank Vatsa. 2018. Unravelling robustness of deep learning based face recognition against adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- [17] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.
- [18] Mahdi M Kalayeh, Misrak Seifu, Wesna LaLanne, and Mubarak Shah. 2015. How to take a good selfie?. In *Proceedings of the 23rd ACM international conference on Multimedia*. 923–926.
- [19] Tero Karras. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv preprint arXiv:1812.04948* (2019).
- [20] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.
- [21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.
- [22] Ruipeng Ma, Jinhao Duan, Fei Kong, Xiaoshuang Shi, and Kaidi Xu. 2023. Exposing the fake: Effective diffusion-generated images detection. *arXiv preprint arXiv:2307.06272* (2023).
- [23] Aleksander Madry. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).
- [24] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725* (2016).
- [25] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*. PMLR, 8162–8171.
- [26] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. 2022. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460* (2022).
- [27] Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. 2024. AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27102–27112.
- [28] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4195–4205.
- [29] Md Awsafur Rahman, Bishmoy Paul, Najibul Haque Sarker, Zaber Ibn Abdul Hakim, and Shaikh Anowarul Fattah. 2023. Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection. In *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2200–2204.
- [30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.
- [31] Min Ren, Yuhan Zhu, Yunlong Wang, and Zhenan Sun. 2022. Perturbation inactivation based adversarial defense for face recognition. *IEEE Transactions on Information Forensics and Security* 17 (2022), 2947–2962.
- [32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- [33] Sergey Sintistsa and Ohad Fried. 2024. Deep Image Fingerprint: Towards Low Budget Synthetic Image Detection and Model Lineage Analysis. *arXiv:2303.10762* [cs.CV] <https://arxiv.org/abs/2303.10762>
- [34] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).
- [35] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. 2024. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 28130–28139.
- [36] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905* (2024).
- [37] tobeweb. 2023. stable-diffusion-face-dataset. <https://github.com/tobeweb/stable-diffusion-face-dataset>.
- [38] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22445–22455.
- [39] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. 2023. Better diffusion models further improve adversarial training. In *International conference on machine learning*. PMLR, 36246–36263.
- [40] Haiwei Wu, Jiantao Zhou, and Shile Zhang. 2023. Generalizable synthetic image detection via language-guided contrastive learning. *arXiv preprint arXiv:2305.13800* (2023).
- [41] Ying Xu, Philipp Terhöst, Marius Pedersen, and Kiran Raja. 2024. Analyzing Fairness in Deepfake Detection With Massively Annotated Databases. *IEEE Transactions on Technology and Society* (2024).
- [42] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. 2024. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8984–8994.
- [43] Daichi Zhang, Zihao Xiao, Shikun Li, Fanzhao Lin, Jianmin Li, and Shiming Ge. 2025. Learning natural consistency representation for face forgery video detection. In *European Conference on Computer Vision*. Springer, 407–424.
- [44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.