# Unsupervised Frequency-Spatial Fusion for Detecting Adversarial Examples

Canran Shen[1], Chenyi Huang[1], Wenbo Fang[2], Keying Li[1], Wengang Ma[1], Xiaolong Lan[1], Junjiang He[1*]
[1] School of Cyber Science and Engineering, Sichuan University, Chengdu 610065, China
[2] College of Computer Science and Artificial Intelligence, Southwest Minzu University, Chengdu 610225, China
Email: shencanran@stu.scu.edu.cn, huangchenyi@stu.scu.edu.cn, fangwenbo@stu.scu.edu.cn,
likeying@stu.scu.edu.cn, mawengang941206@scu.edu.cn, xiaolonglan@scu.edu.cn,
hejunjiang@scu.edu.cn*

***Abstract*—Deep neural networks (DNNs) are vulnerable to adversarial examples generated by adding imperceptible perturbations to the input. To address these threats, detecting adversarial examples has become a promising approach. It ensures that the model runs reliably, even in the presence of adversarial examples. However, existing detection methods still face clear challenges when dealing with adversarial examples built by generative models. The main reason is that these perturbations look more natural in the spatial domain, which weakens the power of spatial-feature-based detection. Notably, such examples are more likely to show abnormal patterns in the frequency domain. After the fast Fourier transform, their spectral distributions often deviate from the statistical patterns of real data. Therefore, we propose SFDD (Spatial-Frequency Domain Detector), an unsupervised detector. It operates in a Class- and Classifier-Free (CCF) setting. SFDD models frequency-domain distributions via normalizing flows and spatial-domain structures through an autoencoder. Then it combines these two anomalous signals for detection. Experiments on CIFAR-10 and Tiny ImageNet against 11 adversarial attacks show that SFDD achieves AUROC scores of 0.85 and 0.94, corresponding to relative improvements of 26.84% and 62.01% over existing methods. For five generative attacks on Tiny ImageNet, our method achieves relative gains of 34.45%, 56.97%, 71.77%, 102.5%, and 73.35%, respectively.**

***Index Terms*—adversarial examples, adversarial defense, unsupervised learning, normalizing flow**

## I. Introduction

Numerous studies have demonstrated the effectiveness of Deep neural networks (DNNs) in various fields such as autonomous driving and medical imaging [1]. However, it is well-known that DNNs remain highly vulnerable to adversarial examples [2]. Adversaries generate adversarial examples by adding imperceptible perturbations to benign images, potentially causing the model to make incorrect predictions. This vulnerability undermines the trustworthiness of deep learning systems. To address this issue, researchers are continuously exploring methods for defending adversarial attacks. [3].

However, defending against adversarial attacks is not trivial. Existing defense methods can be broadly categorized into active defenses and passive defenses [4]. Active defenses enhance model robustness by modifying the training objective or model architecture during training. And adversarial training [5] is a representative active defense strategy that incorporates adversarial examples during training. However, this approach typically incurs degraded clean accuracy and substantial computational overhead [6]. In contrast, passive defenses suppress adversarial examples after training through detection or reconstruction inputs. It provides protection without compromising the classifier's clean accuracy. Adversarial example detection [7] is a representative passive defense approach that aims to identify and reject adversarial inputs before classification. Among such methods, the recently proposed Class- and Classifier-Free (CCF) strategy [4] offers a more practical detection paradigm by employing an unsupervised framework that requires neither labeled data nor access to the target classifier. This design enables effective detection under strict data security and model confidentiality constraints, highlighting its strong applicability in real-world scenarios. However, most existing CCF-based methods are still designed for gradient attacks that exhibit anomalous behavior in the spatial domain, and are powerless against generative attacks. Generative models encode perturbations by implicitly modeling complex high-dimensional data distributions [8], generating adversarial examples that are highly similar to natural images in spatial structure, thereby evading spatial domain detection. Nevertheless, the upsampling operation in generative networks still introduces detectable artifacts in the frequency domain. These artifacts include periodic peaks in the power spectrum and systematic attenuation of high-frequency energy [9] [10].

To address this problem, we propose SFDD (Spatial-Frequency Domain Detector) which brings frequency-domain analysis into the detection framework and complements spatial features. First, SFDD uses Fourier transform (FFT) to obtain spectral representations. It then uses normalizing flows to learn shared frequency-domain representations from normal examples. At the same time, we use a Swin Transformer-based autoencoder to learn the structural features of normal examples and reconstruct them in the spatial domain. Finally, the model completes detection by combining information from both domains.

The main contributions of this work are as follows:

- We design a frequency-domain detection branch, based on latent distribution modeling. It can capture the anomalous frequency-domain patterns of generative attacks. After applying FFT, we employ normalizing flows [11]

to learn a frequency-domain latent representation of the benign example distribution for characterizing spectrum statistical deviations.
- We propose SFDD, an adversarial example detector operating under the unsupervised and CCF setting. SFDD fuses spatial-domain structural information with frequency-domain distribution characteristics to effectively detect both gradient-based and generative attacks.
- We conduct extensive experiments on CIFAR-10 and Tiny ImageNet against 11 adversarial attacks. SFDD achieves AUROC scores of 0.85 and 0.94, representing relative improvements of 26.84% and 62.01% over existing methods. For five generative attacks on Tiny ImageNet, SFDD achieves relative improvements of 34.45%, 56.97%, 71.77%, 102.5%, and 73.35%, respectively.

The remainder of this paper is organized as follows. Section II introduces the relevant techniques. Section III elaborates on the proposed SFDD model and its detection process. Section IV presents thorough adversarial detection experiments on the CIFAR-10 and Tiny ImageNet datasets. Finally, Section V summarizes the paper.

## II. Related Work

### A. Unsupervised Class- and Classifier-Assisted Detection

In the unsupervised setting, Class- and Classifier-Assisted (CCA) methods explicitly leverage class label information or the target classifier to construct detection signals [4]. Common paradigms include reconstruction-based cooperative discrimination, anomaly measures in the feature space, and detection based on consistency under input transformations. Representative methods include MagNet [12], DNR [13], FS [14], and DeepRP [15], which provide strong detection performance under certain threat models. However, a primary limitation is the reliance on the target classifier. The detection performance is often tightly coupled to specific classifier architectures and representations, leading to limited cross-model transferability and generalization. Adversaries can also apply adaptive optimization against the detection signal, reducing robustness and potentially introducing additional inference cost or engineering complexity.

### B. Unsupervised Class- and Classifier-Free Detection

The Class- and Classifier-Free (CCF) paradigm emphasizes independence from class labels and the target classifier [4], modeling the statistical regularities of benign examples and treating distribution-shifted inputs as anomalies. A key advantage is deployment flexibility and model decoupling, where the detector functions as an independent front-end module without assumptions about access to the target classifier, facilitating reuse across models and scenarios.

Representative CCF approaches include statistical feature-based detection methods such as PCL-VLC [16] and LSCF-KDE [17] that extract statistical features from inputs to identify anomalies. Reconstruction error-based detection methods such as autoencoders [12] learn to reconstruct benign examples and use reconstruction error as an anomaly indicator. However, existing CCF detectors are mainly designed for traditional gradient-based perturbations and lack a systematic characterization of the distribution changes caused by perturbations generated by generative models, resulting in insufficient detection stability under generative adversarial examples.

## III. Method

### A. Overview

In the unsupervised CCF setting, given an input image $x \in \mathbb{R}^{H\times W\times C}$, only a set of benign examples $\mathcal{D} = \{x_i\}_{i=1}^{N}$ is available during training. Thus, the detector can only model the benign data distribution. Our goal is to compute a detection score for any input $x$ at inference time to distinguish benign examples from adversarial ones. We further assume a non-adaptive attacker who is unaware of the detector.

The key insight underlying SFDD is that adversarial perturbations induce deviations in both spatial structures and frequency-domain statistics, with such effects often more pronounced for generative model-based attacks. Most existing unsupervised detectors primarily model the spatial domain and fail to capture frequency-domain deviations effectively, making them more likely to miss generative adversarial examples. To address this limitation, we propose a spatial-frequency dual-branch detection framework that models the benign-example distribution in both domains. During inference, SFDD fuses information from both domains to produce the final detection score, improving detection capability and generalization across both gradient-based and generative model-based attacks.

As illustrated in Fig. 1, SFDD consists of three components. A Swin Transformer-style autoencoder [18] learns structural representations of benign examples via reconstruction and builds a spatial-domain structure-residual fingerprint by combining structural cues with reconstruction residuals. A Pyramid Frequency Normalizing Flow [11] performs pyramid FFT decomposition to obtain multi-band spectral representations and learns the frequency-domain distribution of benign examples via normalizing flows. A detection module fuses anomaly measures from the two branches at test time and outputs the final detection score.

### B. Frequency Domain Branch

The frequency-domain branch uses pyramid FFT to obtain spectral representations of the example in three bands, including high-, mid-, and low-frequency bands. Then, we use normalizing flows to learn their latent template representations, respectively, and we propose the anomaly score $d_f$ for detection.

*1) Pyramid FFT Decomposition:* To capture the multi-band perturbations that generative attacks may introduce, we transform the input $x$ to the frequency domain using FFT. We then partition the spectrum into three frequency bands (low, mid, and high) using concentric circular masks $\mathcal{B}_b$ based on the normalized radial frequency $\rho$, where $b \in \{l, m, h\}$. For each

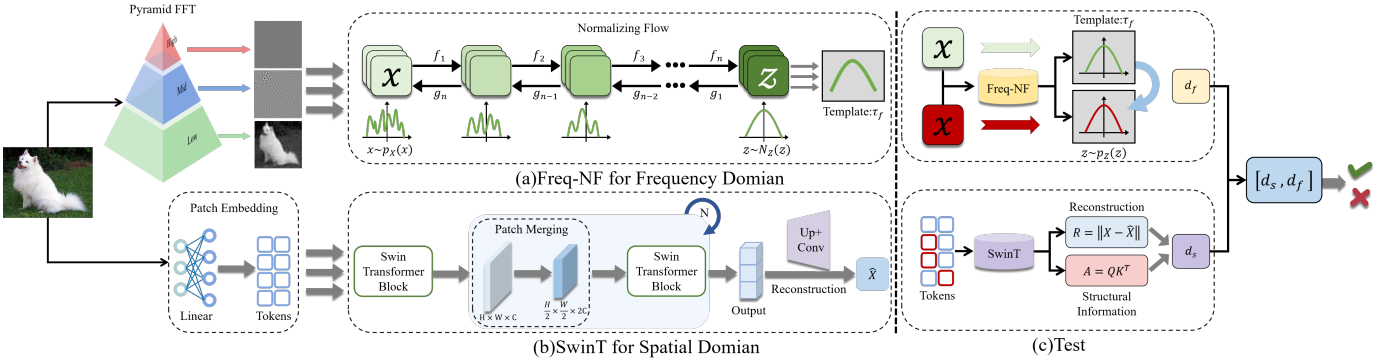


Fig. 1. Overview of the proposed dual-stream unsupervised adversarial example detection framework. (a) Learning latent templates in the frequency domain of benign examples through normalizing flow. (b) Extracting spatial domain structural representations of examples and reconstructing the input based on the Swin Transformer autoencoder. (c) Achieving final detection by fusing $d_f$ and $d_s$.

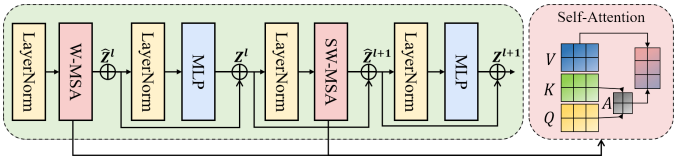


Fig. 2. An illustration of the Swin Transformer architecture, where $A$ denotes the structural information.

band, we compute the log-compressed magnitude spectrum as input to the normalizing flow

$$M_b(x) = \log(1 + |\mathcal{F}(x)|) \odot \mathcal{B}_b, \quad b \in \{l, m, h\}. \tag{1}$$

*2) Normalizing Flow:* We employ invertible affine coupling blocks as the basic building unit [19]. The input features are split along the channel dimension into $x_0$ and $x_1$. The forward mapping $f(\cdot)$ produces outputs $y_0$ and $y_1$, where $y_0 = x_0$ and $x_1$ is transformed by an affine function conditioned on $x_0$. The inverse mapping $g(\cdot) = f^{-1}(\cdot)$ recovers $x_0$ and $x_1$ from $y_0$ and $y_1$. The scale and shift parameters $s(\cdot)$ and $t(\cdot)$ are predicted by a lightweight convolutional network.

A normalizing flow [11] is constructed by cascading multiple coupling blocks. For each band $b \in \{l, m, h\}$, we build a band-specific flow $F_b(\cdot)$ with a Gaussian prior $p_0$ that maps the band input $M_b(x)$ to a latent representation

$$z_b = F_b(M_b(x)), \quad b \in \{l, m, h\}. \tag{2}$$

During training with benign examples, we optimize the flow by minimizing the negative log-likelihood

$$\mathcal{L}_{\text{flow}} = \sum_{b \in \{l,m,h\}} \mathbb{E}_{x \sim \mathcal{D}} \left[ -\log p_0(z_b) - \log \left| \det \left( \frac{\partial F_b}{\partial M_b} \right) \right| \right]. \tag{3}$$

After training, we construct a band-level latent template for benign representations

$$\tau_b = \mathbb{E}_{x \sim \mathcal{D}}(z_b). \tag{4}$$

**Algorithm 1** Training and Inference of SFDD

**Input:** Benign training set $\mathcal{D}$, test example $\mathbf{x}$, clean FPR $\theta$
**Output:** Detection score $d_x$, prediction $y$

```
Train the frequency branch on D using Eq. 1, Eq. 2, and Eq. 3
Compute the benign latent templates {τ_b} using Eq. 4
Train the spatial branch on D using Eq. 6 and Eq. 7
Determine the threshold t on D with clean FPR θ
Compute the frequency-domain score d_f for x using Eq. 1, Eq. 2, and Eq. 5
Compute the spatial-domain score d_s for x using Eq. 6 and Eq. 7
Fuse the two scores and obtain d_x by Eq. 8
if d_x > t then
    y ← adversarial
else
    y ← benign
end if
return d_x, y
```

*3) Anomaly Score Construction:* During inference, we compute the deviation between the test example and the template in each band. The frequency-domain anomaly score is defined as

$$d_f = [\alpha\, D_b(x), \beta\, \mathcal{Z}_b(x), \gamma\, C(x), \sigma\, R_H(x)], \tag{5}$$

where $D_b(x) = \|z_b - \tau_b\|_2$ measures the $L_2$ distance between the latent representation and the template in band $b$. The term $\mathcal{Z}_b(x)$ captures statistical deviation from the expected distribution. The term $C(x)$ quantifies cross-band consistency by measuring the relative relationship across frequency bands. The term $R_H(x)$ represents the ratio of high-frequency energy that often increases under adversarial perturbations. The coefficients $\alpha$, $\beta$, and $\gamma$ balance the contributions of each term.

### C. Spatial Domain Branch

The spatial-domain branch aims to learn the structural information of benign examples. We use a Swin Transformer-style autoencoder to learn the structural information of benign

examples and perform reconstruction. During detection, we combine the structural information and the reconstruction residual to obtain the anomaly score $d_s$.

*1) Swin Transformer Block:* As shown in Fig. 2, we employ Swin Transformer [18] as the encoder backbone. Given an input $x \in \mathbb{R}^{H\times W\times C}$, we first partition it into non-overlapping patches and map them to a token Swin blocks. Unlike global self-attention in the standard Transformer [20], Swin Transformer computes self-attention within local windows of size $M \times M$ and employs shifted windows to enable cross-window interaction. The window-based self-attention is computed as

$$\text{Attention}(x) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V, \tag{6}$$

where $B$ is a learnable relative position bias. We define the pre-softmax correlation term as spatial structure information $A = QK^T$. In the multi-head setting, we concatenate the structure information from all heads for use in the detection score.

*2) Reconstruction:* The encoder incorporates Patch Merging for downsampling, while the decoder performs upsampling for reconstruction. We map the output of the last encoder block $z^n$ back to the input space through a reconstruction function $r(\cdot)$, yielding $\hat{x} = r(z^n)$. The reconstruction residual serves as an anomaly indicator, defined as $R = \|x - \hat{x}\|$. During training with benign examples only, we minimize the reconstruction loss

$$\mathcal{L}_{rec} = \mathbb{E}_{x\sim\mathcal{D}}\|x - \hat{x}\|. \tag{7}$$

The spatial detection score is obtained by fusing the reconstruction residual and the structure information as $d_s = [R, \lambda A]$, where $\lambda$ is a scaling factor balancing their magnitudes.

### D. Detection

The proposed model produces two key anomaly indicators for an input. In the spatial domain, subtle perturbations can disrupt the structural information and prevent proper reconstruction, leading to a larger $d_s$. In the frequency domain, adversarial perturbations can disrupt the flow-based modeling, causing the latent representation to deviate substantially from the benign template and resulting in a larger $d_f$.

To fuse information from both domains, we concatenate the two scores to obtain a matrix and define the final detection score $d_x$

$$d_x = \left\| [d_s,\ d_f] \right\|_p, \tag{8}$$

the detection score $d$ denotes the $p$-norm of the matrix. The larger $d_x$ is, the more likely $x$ is to be adversarial. As SFDD does not rely on adversarial example, we determine the threshold $t$ on benign example set $\mathcal{D}$, such that the clean FPR is controlled to $\theta$.

## IV. Experiments

Comprehensive experiments are conducted on two benchmark datasets to evaluate the detection effectiveness of the proposed SFDD against diverse adversarial attacks. This section describes the experimental setup, presents evaluation results on gradient-based and generative-model-based attacks, and reports ablation studies that validate the contribution of each component.

TABLE I
Classification accuracy of target classifiers on benign examples.

| Dataset | Classifier | Accuracy (%) |
|---|---|---|
| CIFAR-10 | ResNet-18 | 95.14 |
| | DenseNet-121 | 93.12 |
| Tiny ImageNet | ResNet-34 | 77.77 |
| | Swin-Tiny | 82.59 |

### A. Experimental Setup

Datasets. Evaluation is conducted on CIFAR-10 and Tiny ImageNet. CIFAR-10 contains 50,000 training and 10,000 test images across 10 classes, while Tiny ImageNet includes 100,000 training and 10,000 test images spanning 200 classes.

Target Classifiers. Four target classifiers are used across both datasets. On CIFAR-10, ResNet-18 [21] and DenseNet-121 [22] are trained from scratch. On Tiny ImageNet, ResNet-34 [21] and Swin-Tiny [18] are initialized with ImageNet-pretrained weights from timm and fine-tuned. Table I reports their accuracy on benign examples.

Attacks. We evaluate 11 representative attacks from three categories. Gradient-based methods include FGSM [23], BIM [24], PGD [3], CW [25], Square [26], and AutoAttack [27], implemented via torchattacks. GAN-based attacks include CDA [28], M3D [29], and TTP [30], while diffusion-based attacks include Diff-Attack [31] and Diff-PGD [32]. All attacks use $\epsilon = 0.01$ except Square ($\epsilon = 0.03$). Following standard practice [27], evaluation considers only successfully attacked examples.

Baselines. SFDD is compared with eight existing unsupervised adversarial detectors. CCF detectors include PCA-VLC [16], LSCF-KDE [17], AE [12], and CAE [12]. CCA detectors include MagNet [12], DNR [13], FS [14], and DeepRP [15].

Evaluation Metric. Following prior work, AUROC is used as the primary metric, measuring the trade-off between true positive and false positive rates across all thresholds.

Implementation Details. For CIFAR-10 and Tiny ImageNet, the input resolutions are 32×32 and 224×224, and the batch sizes are 128 and 64. Other hyperparameters are the same for both datasets. The spatial branch uses a Swin Transformer encoder with patch size 4, window size 4, depth 4, 8 attention heads, and dropout 0.1. It is trained for 200 epochs with AdamW and a learning rate of 0.001. The frequency branch uses a normalizing flow with 8 affine coupling blocks. It is trained for 200 epochs with AdamW and a learning rate of 0.0002. In all experiments, the clean FPR is fixed at 0.05 to set the threshold $t$.

### B. Detection of Gradient-based Attacks

Table II shows the AUROC results on gradient attacks. Among the baselines, PCA-VLC and LSCF-KDE perform best on CIFAR-10. However, their performance drops a lot on Tiny ImageNet, which shows limited ability on higher dimensional

TABLE II

DETECTION PERFORMANCE (AUROC) AGAINST GRADIENT-BASED ADVERSARIAL ATTACKS. REL. GAIN DENOTES THE RELATIVE IMPROVEMENT OVER THE BASELINE AVERAGE (AVG).

| Dataset | Classifier | Attack | Unsupervised CCF Detector | | | | Unsupervised CCA Detector | | | | Summary | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | PCA-VLC | LSCF-KDE | AE | CAE | MagNet | DNR | FS | DeepRP | AVG | SFDD | Rel. Gain |
| CIFAR-10 | ResNet-18 | FGSM | 0.9451 | **0.9990** | 0.5057 | 0.6828 | 0.5663 | 0.7858 | 0.6852 | 0.5215 | 0.7114 | 0.9100 | 27.91% |
| | | BIM | 0.9097 | **0.9840** | 0.5024 | 0.6173 | 0.8385 | 0.6599 | 0.8665 | 0.5037 | 0.7353 | 0.9074 | 23.41% |
| | | PGD | 0.9205 | **0.9990** | 0.5024 | 0.6243 | 0.8329 | 0.6536 | 0.8671 | 0.5032 | 0.7379 | 0.9328 | 26.42% |
| | | CW | 0.7858 | 0.9190 | 0.5012 | 0.5407 | 0.8286 | **0.9259** | 0.8609 | 0.6260 | 0.7485 | 0.7742 | 3.43% |
| | | Square | 0.8994 | **0.9280** | 0.5397 | 0.8696 | 0.7181 | 0.9067 | 0.7437 | 0.7098 | 0.7894 | 0.9141 | 15.80% |
| | | AutoAttack | 0.8808 | **0.9720** | 0.5015 | 0.6085 | 0.9245 | 0.6787 | 0.9117 | 0.5040 | 0.7477 | 0.8808 | 17.80% |
| | DenseNet-121 | FGSM | 0.9556 | **0.9990** | 0.5058 | 0.7005 | 0.5826 | 0.7241 | 0.7097 | 0.5282 | 0.7132 | 0.9742 | 36.60% |
| | | BIM | 0.9300 | **0.9890** | 0.5020 | 0.6298 | 0.7579 | 0.5521 | 0.8009 | 0.5059 | 0.7084 | 0.9518 | 34.35% |
| | | PGD | 0.9321 | **0.9990** | 0.5020 | 0.6369 | 0.7510 | 0.5563 | 0.7981 | 0.5061 | 0.7102 | 0.9556 | 34.55% |
| | | CW | 0.7329 | 0.8910 | 0.5010 | 0.5457 | 0.7426 | **0.9454** | 0.8319 | 0.7161 | 0.7383 | 0.7833 | 6.09% |
| | | Square | 0.8652 | 0.9030 | 0.5359 | 0.8752 | 0.6833 | 0.8691 | 0.7222 | 0.6995 | 0.7692 | **0.9194** | 19.53% |
| | | AutoAttack | 0.8505 | **0.9610** | 0.5011 | 0.6214 | 0.8015 | 0.5178 | 0.7998 | 0.5091 | 0.6953 | 0.9421 | 35.50% |
| Tiny ImageNet | ResNet-34 | FGSM | 0.5599 | 0.5211 | 0.5008 | 0.5195 | 0.6723 | 0.6869 | 0.6556 | 0.6293 | 0.5932 | **0.8982** | 51.42% |
| | | BIM | 0.5379 | 0.5082 | 0.4993 | 0.5105 | 0.9320 | 0.4881 | 0.6034 | 0.5175 | 0.5746 | **0.9701** | 68.83% |
| | | PGD | 0.5394 | 0.5159 | 0.4994 | 0.5106 | 0.9274 | 0.4827 | 0.6049 | 0.5145 | 0.5744 | **0.9843** | 71.38% |
| | | CW | 0.5059 | 0.5006 | 0.5015 | 0.5019 | 0.5815 | 0.7781 | **0.8652** | 0.5849 | 0.6025 | 0.8015 | 33.04% |
| | | Square | 0.6776 | 0.5565 | 0.4992 | 0.5075 | 0.8647 | 0.7819 | 0.7539 | 0.6536 | 0.6619 | **0.9649** | 45.79% |
| | | AutoAttack | 0.5273 | 0.5115 | 0.5099 | 0.5886 | 0.5837 | 0.5580 | 0.6038 | 0.5617 | 0.5556 | **0.9664** | 73.95% |
| | Swin-Tiny | FGSM | 0.5527 | 0.5224 | 0.5008 | 0.5207 | 0.6658 | 0.6477 | 0.6508 | 0.5975 | 0.5823 | **0.9979** | 71.37% |
| | | BIM | 0.5512 | 0.5197 | 0.4991 | 0.5119 | 0.8292 | 0.6866 | 0.6968 | 0.5515 | 0.6058 | **0.9975** | 64.67% |
| | | PGD | 0.5527 | 0.5182 | 0.4991 | 0.5120 | 0.8294 | 0.6823 | 0.6948 | 0.5488 | 0.6047 | **0.9975** | 64.97% |
| | | CW | 0.5073 | 0.5020 | 0.5012 | 0.5022 | 0.6080 | 0.8003 | **0.9000** | 0.5650 | 0.6108 | 0.8967 | 46.82% |
| | | Square | 0.6784 | 0.5697 | 0.4989 | 0.5090 | 0.8203 | 0.8592 | 0.7868 | 0.6453 | 0.6710 | **0.9687** | 44.38% |
| | | AutoAttack | 0.5411 | 0.5154 | 0.5101 | 0.5884 | 0.5963 | 0.7255 | 0.7391 | 0.5936 | 0.6012 | **0.9967** | 65.79% |
| **Overall Average** | | | 0.7225 | 0.7418 | 0.5050 | 0.5932 | 0.7474 | 0.7064 | 0.7564 | 0.5748 | 0.6684 | **0.9286** | 40.99% |

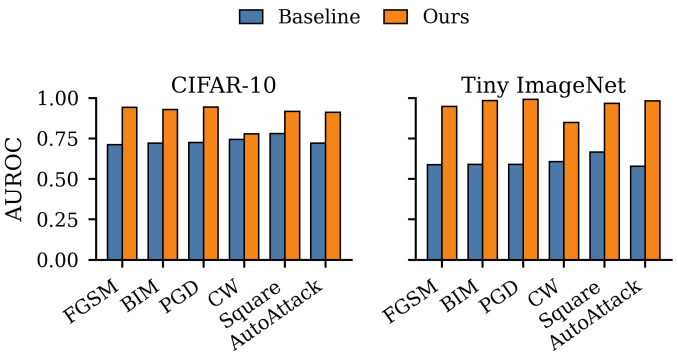


Fig. 3. Average AUROC under gradient-based attacks. The left and right panels correspond to CIFAR-10 and Tiny ImageNet, respectively.

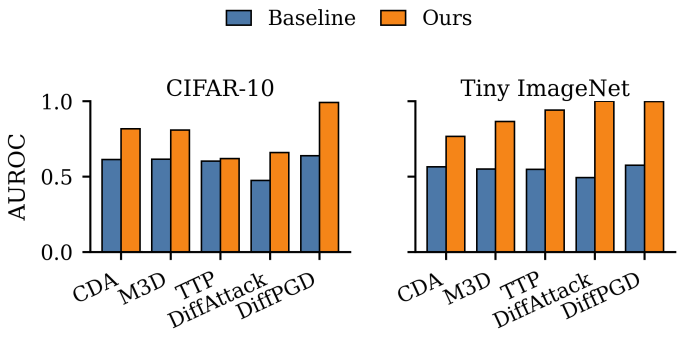


Fig. 4. Average AUROC under generative-model-based attacks. The left and right panels correspond to CIFAR-10 and Tiny ImageNet, respectively.

data. Other baselines do not give stable results across different attacks and classifiers.

SFDD shows strong and stable performance on both datasets. Among 24 settings, SFDD reaches an AUROC above 0.90 in 18 cases. Compared with the baseline average, SFDD improves by 40.99% on average, with a maximum gain of 73.95% on Tiny ImageNet with ResNet-34 under AutoAttack. The stable results across different classifiers and attacks show the robustness of the dual-branch design.

### C. Detection of Generative-based Attacks

Table III reports the detection results on GAN and diffusion attacks. Existing unsupervised detectors are mainly designed for gradient perturbations, so they work poorly on generative attacks. This is because generative attacks have very different noise patterns. The perturbations from GANs and diffusion models are uneven, which makes many methods fail. Among the baselines, PCA-VLC and LSCF-KDE perform better, while others are close to random.

By using frequency information, SFDD detects generative

TABLE III

DETECTION PERFORMANCE (AUROC) AGAINST GENERATIVE-MODEL-BASED ADVERSARIAL ATTACKS. REL. GAIN DENOTES THE RELATIVE IMPROVEMENT OVER THE BASELINE AVERAGE (AVG).

| Dataset | Classifier | Attack | Unsupervised CCF Detector | | | | Unsupervised CCA Detector | | | | Summary | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | PCA-VLC | LSCF-KDE | AE | CAE | MagNet | DNR | FS | DeepRP | AVG | SFDD | Rel. Gain |
| CIFAR-10 | ResNet-18 | CDA | 0.8049 | 0.9296 | 0.5061 | 0.5586 | 0.5299 | 0.4898 | 0.8383 | 0.5892 | 0.6558 | **0.9751** | 48.69% |
| | | M3D | 0.8775 | **0.9467** | 0.4881 | 0.5062 | 0.4661 | 0.5307 | 0.5205 | 0.5196 | 0.6069 | 0.8005 | 31.89% |
| | | TTP | 0.8138 | **0.9648** | 0.5059 | 0.5213 | 0.5214 | 0.4754 | 0.5633 | 0.5423 | 0.6135 | 0.6223 | 1.43% |
| | | DiffAttack | 0.5038 | 0.4998 | 0.3621 | 0.3035 | 0.3912 | 0.4878 | 0.6312 | **0.7534** | 0.4916 | 0.6519 | 32.61% |
| | | DiffPGD | 0.9891 | 0.9928 | 0.5105 | 0.5642 | 0.5477 | 0.5232 | 0.5131 | 0.5727 | 0.6517 | **0.9975** | 53.07% |
| | DenseNet-121 | CDA | 0.6277 | **0.6796** | 0.5079 | 0.5486 | 0.5685 | 0.4887 | 0.5792 | 0.5815 | 0.5727 | 0.6594 | 15.14% |
| | | M3D | 0.9162 | **0.9675** | 0.4997 | 0.5090 | 0.4966 | 0.5233 | 0.5379 | 0.5523 | 0.6253 | 0.8185 | 30.89% |
| | | TTP | 0.7764 | **0.9162** | 0.4830 | 0.5037 | 0.5500 | 0.5141 | 0.4913 | 0.5169 | 0.5940 | 0.6178 | 4.02% |
| | | DiffAttack | 0.4981 | 0.4887 | 0.3555 | 0.2959 | 0.3862 | 0.4492 | 0.5542 | 0.6392 | 0.4584 | **0.6688** | 45.91% |
| | | DiffPGD | **0.9894** | 0.9757 | 0.5059 | 0.5471 | 0.5123 | 0.4754 | 0.5292 | 0.4669 | 0.6252 | 0.9885 | 58.10% |
| Tiny ImageNet | ResNet-34 | CDA | 0.5047 | 0.6030 | 0.5005 | 0.5124 | 0.5689 | 0.5027 | 0.5462 | **0.6368** | 0.5469 | 0.5551 | 1.50% |
| | | M3D | 0.5192 | **0.8686** | 0.4905 | 0.5033 | 0.5052 | 0.4297 | 0.4975 | 0.5701 | 0.5480 | 0.8057 | 47.02% |
| | | TTP | 0.5126 | 0.5625 | 0.5015 | 0.5077 | 0.5668 | 0.4426 | 0.5586 | 0.6196 | 0.5340 | **0.8866** | 66.03% |
| | | DiffAttack | 0.3078 | 0.4166 | 0.4350 | 0.3865 | 0.8591 | 0.4572 | 0.6432 | 0.4563 | 0.4952 | **1.0000** | 101.93% |
| | | DiffPGD | 0.5649 | 0.9369 | 0.5016 | 0.5226 | 0.5308 | 0.4472 | 0.5165 | 0.5357 | 0.5695 | **0.9994** | 75.48% |
| | Swin-Tiny | CDA | 0.5852 | 0.9485 | 0.4962 | 0.5213 | 0.5270 | 0.4980 | 0.4924 | 0.6180 | 0.5858 | **0.9808** | 67.42% |
| | | M3D | 0.5245 | 0.8849 | 0.4957 | 0.4974 | 0.5055 | 0.4971 | 0.4844 | 0.5467 | 0.5545 | **0.9256** | 66.92% |
| | | TTP | 0.6002 | 0.8666 | 0.4959 | 0.5214 | 0.5287 | 0.5058 | 0.4401 | 0.5377 | 0.5621 | **0.9977** | 77.51% |
| | | DiffAttack | 0.3324 | 0.4604 | 0.4405 | 0.3961 | 0.4342 | 0.5081 | 0.7345 | 0.6314 | 0.4922 | **1.0000** | 103.17% |
| | | DiffPGD | 0.5929 | 0.9469 | 0.5012 | 0.5249 | 0.5229 | 0.5012 | 0.4667 | 0.6124 | 0.5836 | **0.9994** | 71.24% |
| **Overall Average** | | | 0.6421 | 0.7928 | 0.4792 | 0.4876 | 0.5260 | 0.4874 | 0.5569 | 0.5749 | 0.5683 | **0.8475** | 50.00% |

TABLE IV

ABLATION STUDY ON $d_f$ AND $d_s$ UNDER GENERATIVE-BASED ATTACKS (AUROC)

| Dataset | Classifier | Attack | w/o $d_f$ | w/o $d_s$ | SFDD |
|---|---|---|---|---|---|
| Tiny ImageNet | ResNet-34 | CDA | 0.5547 | 0.5232 | **0.5551** |
| | | M3D | 0.7876 | 0.6230 | **0.8057** |
| | | TTP | 0.8866 | 0.5308 | **0.8866** |
| | | DiffAttack | 0.6918 | 0.5974 | **1.0000** |
| | | DiffPGD | 0.9970 | 0.8224 | **0.9994** |
| | Swin-Tiny | CDA | 0.9583 | 0.9488 | **0.9808** |
| | | M3D | 0.8876 | 0.7051 | **0.9256** |
| | | TTP | 0.9964 | 0.8090 | **0.9977** |
| | | DiffAttack | 0.6972 | 0.6172 | **1.0000** |
| | | DiffPGD | 0.9993 | 0.9120 | **0.9994** |

TABLE V

FPR@95TPR RESULTS OF DIFFERENT CLASSIFIERS UNDER VARIOUS ATTACKS (TINY IMAGENET)

| Classifier | CDA | M3D | TTP | DiffAttack | DiffPGD |
|---|---|---|---|---|---|
| Swin-Tiny | 0.003 | 0.008 | 0.007 | 0.001 | 0.002 |
| ResNet34 | 0.047 | 0.009 | 0.138 | 0.001 | 0.003 |

attacks well in most cases. Compared with the average of baselines, SFDD achieves up to 103.17% improvement on Tiny ImageNet with Swin-Tiny under DiffAttack. The average AUROC of SFDD is 0.7801 on CIFAR-10 and 0.9089 on Tiny ImageNet, with an overall average of 0.8475 across 20 settings. These results show that the dual-branch design, which uses both spatial and frequency information, can detect both

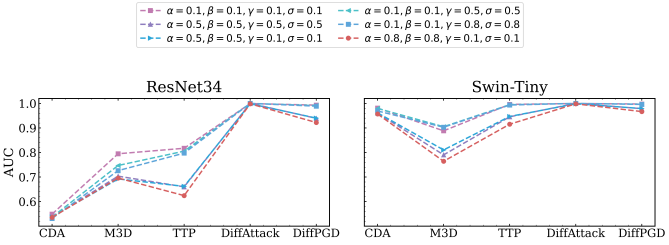


Fig. 5. The AUROC of different $\alpha, \beta, \gamma\, \sigma$ in Eq. 5.

gradient attacks and generative attacks effectively.

We report the detector performance under different parameter settings in Fig. 5. For ResNet34, the AUROC fluctuates by nearly 0.19 on TTP and 0.10 on M3D, while Swin-Tiny still exhibits a considerable variation of about 0.14 on M3D. These results show that detector performance correlates with frequency-domain coefficients, and can be improved through proper hyperparameter tuning. Moreover, incorporating frequency-domain information significantly enhances the model's ability to characterize and detect generative attacks, demonstrating the effectiveness of frequency-domain features for this task.

### *D. Ablation Study*

The detection score in Eq. 8 integrates spatial $d_s$ and frequency $d_f$ components. To assess their individual contributions, we perform ablation studies by removing each

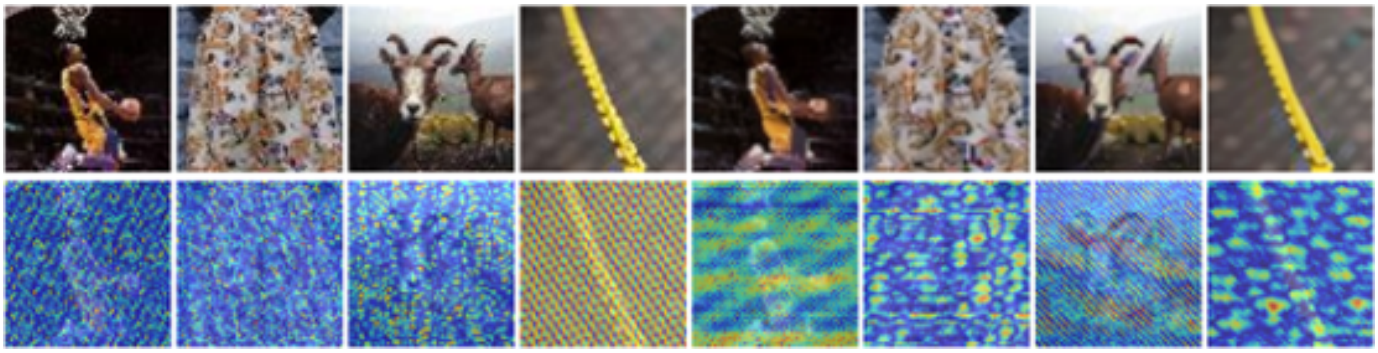

Fig. 6. Visualization of frequency-domain anomaly detection. The first row shows input images, with benign examples on the left and adversarial examples on the right. The second row shows heatmaps from the Freq-NF model, which highlight frequency-domain changes caused by adversarial perturbations.

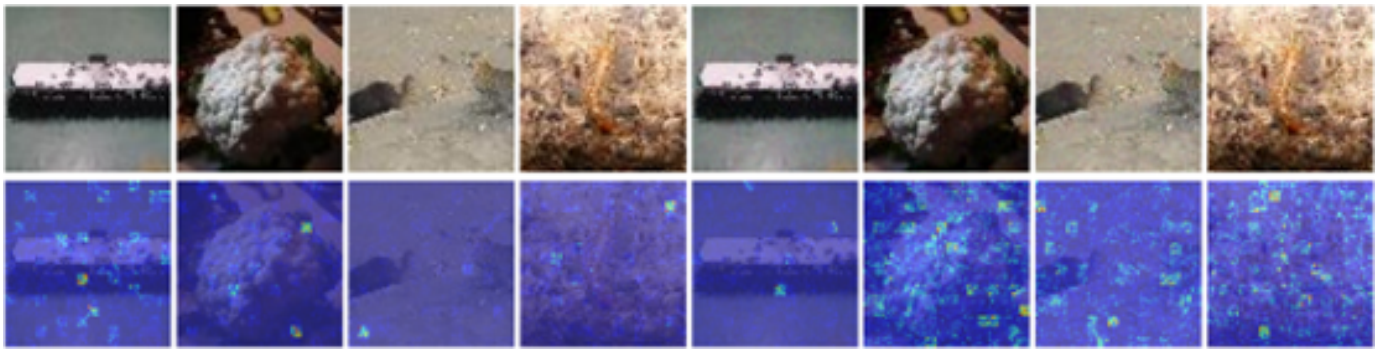

Fig. 7. Visualization of spatial-domain reconstruction. The first row shows input images, with benign examples on the left and adversarial examples on the right. The second row shows attention heatmaps from the Swin Transformer autoencoder, which reveal structural changes caused by adversarial perturbations.

branch separately. Experiments are conducted on Tiny ImageNet under generative-model-based attacks, a challenging setting where frequency-domain cues are expected to be most beneficial.

Effectiveness of Spatial- and Frequency-domain Information. As shown in Figs. 7 and 6, generative-based attacks produce distinctive frequency-domain artifacts, while adversarial perturbations also impair image structure and reconstruction quality. Therefore, the frequency-domain score $d_f$ and spatial-domain score $d_s$ provide complementary cues for detection. Table IV shows that adding $d_f$ improves the AUROC in all 10 configurations, with especially large gains under DiffAttack (from 0.6918 to 1.0000 on ResNet-34 and from 0.6972 to 1.0000 on Swin-Tiny), while adding $d_s$ improves the AUROC in all 10 configurations, with gains exceeding 0.15 in several cases (e.g., from 0.5308 to 0.8866 under TTP on ResNet-34). These results verify that both branches are important and complementary, enabling SFDD to effectively detect both gradient-based and generative-model-based adversarial examples.

The ablation results collectively demonstrate that both the spatial-domain and frequency-domain branches contribute to the overall detection performance. The two branches capture complementary information, enabling SFDD to effectively detect both gradient-based and generative-model-based adversarial examples.

## V. Conclusion

This paper presents SFDD, an unsupervised method for adversarial detection in the CCF setting. SFDD uses a dual-branch design to jointly model spatial structures and frequency statistics, which addresses the weak use of frequency cues in existing unsupervised methods. As a result, it remains effective even when generative attacks look natural in the spatial domain. Experiments on CIFAR-10 and Tiny ImageNet with both gradient-based and generative-model-based attacks show that SFDD achieves up to 103.17% relative AUROC improvement over the average baseline. Ablation studies further show that the frequency branch is crucial for detecting generative attacks, while spatial-only methods perform nearly at random. These results suggest that combining spatial and frequency information is an effective way to improve unsupervised adversarial detection. However, SFDD does not yet consider adaptive attacks, and its generalization and robustness on larger and more complex datasets still need further study.

## Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 62402330); in part

by the Sichuan Provincial Science and Technology Department regional innovation cooperation key project (Grant No.2025YFHZ0265); in part by the Youth Science Foundation of Sichuan,(No.2025ZNSFSC1474).

## REFERENCES


[1] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi, "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability," *Computer Science Review*, vol. 37, p. 100270, 2020.

[2] J. C. Costa, T. Roxo, H. Proença, and P. R. M. Inacio, "How deep learning sees the world: A survey on adversarial attacks & defenses," *IEEE Access*, vol. 12, pp. 61 113–61 136, 2024.

[3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[4] X. Kong, X. Jiang, Z. Song, and Z. Ge, "Data id extraction networks for unsupervised class-and classifier-free detection of adversarial examples," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[5] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International conference on machine learning*. PMLR, 2019, pp. 7472–7482.

[6] Q. Wang, C. Li, Y. Luo, H. Ling, S. Huang, R. Jia, and N. Yu, "Detecting adversarial data using perturbation forgery," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 13 917–13 926.

[7] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey, "Characterizing adversarial subspaces using local intrinsic dimensionality," *arXiv preprint arXiv:1801.02613*, 2018.

[8] J. Liu, Y. Li, Y. Guo, Y. Liu, J. Tang, and Y. Nie, "Generation and countermeasures of adversarial examples on vision: a survey," *Artificial Intelligence Review*, vol. 57, no. 8, p. 199, 2024.

[9] R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva, "Intriguing properties of synthetic images: from generative adversarial networks to diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 973–982.

[10] R. Durall Lopez, M. Keuper, F.-J. Pfreundt, and J. Keuper, "Unmasking deepfakes with simple features," 2019.

[11] J. Lei, X. Hu, Y. Wang, and D. Liu, "Pyramidflow: High-resolution defect contrastive localization using pyramid normalizing flow," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 143–14 152.

[12] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 135–147.

[13] A. Sotgiu, A. Demontis, M. Melis, B. Biggio, G. Fumera, X. Feng, and F. Roli, "Deep neural rejection against adversarial examples," *EURASIP Journal on Information Security*, vol. 2020, no. 1, p. 5, 2020.

[14] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *arXiv preprint arXiv:1704.01155*, 2017.

[15] N. Drenkow, N. Fendley, and P. Burlina, "Attack agnostic detection of adversarial examples via random subspace analysis," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 472–482.

[16] D. Hendrycks and K. Gimpel, "Early methods for detecting adversarial images," *arXiv preprint arXiv:1608.00530*, 2016.

[17] J. Cheng, M. Hussein, J. Billa, and W. AbdAlmageed, "Attack-agnostic adversarial detection," *arXiv preprint arXiv:2206.00489*, 2022.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[19] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803*, 2016.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[23] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[24] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[25] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. Ieee, 2017, pp. 39–57.

[26] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: a query-efficient black-box adversarial attack via random search," in *European conference on computer vision*. Springer, 2020, pp. 484–501.

[27] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International conference on machine learning*. PMLR, 2020, pp. 2206–2216.

[28] M. M. Naseer, S. H. Khan, M. H. Khan, F. Shahbaz Khan, and F. Porikli, "Cross-domain transferability of adversarial perturbations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[29] A. Zhao, T. Chu, Y. Liu, W. Li, J. Li, and L. Duan, "Minimizing maximum model discrepancy for transferable black-box targeted attacks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 8153–8162.

[30] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "On generating transferable targeted perturbations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7708–7717.

[31] J. Chen, H. Chen, K. Chen, Y. Zhang, Z. Zou, and Z. Shi, "Diffusion models for imperceptible and transferable adversarial attack," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[32] H. Xue, A. Araujo, B. Hu, and Y. Chen, "Diffusion-based adversarial sample generation for improved stealthiness and controllability," *Advances in Neural Information Processing Systems*, vol. 36, pp. 2894–2921, 2023.