DIFPA: Boosting Adversarial Transferability via Diffusion-based Importance-aware Feature Pyramid Attack

Bin Chen
College of Computer and Data Science, Fuzhou University
Fuzhou, China
c_chenbin@foxmail.com

Jia-Li Yin✉
College of Computer and Data Science, Fuzhou University
Fuzhou, China
jlyin@fzu.edu.cn

Zhaozhe Hu
College of Computer and Data Science, Fuzhou University
Fuzhou, China
zhaozhehu00@gmail.com

Genna Ba
College of Computer and Data Science, Fuzhou University
Fuzhou, China
bgnbortala@outlook.com

Xiaolei Liu
National Interdisciplinary Research Center of Engineering Physics, China Academy of Engineering Physics
Mianyang, China
luxaole@gmail.com

Wei Lin
College of Mathematics and Computer Science, Fujian University of Technology
Fuzhou, China
wlin@fjut.edu.cn

Yuanlong Yu✉
College of Computer and Data Science, Fuzhou University
Fuzhou, China
yu.yuanlong@fzu.edu.cn

Shouling Ji
College of Computer Science and Technology, Zhejiang University
Hangzhou, China
sji@zju.edu.cn

Abstract

Deep neural networks (DNNs) are known to be vulnerable to adversarial examples, raising serious security concerns in real-world applications. In black-box settings, transfer-based attacks exploit the transferability of adversarial examples crafted on a surrogate model to fool unknown target models, making transferability a central issue in practical adversarial attacks. Recent feature-level transfer attacks improve transferability by perturbing intermediate representations that are more likely to be shared across architectures. However, their effectiveness is still limited by unreliable feature-importance estimation and an over-reliance on single-layer features, which together hinder the accurate localization and suppression of transferable intrinsic features. In particular, existing methods often estimate feature importance from a single input or heuristic transformations, making the resulting indicators easily contaminated by unstable non-robust responses. We observe that, as the diffusion forward process progressively perturbs an input, unstable non-robust responses are gradually attenuated, while semantically relevant robust responses become more concentrated. Inspired by this observation, we propose the **D**iffusion-based **I**mportance-aware **F**eature **P**yramid **A**ttack (**DIFPA**), a novel feature-level attack that improves transferability from both importance estimation and hierarchical feature modeling. Specifically, we introduce a Diffusion-Guided Importance Estimator (DGIE), which exploits the diffusion forward marginal as a principled and controllable neighborhood sampler to probe local feature stability, suppress unstable non-robust responses, and produce more reliable feature-importance indicators. We further propose a Hierarchical Feature Pyramid Attack (HFPA), which systematically aggregates important features across multiple layers and exploits the hierarchical semantic diversity of DNNs to suppress transferable intrinsic features in a pyramid manner rather than at a single layer. Extensive experiments demonstrate that DIFPA consistently outperforms state-of-the-art methods, achieving average improvements of 8.17% against normally trained models and 7.92% against defended models.

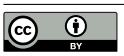
CCS Concepts

• **Security and privacy → Domain-specific security and privacy architectures**.

Keywords

Adversarial examples, transfer-based attack, feature-level attack, diffusion process, feature pyramid

ACM Reference Format:
Bin Chen, Jia-Li Yin, Zhaozhe Hu, Genna Ba, Xiaolei Liu, Wei Lin, Yuanlong Yu, and Shouling Ji. 2026. DIFPA: Boosting Adversarial Transferability via Diffusion-based Importance-aware Feature Pyramid Attack. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835321



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835321

1 Introduction

Deep neural networks (DNNs) achieve strong performance in image classification [39], object detection [6, 45], and image generation [47, 72], but remain vulnerable to imperceptible adversarial

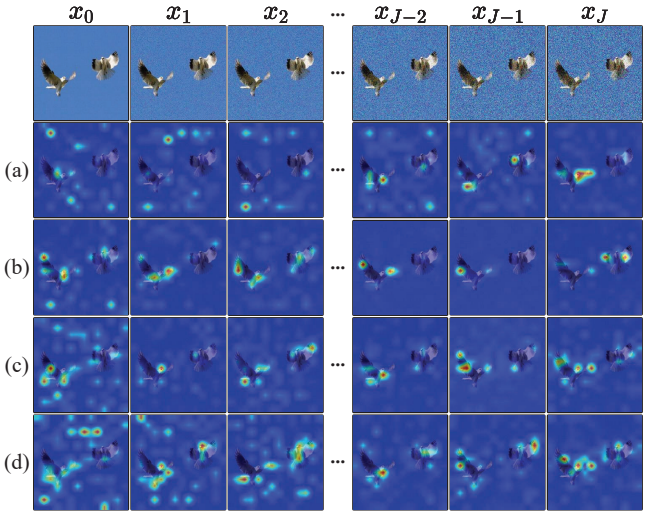


Figure 1: Visualization of the image diffusion forward process and its impact on feature importance localization across the layers of ViT-B. The top row shows progressively noised samples from left to right. Rows (a)–(d) report the corresponding feature-importance indicators computed at *Layers 1*, *4*, *9*, and *11*, respectively. Red represents important features, while blue represents unimportant features. It is noteworthy that different layers of the same model focus on significantly different regions, and as diffusion noise increases, attention across all layers becomes increasingly concentrated on semantically relevant regions.

perturbations that induce erroneous predictions and threaten deployed systems [11, 53, 58, 80]. Evaluating this vulnerability is therefore essential to improving model robustness.

Adversarial attacks operate in white-box [15, 41] or black-box [64, 65, 78] settings. Transfer-based attacks are particularly practical in the latter because adversarial examples crafted on a surrogate model can attack unknown targets without queries. Existing methods improve transferability through optimization [59], input transformations [61], or feature-level perturbations [27, 31, 57]. Feature-level attacks directly manipulate intermediate representations that are often shared across architectures.

Despite their progress, existing feature-level attacks still suffer from two limitations. First, feature-importance estimation is often unreliable. Most methods estimate importance from a single input or heuristic transformations, making the resulting indicators easily contaminated by unstable non-robust responses. As shown in the first column of Fig. 1, the resulting importance maps often contain considerable irrelevant noise. Second, most methods rely on a single layer, overlooking the hierarchical structure of DNNs. Fig. 1(a)–(d) shows that different layers focus on distinct regions, suggesting that transferable cues are inherently multi-level rather than confined to one layer.

A key challenge, therefore, is to distinguish stable transferable cues from locally unstable model-specific responses when estimating feature importance. We observe that, as the diffusion forward process progressively perturbs an input, unstable non-robust responses are gradually attenuated, while semantically relevant robust responses become more concentrated, as illustrated in Fig. 1. This observation suggests that transferable importance should be estimated from a local neighborhood rather than from a single point. Among possible neighborhood samplers, the diffusion forward marginal provides a closed-form and controllable perturbation distribution with negligible additional overhead. Unlike diffusion-based adversarial synthesis, our method uses only the forward diffusion process to obtain diffusion-corrupted samples for feature-importance estimation.

Motivated by this observation, we propose the **D**iffusion-based **I**mportance-aware **F**eature **P**yramid **A**ttack (**DIFPA**), a novel feature-level transfer attack. Specifically, we first introduce a **D**iffusion-**G**uided **I**mportance **E**stimator (DGIE), which exploits the diffusion forward marginal as a principled neighborhood sampler and averages layer-wise importance indicators over diffusion-corrupted local samples, thereby suppressing unstable non-robust responses and producing more reliable importance cues. We further propose a **H**ierarchical **F**eature **P**yramid **A**ttack (HFPA), which aggregates important features across multiple layers and performs hierarchical feature suppression instead of attacking a single feature map. By combining diffusion-guided importance estimation with hierarchical multi-layer suppression, DIFPA substantially improves adversarial transferability.

Experiments on normally trained and defended models show that DIFPA outperforms state-of-the-art attacks. With Res-18 as the surrogate, it improves average black-box attack success rate by 8.17% and 7.92% in the two settings, respectively, while retaining a favorable effectiveness–efficiency trade-off. Our contributions are:

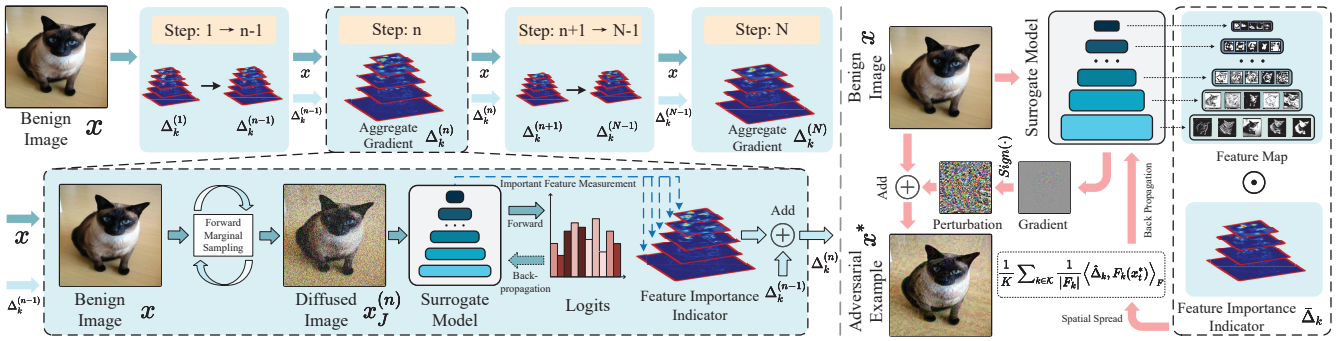
- We propose DGIE, which incorporates the diffusion forward process into feature-importance estimation and produces more reliable transferable importance indicators by suppressing unstable non-robust responses.
- We propose HFPA, which overcomes the limitation of single-layer feature attacks by aggregating important features across multiple layers and exploiting the hierarchical semantic diversity of DNNs.
- We unify diffusion-guided neighborhood-based importance estimation and hierarchical multi-layer feature suppression into a single transferable attack framework under standard ℓ_∞ constraints.
- Extensive experiments demonstrate that DIFPA consistently outperforms previous state-of-the-art methods across normally trained and defended models.

2 Related Work

In this section, we briefly review adversarial attacks, defenses, and diffusion-based adversarial methods related to this work.

2.1 Adversarial Attacks

Adversarial attacks are white-box or black-box depending on access to the victim model. White-box methods, including FGSM [15] and PGD [41], directly use model gradients but are not applicable without full access. Black-box attacks instead rely on victim queries



(a) Diffusion-Guided Importance Estimator (DGIE) (b) Hierarchical Feature Pyramid Attack (HFPA)

Figure 2: Framework of DIFPA. (a) DGIE samples diffusion-corrupted neighborhood points using the closed-form forward diffusion marginal, computes layer-wise feature-importance indicators, and averages them over N samples. (b) HFPA uses the resulting indicators to suppress important features across multiple layers and generate the adversarial example.

through random search [48], evolutionary strategies [23], or gradient estimation [56], or exploit adversarial transferability [44]. Transfer attacks enhance transferability via optimization [8, 12, 13, 35], input transformations [9, 33, 40, 46, 60, 63, 69], surrogate ensembles [3, 71], or feature-level perturbations.

Feature-level attacks manipulate intermediate representations shared across architectures. Prior work perturbs intermediate features [81], refines adversarial examples in feature space [17, 22], estimates and attacks important features [30, 65, 66, 78, 79], or optimizes intermediate directions and regions [31, 32]. DIFPA instead targets two underexplored issues: noisy importance estimation and single-layer structural bias.

2.2 Adversarial Defenses

Representative model-side defenses include feature-aware and ensemble adversarial training, mixing-based augmentation, and adversarially robust class-incremental learning [20, 21, 75, 77]. Input preprocessing defenses employ denoising [34], random resizing [68], smoothing [26], compression [38, 73], and purification [42]. We evaluate transfer attacks against both categories.

2.3 Diffusion-based Adversarial Methods

Diffusion models have been used for adversarial generation [4, 7], purification or denoising [5, 28, 43], and robustness-oriented representation learning [1, 62, 74]. In contrast, DIFPA neither trains a diffusion model nor performs reverse denoising or generative sampling. It uses only the forward marginal $q(x_J|x_0)$ to define local neighborhoods for stable feature-importance estimation under standard ℓ_∞ constraints; diffusion is therefore estimation-oriented rather than generation-oriented.

3 Methodology

Fig. 2 illustrates the overall pipeline of DIFPA, which consists of two stages: DGIE in Fig. 2(a) and HFPA in Fig. 2(b). We describe them in detail below.

3.1 Preliminaries

3.1.1 General Adversarial Attacks. Assume a victim model $f_v : x \mapsto y$, where x and y denote the benign image and true label, respectively. Our objective is to craft adversarial examples $x^* = x+\delta$ using a surrogate model f_s (denoted as f hereafter), where the perturbation δ satisfies:

$$\arg \max_{\|\delta\|_p \le \epsilon} \mathcal{L}(f(x+\delta), y)\,, \tag{1}$$

where $\mathcal{L}(\cdot)$ denotes the cross-entropy loss, ϵ is the perturbation budget, and p specifies the order of the ℓ_p-norm used to constrain the perturbation magnitude. Following prior transfer attacks [27, 31, 36, 57], we set $p = \infty$ in this work.

3.1.2 Feature-Level Transfer Attacks. Instead of manipulating only the final logits, recent methods improve transferability by perturbing *intermediate representations* that are more likely shared across architectures [65, 78, 79]. Assuming the model has a total of L layers, let $F_k(\cdot)$ denote the feature map at the k-th layer, where $k \in \{1, \ldots, L\}$, and let $f(\cdot,\cdot)$ denote the logit output with respect to the true label y. A common strategy is to identify *important intrinsic features* via a feature-importance indicator:

$$\Delta_k(x) \triangleq \frac{\partial f(x,y)}{\partial F_k(x)}, \tag{2}$$

and then suppress these features in the adversarial example:

$$\arg \min_{\|\delta\|_p \le \epsilon} \sum \big(\Delta_{k^*}(x) \odot F_{k^*}(x+\delta)\big). \tag{3}$$

where k^* denotes a pre-selected layer, and the summation reduces the feature map over all spatial locations and channels to a scalar loss.

While Eq. (3) forms the basis of feature-level attacks, two fundamental limitations hinder their transferability: (1) **Susceptibility to Non-Robust Features:** The gradient $\Delta_k(x)$ derived directly from original data is inherently *noisy*. It entangles robust semantic structures with model-specific high-frequency patterns, resulting in importance maps that fail to pinpoint intrinsic features shared across diverse architectures. (2) **Structural Overfitting via Single-Layer Reliance:** Relying on a single layer neglects the hierarchical nature of DNNs. High-level features are spatially sparse and abstract,

while low-level features contain rich structural details. Focusing on a specific depth induces structural overfitting, limiting the adaptability of the attack to target models with different receptive fields. These two limitations respectively motivate the two components of our method: DGIE addresses the reliability of feature-importance estimation, while HFPA addresses the structural bias introduced by single-layer reliance.

3.2 Diffusion-Guided Importance Estimator

To address the unreliability of single-point feature-importance estimation described above, DGIE estimates more reliable feature-importance indicators from diffusion-corrupted neighborhood samples, as shown in Fig. 2(a).

3.2.1 Robust vs. Non-Robust Features under the Manifold Hypothesis. Following the feature-based view of adversarial vulnerability [24], a classifier may exploit both *robust* and *non-robust* predictive cues. From the manifold perspective, natural images concentrate near a low-dimensional data manifold $\mathcal{M}$. Robust cues tend to be consistent with semantic variations that stay on (or close to) $\mathcal{M}$, whereas non-robust cues are often highly sensitive to directions that deviate from $\mathcal{M}$. Consistently, prior work distinguishes regular (unconstrained) adversarial examples from on-manifold ones, and shows that regular adversarial examples typically leave the manifold [51]. Moreover, gradient-based explanations that align with the manifold tangent space are empirically more perceptually meaningful [2]. These observations motivate estimating feature importance in a way that downweights locally unstable, off-manifold-sensitive components.

3.2.2 Diffusion-Induced Local Neighborhood. Given a benign image x_0, we draw a neighborhood point at diffusion step J using the closed-form forward-diffusion marginal [19, 50]:

$$x_J = \sqrt{\bar{\alpha}_J}\, x_0 + \sqrt{1-\bar{\alpha}_J}\, z, \quad z \sim \mathcal{N}(m, \sigma^2 \mathbf{I}), \tag{4}$$

where $\bar{\alpha}_J$ controls the neighborhood size and (m, σ) parameterize the corruption strength. Since Eq. (4) can be represented as a closed-form forward-diffusion process, the result of J steps can be calculated in a single step. This ensures that the method does not incur excessive computational overhead compared to other feature-level adversarial attacks, as demonstrated in subsequent experiments and Fig. 6. Following the previous work [31, 65, 78], we performed N independent samples and averaged the corresponding indicators to obtain the final k-layer important feature indicators, where n indexes the n-th sample:

$$\bar{\Delta}_k = \frac{1}{N}\sum_{n=1}^{N} \Delta_k^{(n)}, \qquad \Delta_k^{(n)} = \frac{\partial f(x_J^{(n)}, y)}{\partial F_k(x_J^{(n)})}. \tag{5}$$

3.2.3 Why Diffusion Helps Separate Robust/Non-Robust Features? As illustrated in Fig. 3, diffusion-style noising suppresses unstable non-robust cues while preserving relatively stable robust semantic features, so that the resulting feature-importance indicators become less contaminated by spurious responses. Under the manifold hypothesis, deviations from $\mathcal{M}$ correspond to directions in which model responses can change sharply. Sampling $x_J \sim q(\cdot \mid x_0)$ explores a local neighborhood that includes such off-manifold deviations. Importance components tied to off-manifold-sensitive (non-robust) cues tend to fluctuate with higher variance across nearby neighborhood points, and are thus attenuated when importance is measured in a neighborhood-based manner. In contrast, importance components associated with manifold-consistent (robust) cues are more stable under local neighborhood sampling and remain prominent. Therefore, diffusion-induced neighborhood sampling biases the resulting importance signal toward more stable, cross-model consistent evidence.

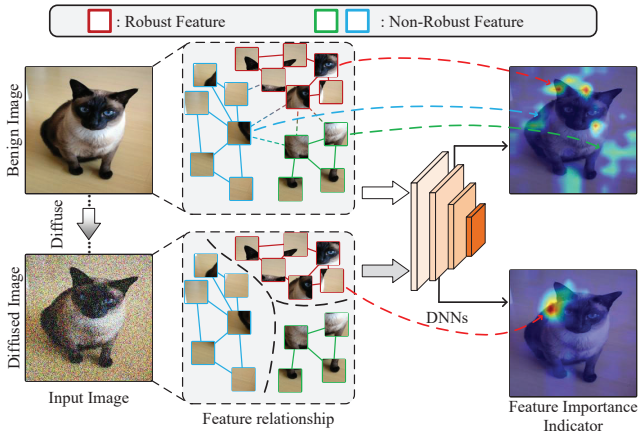


Figure 3: Illustration of separating robust features from non-robust features using diffusion-style noising in DGIE. Red represents robust features, while blue and green represent non-robust features.

3.3 Hierarchical Feature Pyramid Attack

To address the structural overfitting caused by single-layer reliance, HFPA uses the diffusion-guided indicators to jointly suppress important features across multiple layers, as shown in Fig. 2(b). In addition, because transferable guidance may still suffer from sparse and slightly shifted activations across architectures, HFPA further incorporates a spatial spread operation before multi-layer aggregation.

3.3.1 Spatial Spread for Sparsity and Feature Misalignment. In practice, intermediate activations are often sparse due to rectifier nonlinearities [14], and salient spatial responses may not be perfectly aligned across architectures even when representations are related [29, 76]. Such sparsity and slight spatial misalignment can cause the guidance derived from a surrogate model to overfit to model-specific peaks and become less reliable for transfer. To alleviate this issue, we perform a spatial spread operation by convolving each diffusion-guided importance indicator with a fixed kernel $W_p \in \mathbb{R}^{s_w \times s_w}$ which can be written as:

$$W_p[i,j] = \begin{cases} 0, & i = \lfloor s_w/2 \rfloor, j = \lfloor s_w/2 \rfloor \\ 1/\left(s_w^2 - 1\right). & \text{else} \end{cases} \tag{6}$$

Then we can preserve original salient responses via an element-wise maximum:

$$\tilde{\Delta}_k = W_p * \bar{\Delta}_k, \qquad \hat{\Delta}_k^{(i,j)} = \max\big(\bar{\Delta}_k^{(i,j)}, \tilde{\Delta}_k^{(i,j)}\big). \tag{7}$$

This spread makes the guidance less sensitive to sparse, slightly shifted activations, thereby improving cross-model robustness of the importance signal.

3.3.2 Multi-Layer Feature Pyramid. Deep networks form an inherent feature hierarchy: shallow layers capture fine spatial patterns while deeper layers encode higher-level semantics. Leveraging multiple layers therefore provides complementary supervision signals and improves black-box transferability for feature-space attacks compared to single-layer constraints [22, 25]. Based on the diffusion-guided importance $\{\bar{\Delta}_k\}_{k\in\mathcal{K}}$ from DGIE, HFPA constructs an importance pyramid over the selected layer set

$$\mathcal{K} = \{k_1, \ldots, k_M\}, \qquad M = |\mathcal{K}|. \tag{8}$$

For each layer $k \in \mathcal{K}$, we apply the same spatial spread in Eq. (7) to obtain $\hat{\Delta}_k$, and use them to jointly suppress the corresponding weighted features.

3.3.3 Multi-Layer Suppression Objective. To avoid layer-scale dominance, we normalize $\hat{\Delta}_k$ and use a size-normalized uniform aggregation:

$$\begin{aligned} \hat{\Delta}_k &\leftarrow \hat{\Delta}_k / \|\hat{\Delta}_k\|_2, \\ \mathcal{L}_{\mathrm{DIFPA}}(x) &= \frac{1}{M} \sum_{k\in\mathcal{K}} \frac{1}{|F_k|} \left\langle \hat{\Delta}_k,\ F_k(x) \right\rangle_F, \end{aligned} \tag{9}$$

where $|F_k|$ is the number of elements in the feature map, and $\langle\cdot,\cdot\rangle_F$ is the Frobenius inner product. HFPA then crafts adversarial examples by minimizing $\mathcal{L}_{\mathrm{DIFPA}}(x)$ under an ℓ_∞ constraint using projected gradient descent:

$$x^* = \arg\min_{\|\delta\|_\infty \le \epsilon} \mathcal{L}_{\mathrm{DIFPA}}(x). \tag{10}$$

3.4 Attack Algorithm

DIFPA integrates the Diffusion-Guided Importance Estimation (DGIE) module and the Hierarchical Feature Pyramid Attack (HFPA) module to construct a transferable feature-importance pyramid. Based on this pyramid, it iteratively updates the adversarial example x^* under the ℓ_∞ constraint. The complete procedure is summarized in Algorithm 1.

Algorithm 1: Diffusion-based Importance-aware Feature Pyramid Attack (DIFPA)

Input : A surrogate model f; a benign image x and ground-truth label y; perturbation budget ϵ; iterations T; momentum factor β; total diffusion steps J; ensemble number N; selected layer set $\mathcal{K}$ (thus $M = |\mathcal{K}|$).

Output : An adversarial example x^*.

Let $x_0^* = x$, $g_0 = 0$, $\alpha = \epsilon/T$;
Initialize $\bar{\Delta}_k = 0$ for all $k \in \mathcal{K}$;
for $n = 1$ *to* N **do**
  Generate diffused image $x_J^{(n)}$ by Eq. (4);
  foreach $k \in \mathcal{K}$ **do**
    $\Delta_k^{(n)} = \frac{\partial f(x_J^{(n)}, y)}{\partial F_k(x_J^{(n)})}$;
    $\bar{\Delta}_k \leftarrow \bar{\Delta}_k + \Delta_k^{(n)}$;
$\bar{\Delta}_k \leftarrow \bar{\Delta}_k / N$ for all $k \in \mathcal{K}$;
Obtain $\{\hat{\Delta}_k\}_{k\in\mathcal{K}}$ using Eq. (6)–(7);
$\hat{\Delta}_k \leftarrow \hat{\Delta}_k / \|\hat{\Delta}_k\|_2$ for all $k \in \mathcal{K}$;
for $t = 0$ *to* $T-1$ **do**
  $\mathcal{L}_{\mathrm{DIFPA}}(x_t^*) = \frac{1}{M}\sum_{k\in\mathcal{K}} \frac{1}{|F_k|} \left\langle \hat{\Delta}_k, F_k(x_t^*) \right\rangle_F$;
  $g_{t+1} = \beta \cdot g_t + \frac{\nabla_x \mathcal{L}_{\mathrm{DIFPA}}(x_t^*)}{\|\nabla_x \mathcal{L}_{\mathrm{DIFPA}}(x_t^*)\|_1}$;
  $x_{t+1}^* = \mathrm{Clip}_{x,\epsilon}\{x_t^* - \alpha \cdot \mathrm{sign}(g_{t+1})\}$;
return x_T^*;

4 Experiments

4.1 Experiment Setup

4.1.1 Datasets. Following prior work [69], we evaluate on 1,000 ImageNet validation images spanning all classes. We report attack success rate (ASR), the fraction of initially correctly classified images misclassified after attack.

4.1.2 Models. We adopt ten normally pre-trained models, including VGG-19 [49], ResNet-18/50 (Res-18/50) [18], ResNext-50 (ResN-50) [70], Inception-V3/V4 (Inc-V3/V4) [52], ConvNext-T (ConvN-T) [39], ViT-B [10], DeiT-B [54], and Swin-T [37]. For evaluation under various defenses, we consider advanced defense methods, including Bit-R [73], R&P [68], FD [38], HGD [34], JPEG [16], NRP [42], RS [26] and DiffPure [43]. Also, we consider two adversarially trained models [55], including Adv-Inc-v3 and Adv-Inc-v2, to further validate the effectiveness of each method. All pre-trained weights are obtained from the TIMM library [67].

4.1.3 Baselines. We compare DIFPA with seven feature-level attacks: FIA [65], NAA [78], RPA [79], ILPD [31], DANAA [27], BFA [57], and P2FA [36]. We also include DI^2-FGSM [69], TI-FGSM [9], and NI-FGSM [35].

4.1.4 Implementation Details. Unless stated otherwise, all attacks use an ℓ_∞ budget $\epsilon = 16/255$, $T = 10$ iterations, and step size $\alpha = \epsilon/T$ [65]. We use official implementations and default method-specific settings. For DIFPA, we set $m = 0$, $\sigma = 0.06$, $J = 6$, $N = 30$, and use five selected intermediate layers. These settings are fixed across all black-box victim models, without using target-model information or query feedback. Experiments run in PyTorch on an Intel Xeon Silver 4210 CPU and an NVIDIA RTX A6000 GPU.

4.2 Main Results

In this section, we compare DIFPA with baseline attacks on normally trained models, defense models, and different attack backbones using ASR as the evaluation metric.

4.2.1 Attacking Normally Trained Models. To evaluate the effectiveness of the proposed DIFPA, we compare it with existing methods in a black-box setting. The experimental results are summarized in Table 1. As shown, DIFPA significantly outperforms previous SOTA methods, including input-transformation-based, optimization-based,

Table 1: Attack success rates (%) of our proposed DIFPA with eleven SOTA adversarial attack methods. The rightmost column represents the average results on the black-box models. The data marked with an asterisk indicates the result of the white-box attack, and the bolded data denotes the best results under the same setting. Δ represents the performance improvement compared to the previous best-performing SOTA method.

Surrogate Models	Attacks	Victim Models										Avg. (Δ)
		VGG-19	Res-18	Res-50	ResN-50	Inc-V3	Inc-V4	ConvN-T	ViT-B	DeiT-B	Swin-T	
Res-18	DI^2-FGSM	81.30	**100.00***	72.00	66.40	68.90	68.70	53.40	27.40	36.10	53.60	58.64
	TI-FGSM	66.90	**100.00***	41.70	37.90	46.80	43.40	34.10	13.70	14.60	36.10	37.24
	NI-FGSM	65.80	**100.00***	51.40	44.50	49.80	42.80	32.20	14.20	16.90	36.70	39.37
	VMI-FGSM	78.30	99.90*	68.50	63.70	64.80	62.20	49.00	26.60	34.20	53.50	55.64
	FIA	65.90	98.60*	44.20	36.20	41.90	42.40	25.00	8.50	11.50	33.10	34.30
	NAA	70.40	98.40*	56.70	52.60	55.70	52.20	38.50	18.70	25.10	48.10	46.44
	RPA	90.60	**100.00***	82.30	78.50	79.60	78.30	56.50	29.30	40.10	63.80	66.56
	ILPD	90.80	99.90*	86.10	81.50	81.90	80.40	65.80	36.70	49.30	68.70	71.24
	DANAA	88.40	99.90*	77.30	70.80	73.50	73.00	46.50	19.70	28.90	53.80	59.10
	BFA	94.30	**100.00***	87.50	83.40	83.20	83.90	66.70	32.19	49.50	68.19	72.10
	P2FA	88.00	**100.00***	72.00	67.00	61.20	62.00	36.29	11.50	18.50	43.70	51.13
	DIFPA (Ours)	**94.40**	99.90*	**89.60**	**86.60**	**91.00**	**88.90**	**77.10**	**52.10**	**64.90**	**77.80**	**80.27 (+8.17)**
Inc-V4	FIA	67.00	60.50	47.60	44.40	70.90	96.90*	43.40	17.10	22.00	45.40	46.48
	NAA	62.70	56.10	48.90	46.10	66.90	91.70*	44.10	22.10	25.50	43.60	46.22
	RPA	82.10	77.90	70.00	67.90	86.10	99.20*	62.70	31.30	41.00	59.50	64.28
	ILPD	82.20	80.60	72.90	73.20	86.70	97.00*	67.00	41.90	48.00	62.60	68.34
	DANAA	79.60	73.70	63.10	61.90	83.50	98.90*	59.30	27.60	35.40	54.30	59.82
	BFA	82.80	80.90	73.20	73.30	91.90	99.90*	68.90	42.30	50.50	62.60	69.60
	P2FA	77.20	67.60	53.60	51.80	66.90	96.10*	35.60	13.10	20.00	40.00	47.31
	DIFPA (Ours)	**85.10**	**84.60**	**79.10**	**77.70**	**94.70**	**99.90***	**73.10**	**49.10**	**58.10**	**71.80**	**74.81 (+5.21)**
DeiT-B	FIA	87.10	85.40	78.80	78.10	85.70	85.10	78.90	87.60	99.00*	90.00	84.08
	NAA	75.20	72.00	68.70	65.60	72.20	70.10	63.80	73.10	89.30*	76.40	70.79
	RPA	83.80	84.90	77.60	75.50	81.40	80.20	75.50	85.50	98.30*	86.70	81.23
	ILPD	85.80	84.40	81.20	82.80	82.50	82.90	85.50	92.10	99.30*	88.60	85.09
	DANAA	85.20	84.50	77.70	74.30	83.70	82.10	76.30	86.60	98.30*	86.70	81.90
	BFA	85.50	86.80	83.10	82.40	81.30	83.80	81.90	92.80	**100.00***	89.30	85.21
	P2FA	90.20	87.60	83.10	80.50	81.90	82.50	80.10	92.50	**100.00***	91.60	85.56
	DIFPA (Ours)	**93.70**	**92.30**	**92.10**	**91.50**	**89.50**	**89.00**	**90.80**	**97.40**	99.80*	**98.10**	**92.71 (+7.15)**

and diverse feature-level attacks. Specifically, when using Res-18, Inc-V4, and DeiT-B as the surrogate model, our method improves the average black-box attack success rate by 8.17%, 5.21% and 7.15%, respectively, compared with the previously best-performing method. When using Res-18 as the surrogate model, DIFPA surpasses FIA, NAA, and RPA by 45.97%, 33.83%, and 13.71%, respectively, and still improves over recent strong baselines BFA and P2FA by 8.17% and 29.14%. These gains support that diffusion-guided importance estimation and hierarchical multi-layer suppression both improve transferability.

To further validate the superiority of our proposed DIFPA, we combine our method and other SOTA methods with different base attacks to compare the transferability performance, and the experimental results are shown in Table 2. Results indicate that, when using different base attack methods, DIFPA consistently outperforms previous SOTA methods. Specifically, when based on NI-FGSM, our method achieves a 5.49% improvement over the previous best method. This indicates that our method can be readily combined with a broad range of base attack methods, highlighting both its effectiveness and strong compatibility.

4.2.2 Attacking Defense Methods. To further evaluate the robustness of DIFPA under protected settings, we test it against ten defense models, including eight input-transformation-based defenses and two adversarially trained models. As shown in Table 3, DIFPA consistently achieves the best attack success rate across all evaluated defenses. In particular, it attains an average ASR of 76.75%, outperforming the previous best baseline BFA (68.83%) by 7.92%. This advantage is consistent not only on input-transformation-based defenses such as HGD, JPEG, R&P, RS, and DiffPure, but also on adversarially trained models, where DIFPA reaches 79.90% on Adv-Inc-v3 and 60.60% on Adv-Inc-v2. These results demonstrate that diffusion-guided importance estimation and hierarchical multi-layer suppression produce more robust and transferable perturbations, allowing DIFPA to better survive diverse defense mechanisms.

4.2.3 Visualization Comparison. The visualization comparison of the proposed DIFPA and other SOTA methods is shown in Fig. 4. As illustrated in Fig. 4(a), the black-box model accurately classifies benign images, with its regions of interest (ROIs) correctly located. In contrast, the adversarial examples generated by other attack methods, shown in Fig. 4(b)-(d), exhibit shifted and reduced ROIs. As a comparison, in Fig. 4(e), the proposed DIFPA significantly

Table 2: Attack success rates (%) of our proposed DIFPA with seven SOTA feature-level adversarial attack methods using different base attacks. The data marked with an asterisk indicates the result of the white-box attack, and the bolded data denotes the best results under the same setting. Δ represents the performance improvement compared to the previous best-performing SOTA method.

Base Attacks	Attacks	Victim Models										Avg. (Δ)
		VGG-19	Res-18	Res-50	ResN-50	Inc-V3	Inc-V4	ConvN-T	ViT-B	DeiT-B	Swin-T	
NI-FGSM	FIA	65.30	98.10*	42.50	36.30	43.20	41.10	26.60	8.10	11.40	32.00	34.06
	NAA	70.90	98.10*	57.30	53.40	55.80	52.50	39.00	18.70	25.50	48.20	46.81
	RPA	71.60	95.20*	41.40	35.20	51.60	41.60	29.80	13.30	17.40	34.90	37.42
	ILPD	82.70	99.80*	64.50	59.60	66.40	66.90	47.00	28.10	35.30	50.50	55.67
	DANAA	82.00	99.80*	66.10	58.20	70.10	65.40	44.70	19.40	28.80	51.80	54.06
	BFA	81.00	**100.00***	62.20	57.80	65.30	66.30	46.30	33.20	37.50	51.70	55.70
	P2FA	45.20	40.40*	18.70	14.40	30.00	23.50	14.10	6.30	6.40	17.00	19.51
	DIFPA (Ours)	**84.90**	99.00*	**69.30**	**66.30**	**74.70**	**69.10**	**54.10**	**36.00**	**38.70**	**57.60**	**61.19 (+5.49)**

Table 3: Average attack success rates (%) of our proposed DIFPA with seven SOTA feature-level adversarial attack methods under ten different adversarial defense methods. The rightmost column represents the average results under all defense methods. Bolded data denotes the best results under the same setting. Δ represents the performance improvement compared to the previous best-performing SOTA method.

Attack	No Defense	Input-Transform-Based Adversarial Defense								Adversarial Trained Model		Avg. (Δ)
		Bit-R	FD	HGD	JPEG	NRP	R&P	RS	DiffPure	Adv-Inc-v3	Adv-Inc-v2	
FIA	34.30	39.42	34.44	40.52	40.86	37.15	38.87	40.79	44.50	37.79	22.29	37.66
NAA	46.44	46.17	42.40	51.91	51.96	45.87	50.69	42.35	44.70	42.30	27.10	44.55
RPA	66.56	63.16	61.56	69.57	69.69	60.01	68.87	56.17	52.04	61.90	43.40	60.64
ILPD	71.24	65.99	65.56	73.88	73.82	62.48	73.01	57.24	54.16	66.30	49.00	64.14
DANAA	59.10	58.50	56.40	63.38	63.34	54.84	62.39	53.47	51.03	54.40	34.00	55.18
BFA	72.10	71.50	69.70	75.40	75.30	70.10	74.80	65.90	63.80	66.40	55.40	68.83
P2FA	51.13	50.15	48.82	52.14	52.27	50.17	52.16	47.56	43.28	47.87	33.10	47.75
DIFPA (Ours)	**80.27**	**76.79**	**77.05**	**82.11**	**82.21**	**73.43**	**82.40**	**74.50**	**78.55**	**79.90**	**60.60**	**76.75 (+7.92)**

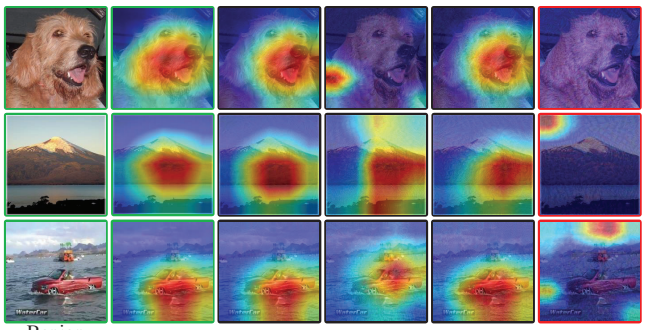


Figure 4: Visualization comparison of ROIs on Inc-V4 for adversarial examples crafted from Res-18. Columns (a)–(e) show the victim-model ROIs for the benign image and the adversarial examples generated by NAA, RPA, ILPD, and DIFPA, respectively. Red indicates high attention and blue indicates low attention.

disrupts the ROIs of the black-box model, causing more substantial shifts and reductions. This observation validates that DIFPA effectively disrupts the semantic consistency of shared features across disparate architectures, thereby achieving superior cross-model transferability.

4.2.4 Attack Efficiency Comparison. The attack efficiency comparison of the proposed DIFPA and other SOTA methods is shown in Fig. 6. As illustrated in Fig. 6, DIFPA achieves significant performance improvements compared to FIA, NAA, DANAA, and BFA, while maintaining similar time costs, with performance doubling relative to FIA and NAA. For ILPD, RPA, and P2FA, which employ more complex attack strategies and thus incur additional time costs, the performance remains inferior to that of DIFPA. This demonstrates the superiority of DIFPA, which improves the reliability of feature-importance estimation and exploits complementary multi-layer information within a moderate additional cost, thereby achieving a better effectiveness-efficiency trade-off. The results highlight the advantages of DIFPA in terms of attack efficiency. Specifically, DGIE and HFPA improve the reliability of feature-importance estimation and exploit complementary multi-layer information with only moderate overhead. As a result, DIFPA achieves consistently higher ASR and a better effectiveness-efficiency trade-off than strong baselines.

4.3 Ablation Studies

We conduct ablations to evaluate the key components of DIFPA.

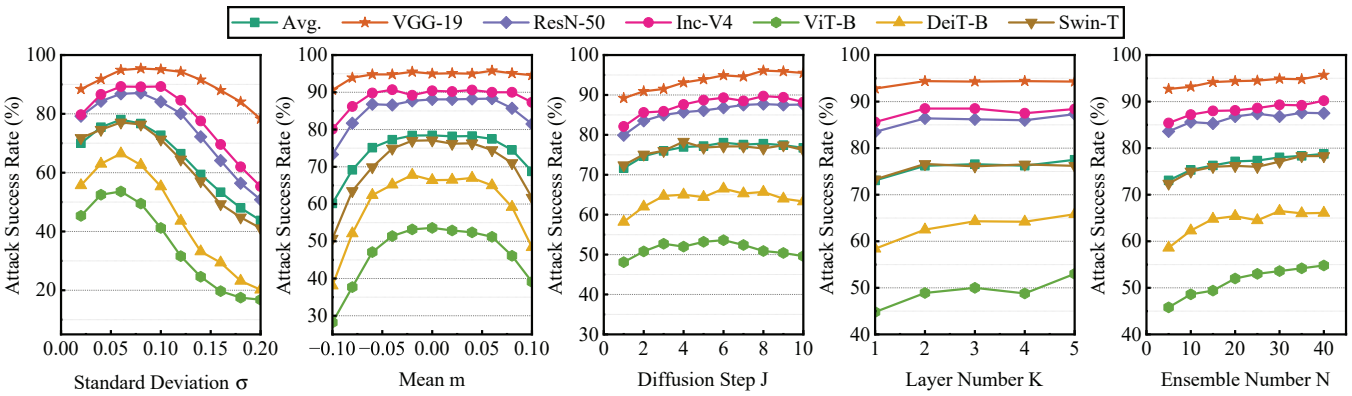


Figure 5: Ablation of DIFPA hyper-parameters. Using Res-18 as the surrogate model, we vary σ, m, J, K, and N while keeping other settings fixed, and report the ASR on six black-box victim models.

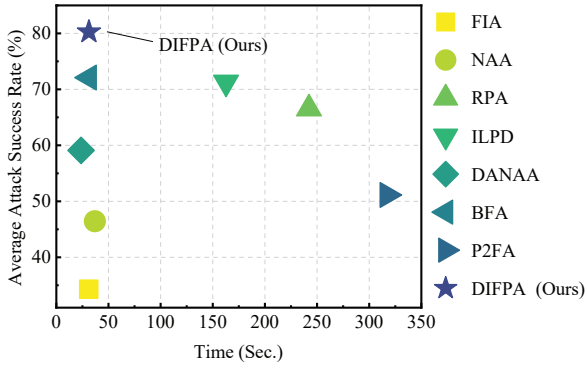


Figure 6: Attack efficiency comparison. The y-axis denotes the average ASR on black-box models and the x-axis denotes the average runtime per image. Points closer to the top-left indicate better effectiveness-efficiency trade-off.

Table 4: Average attack success rate (%) of ablation studies. The black-box models used are consistent with the previous experiment, and the white-box model is Res-18. Δ denotes the absolute ASR rise compared with the baseline.

HFPA	DGIE	Avg. (%) ↑	Δ
✗	✗	54.18	-
✓	✗	56.01	1.83
✗	✓	67.27	13.09
✓	✓	80.27	26.09

4.3.1 Ablation Study of DGIE and HFPA. We first verify the individual and joint contributions of DGIE and HFPA by enabling each module separately on top of the baseline feature-level attack, as shown in Table 4. Compared with the baseline (54.18%), enabling HFPA alone yields a modest gain of 1.83%, while enabling DGIE alone brings a much larger improvement of 13.09%, indicating that diffusion-guided importance estimation is the primary source of the performance gain. When both modules are activated, the average ASR further rises to 80.27%, achieving a total improvement of 26.09% over the baseline. These results show that DGIE and HFPA are complementary, and their combination yields the strongest transferability.

4.3.2 Ablation Studies of Hyper-Parameters. Figure 5 analyzes five hyperparameters: standard deviation σ, mean m in Eq. (4), diffusion steps J, layer number K in Eq. (9), and ensemble number N in Eq. (5). In general, increasing the standard deviation σ and the number of diffusion steps J introduces more diffusion noise, thereby better decoupling relevant from irrelevant features. However, setting these parameters too high will distort useful information in the image, leading to a decrease in final transferability performance. Based on this observation, we set $\sigma = 0.06$ and $J = 6$. For the mean m, setting it too high or too low will affect the distribution of the added diffuse noise. Therefore, the best transferability performance is achieved when $m = 0$. As the layer number K and the ensemble number N increase, the final transferability performance increases as K and N increase. This suggests that our method exploits more transferable information in more intermediate layers and produces more stable importance estimates when more samples are used for averaging. Following previous works, we set $K = 5$ and $N = 30$.

5 Conclusion

In this paper, we propose DIFPA, a transferable attack that combines diffusion-guided importance estimation with a hierarchical feature pyramid. The key idea is that transferability depends not only on stronger optimization, but also on converting single-point, single-layer feature attacks into neighborhood-based and hierarchical feature suppression. DGIE leverages the diffusion forward marginal to reduce indicator fluctuation and produce more reliable layer-wise cues. Based on them, HFPA aggregates cross-layer information and mitigates spatial mismatch via spatial spreading, enabling more effective suppression of transferable intrinsic features. Extensive experiments show that DIFPA consistently outperforms strong baselines across diverse black-box models, improving the average transfer performance by up to 8.17% with Res-18 as the surrogate model.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grants U2441239, 62172243, 62402425, 62402418, and U21A20471); the Fujian Provincial Industrial Technology Development and Application Program (Grant 2025H0043); the China Postdoctoral Science Foundation (Grant 2024M762829); the Zhejiang Provincial Natural Science Foundation (Grant LD24F020002); the "Pioneer and Leading Goose" R&D Program of Zhejiang (Grants 2025C01082, 2025C02033, and 2025C02263); and the Zhejiang Provincial Priority-Funded Postdoctoral Research Project (Grant ZJ2024001).

References

[1] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2022. Label-Efficient Semantic Segmentation with Diffusion Models. In *ICLR*.

[2] Sebastian Bordt, Uddeshya Upadhyay, Zeynep Akata, and Ulrike von Luxburg. 2023. The Manifold Hypothesis for Gradient-Based Explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 3697–3702.

[3] Bin Chen, Jiali Yin, Shukai Chen, Bohao Chen, and Ximeng Liu. 2023. An Adaptive Model Ensemble Adversarial Attack for Boosting Adversarial Transferability. In *ICCV*. 4489–4498.

[4] Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. 2025. Diffusion Models for Imperceptible and Transferable Adversarial Attack. *PAMI* 47, 2 (2025), 961–977. doi:10.1109/TPAMI.2024.3480519

[5] Yong Chen, Xuedong Li, Peng Hu, Dezhong Peng, and Xu Wang. 2024. DifFilter: Defending Against Adversarial Perturbations With Diffusion Filter. *TIFS* 19 (2024), 6779–6794. doi:10.1109/TIFS.2024.3422923

[6] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. 2024. YOLO-World: Real-Time Open-Vocabulary Object Detection. In *CVPR*. 16901–16911.

[7] Xuelong Dai, Kaisheng Liang, and Bin Xiao. 2025. AdvDiff: Generating Unrestricted Adversarial Examples Using Diffusion Models. In *ECCV*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 93–109.

[8] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting Adversarial Attacks With Momentum. In *CVPR*. 9185–9193.

[9] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. 2019. Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks. In *CVPR*. 4312–4321.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*. https://openreview.net/forum?id=YicbFdNTTy

[11] Xia Du, Xiaoyuan Liu, Jizhe Zhou, Zheng Lin, Chi-man Pun, Cong Wu, Tao Li, Zhe Chen, Wei Ni, and Jun Luo. 2026. Defensive Adversarial CAPTCHA: A Semantics-Driven Framework for Natural Adversarial Example Generation. *TDSC* 23, 2 (2026), 3423–3435. doi:10.1109/TDSC.2025.3636741

[12] Mingyuan Fan, Xiaodan Li, Cen Chen, and Yinggui Wang. 2024. Flatness-aware Adversarial Attack.

[13] Fuquan Gan and Yan Wo. 2025. Boosting the Transferability of Adversarial Examples Through Gradient Aggregation. *TIFS* 20 (2025), 5563–5576. doi:10.1109/TIFS.2025.3574989

[14] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep Sparse Rectifier Neural Networks. In *AISTATS (Proceedings of Machine Learning Research, Vol. 15)*, Geoffrey Gordon, David Dunson, and Miroslav Dudík (Eds.). PMLR, Fort Lauderdale, FL, USA, 315–323.

[15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572 [stat.ML]

[16] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. 2018. Countering Adversarial Images using Input Transformations. In *ICLR*.

[17] Yiwen Guo, Qizhang Li, Wangmeng Zuo, and Hao Chen. 2023. An Intermediate-Level Attack Framework on the Basis of Linear Regression. *PAMI* 45, 3 (2023), 2726–2735.

[18] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *NIPS*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 6840–6851.

[20] Zhaozhe Hu, Bin Chen, Jia-Li Yin, Bo-Hao Chen, Yaguan Qian, and Shouling Ji. 2026. Mix2Aug: Revisiting Mixing-Based Augmentations for Improving Robust Generalization of Adversarial Training. *TDSC* 23, 2 (2026), 3036–3050. doi:10.1109/TDSC.2025.3632109

[21] Zhaozhe Hu, Jia-Li Yin, Bin Chen, Luojun Lin, Bo-Hao Chen, and Ximeng Liu. 2024. MEAT: Median-Ensemble Adversarial Training for Improving Robustness and Generalization. In *ICASSP*. 5600–5604.

[22] Qian Huang, Isay Katsman, Zeqi Gu, Horace He, Serge Belongie, and Ser-Nam Lim. 2019. Enhancing Adversarial Example Transferability With an Intermediate Level Attack. In *ICCV*. 4732–4741. doi:10.1109/ICCV.2019.00483

[23] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box Adversarial Attacks with Limited Queries and Information. In *ICML (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 2137–2146.

[24] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial Examples Are Not Bugs, They Are Features. In *NIPS*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc., 125–136. https://proceedings.neurips.cc/paper_files/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf

[25] Nathan Inkawhich, Kevin Liang, Binghui Wang, Matthew Inkawhich, Lawrence Carin, and Yiran Chen. 2020. Perturbing Across the Feature Hierarchy to Improve Standard and Strict Blackbox Attack Transferability. In *NIPS*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 20791–20801.

[26] Jinyuan Jia, Xiaoyu Cao, Binghui Wang, and Neil Zhenqiang Gong. 2020. Certified Robustness for Top-k Predictions against Adversarial Perturbations via Randomized Smoothing. In *ICLR*.

[27] Zhibo Jin, Zhiyu Zhu, Xinyi Wang, Jiayu Zhang, Jun Shen, and Huaming Chen. 2023. DANAA: Towards Transferable Attacks with Double Adversarial Neuron Attribution. In *Advanced Data Mining and Applications*. Springer Nature Switzerland, Cham, 456–470.

[28] Caixin Kang, Yinpeng Dong, Zhengyi Wang, Shouwei Ruan, Yubo Chen, Hang Su, and Xingxing Wei. 2025. DIFFender: Diffusion-Based Adversarial Defense Against Patch Attacks. In *ECCV*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 130–147.

[29] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of Neural Network Representations Revisited. In *ICML (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 3519–3529.

[30] Jian-Wei Li, Wen-Ze Shao, Yu-Bao Sun, Li-Qian Wang, Qi Ge, and Liang Xiao. 2025. Boosting Adversarial Transferability via Relative Feature Importance-Aware Attacks. *TIFS* 20 (2025), 3489–3504. doi:10.1109/TIFS.2025.3552030

[31] Qizhang Li, Yiwen Guo, Wangmeng Zuo, and Hao Chen. 2023. Improving Adversarial Transferability via Intermediate-level Perturbation Decay. In *NIPS*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 32900–32912. doi:10.52202/075280-1425

[32] Zhiwei Li, Min Ren, Qi Li, Fangling Jiang, and Zhenan Sun. 2024. Improving Transferability of Adversarial Samples via Critical Region-Oriented Feature-Level Attack. *TIFS* 19 (2024), 6650–6664.

[33] Zhankai Li, Weiping Wang, Jie Li, Kai Chen, and Shigeng Zhang. 2024. Foolmix: Strengthen the Transferability of Adversarial Examples by Dual-Blending and Direction Update Strategy. *TIFS* 19 (2024), 5286–5300. doi:10.1109/TIFS.2024.3393745

[34] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. 2018. Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser. In *CVPR*. 1778–1787.

[35] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. 2020. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. In *International Conference on Learning Representations*.

[36] Renpu Liu, Hao Wu, Jiawei Zhang, Xin Cheng, Xiangyang Luo, Bin Ma, and Jinwei Wang. 2025. Pixel2Feature Attack (P2FA): Rethinking the Perturbed Space to Enhance Adversarial Transferability. In *ICML (Proceedings of Machine Learning Research, Vol. 267)*, Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR, 39853–39870. https://proceedings.mlr.press/v267/liu25cd.html

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *ICCV*. 10012–10022.

[38] Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. 2019. Feature Distillation: DNN-Oriented JPEG Compression Against Adversarial Examples. In *CVPR*. 860–868.

[39] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. In *CVPR*. 11976–11986.

[40] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. 2022. Frequency Domain Model Augmentation for Adversarial Attack. In *ECCV*. Springer Nature Switzerland, Cham, 549–566.

[41] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*.

[42] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. 2020. A Self-supervised Approach for Adversarial Robustness. In *CVPR*. 259–268.

[43] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. 2022. Diffusion Models for Adversarial Purification. In *ICML (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 16805–16827.

[44] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. arXiv:1605.07277 [cs.CR]

[45] Yifan Pu, Weicong Liang, Yiduo Hao, YUHUI YUAN, Yukang Yang, Chao Zhang, Han Hu, and Gao Huang. 2023. Rank-DETR for High Quality Object Detection. In *NIPS*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 16100–16113.

[46] Yaguan Qian, Kecheng Chen, Bin Wang, Zhaoquan Gu, Shouling Ji, Wei Wang, and Yanchun Zhang. 2024. Enhancing Transferability of Adversarial Examples Through Mixed-Frequency Inputs. *TIFS* 19 (2024), 7633–7645. doi:10.1109/TIFS.2024.3430508

[47] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2024. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning. In *CVPR*. 8543–8552.

[48] Yucheng Shi, Yahong Han, Qinghua Hu, Yi Yang, and Qi Tian. 2023. Query-Efficient Black-Box Adversarial Attack With Customized Iteration and Sampling. *PAMI* 45, 2 (2023), 2226–2245.

[49] K Simonyan and A Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

[50] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *ICML (Proceedings of Machine Learning Research, Vol. 37)*, Francis Bach and David Blei (Eds.). PMLR, Lille, France, 2256–2265.

[51] David Stutz, Matthias Hein, and Bernt Schiele. 2019. Disentangling Adversarial Robustness and Generalization. In *CVPR*. 6976–6987.

[52] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper With Convolutions. In *CVPR*. 1–9.

[53] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. arXiv:1312.6199 [cs.CV]

[54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. 2021. Training data-efficient image transformers & distillation through attention. In *ICML (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 10347–10357. https://proceedings.mlr.press/v139/touvron21a.html

[55] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2018. Ensemble Adversarial Training: Attacks and Defenses. In *ICLR*.

[56] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. 2019. AutoZOOM: Autoencoder-Based Zeroth Order Optimization Method for Attacking Black-Box Neural Networks. *AAAI* 33, 01 (Jul. 2019), 742–749.

[57] Maoyuan Wang, Jinwei Wang, Bin Ma, and Xiangyang Luo. 2024. Improving the transferability of adversarial examples through black-box feature attacks. *Neurocomputing* 595 (2024), 127863.

[58] Ping Wang, Haichang Gao, Xiaoyan Guo, Zhongni Yuan, and Jiawei Nian. 2024. Improving the Security of Audio CAPTCHAs With Adversarial Examples. *TDSC* 21, 2 (2024), 650–667. doi:10.1109/TDSC.2023.3236367

[59] Xiaosen Wang and Kun He. 2021. Enhancing the Transferability of Adversarial Attacks Through Variance Tuning. In *CVPR*. 1924–1933.

[60] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. 2021. Admix: Enhancing the Transferability of Adversarial Attacks. In *ICCV*. 16158–16167.

[61] Xiaosen Wang, Zeliang Zhang, and Jianping Zhang. 2023. Structure Invariant Transformation for better Adversarial Transferability. In *ICCV*. 4607–4619.

[62] Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. 2023. InfoDiffusion: Representation Learning Using Information Maximizing Diffusion Models. In *ICML (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, 36336–36354.

[63] Yajie Wang, Yi Wu, Shangbo Wu, Ximeng Liu, Wanlei Zhou, Liehuang Zhu, and Chuan Zhang. 2024. Boosting the Transferability of Adversarial Attacks With Frequency-Aware Perturbation. *TIFS* 19 (2024), 6293–6304. doi:10.1109/TIFS.2024.3411921

[64] Yajie Wang, Chuan Zhang, Huipeng Zhou, Zuobin Ying, Zehui Xiong, Wanlei Zhou, and Liehuang Zhu. 2026. Boosting Adversarial Transferability of Vision Transformers. *TDSC* 23, 1 (2026), 329–342. doi:10.1109/TDSC.2025.3605453

[65] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. 2021. Feature Importance-Aware Transferable Adversarial Attacks. In *ICCV*. 7639–7648.

[66] Juanjuan Weng, Zhiming Luo, and Shaozi Li. 2025. Improving Transferable Targeted Adversarial Attack via Normalized Logit Calibration and Truncated Feature Mixing. *TIFS* 20 (2025), 4595–4609. doi:10.1109/TIFS.2025.3563820

[67] Ross Wightman. 2019. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models. doi:10.5281/zenodo.4414861

[68] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. 2018. Mitigating Adversarial Effects Through Randomization. In *ICLR*.

[69] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. 2019. Improving Transferability of Adversarial Examples With Input Diversity. In *CVPR*. 2730–2739.

[70] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *CVPR*. 1492–1500.

[71] Yifeng Xiong, Jiadong Lin, Min Zhang, John E. Hopcroft, and Kun He. 2022. Stochastic Variance Reduced Ensemble Adversarial Attack for Boosting the Adversarial Transferability. In *CVPR*. 14983–14992.

[72] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In *NIPS*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 15903–15935.

[73] Weilin Xu, David Evans, and Yanjun Qi. 2018. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *NDSS*. The Internet Society. https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-4_Xu_paper.pdf

[74] Xingyi Yang and Xinchao Wang. 2023. Diffusion Model as Representation Learner. In *ICCV*. 18938–18949.

[75] Jia-Li Yin, Bin Chen, Wanqing Zhu, Bo-Hao Chen, and Ximeng Liu. 2023. Push Stricter to Decide Better: A Class-Conditional Feature Adaptive Framework for Improving Adversarial Robustness. *TIFS* 18 (2023), 2119–2131.

[76] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *NIPS*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (Eds.), Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2014/file/532a2f85b6977104bc93f8580abbb330-Paper.pdf

[77] Yeliang You, Bin Chen, Jia-li Yin, Ximeng Liu, and Wei Lin. 2025. Towards Adversarial-Robust Class-Incremental Learning via Progressively Volume-Up Perturbation Generation. In *PRCV*, Zhouchen Lin, Ming-Ming Cheng, Ran He, Kurban Ubul, Wushouer Silamu, Hongbin Zha, Jie Zhou, and Cheng-Lin Liu (Eds.). Springer Nature Singapore, Singapore, 61–75.

[78] Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R. Lyu. 2022. Improving Adversarial Transferability via Neuron Attribution-Based Attacks. In *CVPR*. 14993–15002.

[79] Yaoyuan Zhang, Yu-an Tan, Tian Chen, Xinrui Liu, Quanxin Zhang, and Yuanzhang Li. 2022. Enhancing the Transferability of Adversarial Examples with Random Patch. In *IJCAI*, Lud De Raedt (Ed.). 1672–1678. doi:10.24963/ijcai.2022/233 Main Track.

[80] Zeyu Zhao, Ke Xu, Tanfeng Sun, and Xinghao Jiang. 2026. ASGA: Attention-Based Sparse Global Attack to Video Action Recognition. *TDSC* 23, 1 (2026), 1191–1202. doi:10.1109/TDSC.2025.3615158

[81] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. 2018. Transferable Adversarial Perturbations. In *ECCV*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 471–486.