FAST PREEMPTIVE ROBUSTIFICATION: HIGH-FREQUENCY RESPONSE ANTI-ALIGNS SHARED VULNERABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial attacks can readily compromise deep neural networks (DNNs). In particular, transferable attacks (TAs) exploit the shared vulnerabilities among DNNs, enabling perturbations crafted on surrogates to transfer to unseen models. Training-time and post-attack defenses have been extensively studied for combating TAs. Orthogonal to these approaches, preemptive robustification (PR) has emerged as a pre-attack defense that enhances the robustness of benign samples by superimposing protective variations before attacks. Despite its promise, PR remains underexplored and faces several important limitations. First, dependence on well-trained surrogate classifiers limits applicability, as surrogates are task-specific and may even be unavailable in some practical settings. Moreover, the required iterative optimization or dedicated PR generator training incurs substantial costs. Third, the variations generated by existing methods lack human interpretability. To address these, we seek an efficient PR that is surrogate-free, optimization-free, training-free, and human-interpretable. Intriguingly, we discover a numerical correlation between the shared vulnerabilities of DNNs and Laplacian responses, with their cosine similarity being significantly negative. This indicates that negated high-frequency response constitutes an important component of shared vulnerabilities. Consequently, strengthening Laplacian responses counteracts this component, improving resistance to TAs. Building upon this insight, we propose **F**ast **P**reemptive **R**obustification (**FPR**), which runs efficiently on commodity CPUs, enabling real-time PR without dedicated GPUs. Despite its simplicity, FPR provides substantial robustness gains across diverse attack settings. Specifically, FPR reduces the attack success rate (ASR) of untargeted TAs by 12.7% and that of targeted TAs from 10.7% to 4.1%. The code will be released.

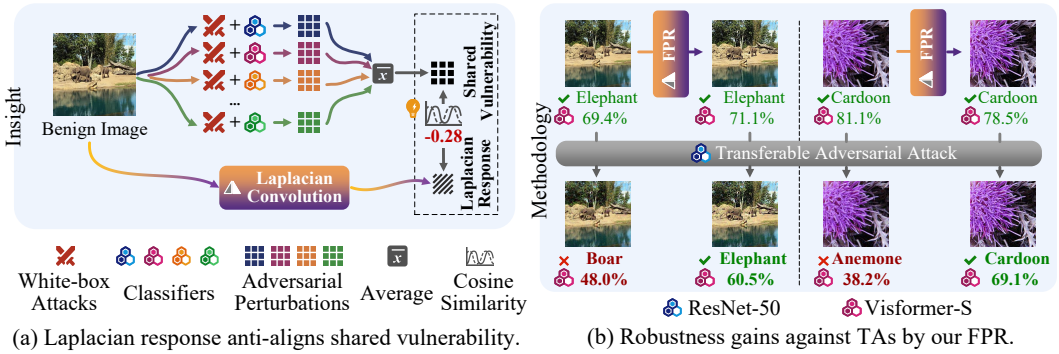
1 INTRODUCTION

> *Poisons and medicine are oftentimes the same substance given with different intents.* — Peter Mere Latham (1789–1875)

Deep neural networks (DNNs) Szegedy et al. (2015); He et al. (2016); Vaswani et al. (2017) have revolutionized numerous applications Kaltenecker et al. (2026); Li et al. (2026). Unfortunately, they remain highly vulnerable to adversarial attacks due to high-dimensional non-robust representations Szegedy et al. (2013); Carlini & Wagner (2017). In particular, transferable attacks (TAs) Goodfellow et al. (2015); Liu et al. (2017); Dong et al. (2018); Xie et al. (2019) approximate shared vulnerabilities across DNNs, enabling adversarial perturbations crafted on surrogate models to effectively transfer to unseen models, thereby hindering reliable DNN deployment even in black-box settings. Mitigating TAs has therefore emerged as a critical challenge in trustworthy AI.

To address this, training-time and test-time defenses have been extensively studied. Training-time defenses include adversarial training Sanyal et al. (2026), robust architecture design Niu et al. (2024), and robustness-oriented regularization Chen et al. (2026). Post-attack defenses mitigate adversarial effects through input transformations Liang et al. (2025), input purification Jiang et al. (2026), or output correction Shanmugam et al. (2025), or identify malevolence via adversarial detection Fares et al. (2026). However, both training-time and post-attack defenses typically incur

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Laplacian response anti-aligns shared vulnerability. (b) Robustness gains against TAs by our FPR.

Figure 1: Illustrations of (a) the numerical relationship between Laplacian response and shared vulnerability, and (b) the robustness gains against TAs induced by our Fast Preemptive Robustification.

substantial training or inference overhead. More importantly, these approaches require significant modifications to the parameters, architectures, or pipelines of original models, which may introduce undesirable changes and considerably compromise clean accuracy.

Orthogonal to training-time and post-attack defenses, recent preemptive robustification (PR) approaches Moon et al. (2022); Frosio & Kautz (2023) introduce a pre-attack paradigm that lies between the two. Specifically, PR superimposes small protective variations onto benign samples, making them more resistent to attacks. Instead of imposing robustness requirements on models, PR shifts the burden to samples. This enables defenders to achieve sample-tailored robustness rather than requiring a single model to maintain robustness across diverse inputs, while providing an alternative defense strategy and alleviating training and inference overhead.

However, as an emerging paradigm, PR remains underexplored and faces several challenges. First, existing PR methods are tightly coupled with well-trained surrogate classifiers, limiting their applicability in scenarios where such surrogates are unavailable and hindering adaptation to new tasks. Moreover, existing methods require iterative optimization or dedicated PR generator training, resulting in substantial computational overhead, particularly for real-time applications. On top of that, protective variations generated by existing methods are opaque to humans, raising concerns about their reliability. Therefore, we seek to develop an efficient PR method that is highly effective against TAs, while being surrogate-free, optimization-free, training-free, and human-interpretable.

To defend against TAs, an ideal protective variation should neutralize the shared vulnerability among classifiers. Alternatively, a relaxed objective is to identify the dominant components contributing to the shared vulnerability and aggregate their negated counterparts as protective variations. Several studies provide clues to one of the key components underlying this shared vulnerability: (i) Extensive studies Yin et al. (2019); Wang et al. (2020); Caro et al. (2020); Pei et al. (2025) have shown that adversarial perturbations often exhibit dominant energy in the high frequency spectrum. (ii) Protective variations generated by previous PR methods Frosio & Kautz (2023) exhibit texture-like patterns, and the robustified samples resemble texture- and contrast-enhanced images. (iii) OTI Liang et al. (2026) reveals that images with stronger textures tend to be more resistant to attacks.

Collectively, these findings lead us to a hypothesis: The negated high-frequency response of an image constitutes a critical component of the shared vulnerability on the image across models. However, prior studies have yet to establish this relationship quantitatively. We provide the first precise empirical characterization of this relationship (see Section 3.2), as shown in Figure 1(a):

The average white-box perturbation (shared vulnerabilities) across multiple classifiers on the same image exhibits a strongly negative cosine similarity with the image's Laplacian response, and its magnitude increases with more classifiers and iterations.

Therefore, we can achieve PR by enhancing the high-frequency response of images to counteract this component of shared vulnerabilities. With an appropriate scale, Laplacian sharpening is not only imperceptible to humans but also achieves strong PR performance. Therefore, we adopt Laplacian sharpening to enhance high-frequency responses as a PR strategy (see Section 3.3). Laplacian

sharpening requires only a single channel-wise 3×3 convolution, making it highly efficient even on commodity CPUs. Therefore, we name this method as **F**ast **P**reemptive **R**obustification (**FPR**).

Extensive experiments (see Section 4) demonstrate the effectiveness of FPR, with representative visual examples in Figure 1(b). FPR has a negligible impact on benign accuracy while substantially improving robustness against TAs. On average, FPR reduces the attack success rate (ASR) of single-surrogate untargeted TAs by 12.7%, ensemble-surrogate untargeted ASR by 14.7%, and targeted ASR from 10.7% to 4.1%. Moreover, FPR is complementary to training-time and test-time defenses. For example, when combined with state-of-the-art adversarial training methods, FPR further reduces the ASR from 12.1% to 10.3%. It is remarkable that this classical image processing technique can take on new relevance in the deep learning era. The main contributions are summarized as follows:

- To the best of our knowledge, this is the first work to reveal a quantitative similarity between an image's high-frequency response and shared model vulnerabilities.
- Based on this similarity, we develop FPR with Laplacian sharpening. FPR is the first surrogate-free, optimization-free, training-free, and human-interpretable PR method.
- Extensive experiments demonstrate that FPR achieves strong robustness gains against diverse TAs while incurring negligible computational costs, enabling real-time PR.

2 Related Work

2.1 Transferable Attacks

In the real world, attackers typically lack access to the architecture and parameters of the victim model, giving rise to the black-box threat model. Two mainstream paradigms, query-based attacks and transferable attacks, have been developed for black-box attacks. Query-based attacks refine perturbations by exploiting the feedback from the victim model. In contrast, transferable attacks enhance the adversarial transferability so that perturbations crafted on surrogate models remain effective against unseen models. Diverse strategies have been proposed to improve transferability, including gradient-based Gan & Wo (2025); Peng et al. (2025), model-related Wang et al. (2025); Liu et al. (2026), transformation-based Liang & Pun (2025); Guo et al. (2025), advanced objective Zheng et al. (2025); Liu et al. (2025b), generative Nakka & Alahi (2025); Li et al. (2025), and ensemble-based Chen et al. (2023b); Tang et al. (2024) approaches.

2.2 Adversarial Defenses

In response to the growing adversarial threat, various defenses have been developed to protect deep learning models. Depending on the deployment stage, these defense strategies can be classified as training-time defenses, post-attack defenses, and recent pre-attack defenses.

Training-time defenses enhance model robustness during the training phase. Adversarial training Zhang et al. (2025); Sanyal et al. (2026) does so by incorporating adversarial examples into the training set. Robust architecture design Peng et al. (2023); Niu et al. (2024) leverages structural designs to mitigate adversarial effects. Robustness-oriented regularization Zühlke & Kudenko (2025); Chen et al. (2026) leverages tailored terms to encourage models to learn more robust representations.

Post-attack defenses alleviate adversarial threats by mitigating or detecting adversarial effects after attacks. Input purification defenses Yang et al. (2024); Lei et al. (2025) adjust adversarial examples toward the benign data distribution via generative methods such as GANs and diffusion models. Input transformation defenses Cohen & Giryes (2024); Liang et al. (2025) transform inputs to disrupt the structure of adversarial perturbations and consequently mitigate their impact on model predictions. Adversarial detection Liu et al. (2024); Kong et al. (2025) prevents malicious access by identifying adversarial effects in the input. Output correction defenses Shanmugam et al. (2025) aggregate multiple prediction copies to produce more reliable outputs.

Pre-attack defenses. Beyond these two stages, there exists another stage after model deployment but before inference, which is responsible for sample storage and transmission. We term defenses operating at this stage as pre-attack defenses. Unlike training-time and post-attack defenses, which primarily improve robustness from a model-centric perspective, pre-attack defenses shift the focus toward enhancing the intrinsic robustness of the samples themselves.

Specifically, preemptive robustification superimposes protective variations onto samples, making them more resistant to adversarial attacks. Unfortunately, existing PR methods remain scarce. Bi-Level Moon et al. (2022) preemptively robustifies images by solving a bi-level optimization problem to find nearby robust points. A^5 Frosio & Kautz (2023) leverages automatic perturbation analysis to craft certified defensive perturbations against bounded attacks. Despite these advances, as discussed above, their limitations hinder practical deployment. Therefore, this work aims to develop an efficient PR method to overcome their limitations.

2.3 Image Sharpening in Deep Learning Security

Traditionally, image sharpening has been widely studied as a means of enhancing perceptual quality. Such approaches range from classical spatial-domain and frequency-domain filtering Gonzalez (2009) to adaptive filtering Schavemaker et al. (2000); Zhang & Allebach (2008) methods, and more recently, DNN-based sharpening techniques Xu et al. (2025). In recent years, image sharpening has found renewed applications in the deep learning era. It has been explored as an augmentation strategy to improve model generalization and robustness He et al. (2023b;a); Heinert et al. (2024), and as a post-attack defense to mitigate adversarial effects Mustafa et al. (2019); He et al. (2023a). These studies suggest that image sharpening, beyond its traditional role in perceptual enhancement, can also contribute to deep learning security. In this work, we revisit image sharpening from the pre-attack defense perspective and uncover its potential for PR.

3 Theory and Methodology

3.1 Preliminaries

Transferable Attacks (TA). Given a benign sample set $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^N$, a TA $A \in \mathcal{A}$ crafts highly transferable perturbations on a surrogate f_s, yielding the adversarial example set:

$$\mathcal{X}_\mathrm{adv} = \{\boldsymbol{x} + \boldsymbol{\delta}_\mathrm{adv} \mid \boldsymbol{\delta}_\mathrm{adv} = A(\boldsymbol{x}; f_\mathrm{s}), \boldsymbol{x} \in \mathcal{X}\}, \tag{1}$$

such that unseen classifiers $\mathcal{F} = \{f_j\}_{j=1}^N$ can be severely misled, and the attack success rate (ASR) are calculated by:

$$\mathrm{ASR}(\mathcal{X}, A, f_\mathrm{s}, \mathcal{F}) = \mathbb{E}_{\boldsymbol{x},f} \mathbf{1}\{f(\boldsymbol{x} + \boldsymbol{\delta}_\mathrm{adv}) \neq y\}, \tag{2}$$

where y is the true label of corresponding benign sample $\boldsymbol{x}$ and $\mathbf{1}\{\cdot\}$ is the indicator function. Here, the adversarial perturbations are constrained by ℓ_p-norm attack budget ϵ_adv:

$$\|\boldsymbol{\delta}_\mathrm{adv}\|_p \leq \epsilon_\mathrm{adv}. \tag{3}$$

Following previous work, we primarily consider $p = \infty$.

Preemptive Robustification (PR). A PR method $R \in \mathcal{R}$ aims to generate a protective variation $\boldsymbol{\delta}_\mathrm{r}$ for each benign sample $\boldsymbol{x}$, yielding a preemptively robustified sample set:

$$\mathcal{X}_\mathrm{r} = \{\boldsymbol{x} + \boldsymbol{\delta}_\mathrm{r} \mid \boldsymbol{\delta}_\mathrm{r} = R(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}\}, \tag{4}$$

that is more resistant to TAs. The robustness gains are measured by:

$$\Delta = \mathrm{ASR}(\mathcal{X}, A, f_\mathrm{s}, \mathcal{F}) - \mathrm{ASR}(\mathcal{X}_\mathrm{r}, A, f_\mathrm{s}, \mathcal{F}). \tag{5}$$

3.2 Quantifying Shared Vulnerability with High-Frequency Response

We begin with an idealized setting. Suppose the defender has full knowledge of the classifier set $\mathcal{F}$ to be attacked. For a benign sample $\boldsymbol{x}$, the ideal protective variation $\boldsymbol{\delta}_\mathrm{r}^*$ should correspond to the opposite of the expected vulnerability of classifiers in $\mathcal{F}$ with respect to $\boldsymbol{x}$. We characterize the vulnerability of a classifier on $\boldsymbol{x}$ by its white-box perturbation. Thus, the ideal $\boldsymbol{\delta}_\mathrm{r}^*$ is:

$$\boldsymbol{\delta}_\mathrm{r}^* = -\mathbb{E}_{f \in \mathcal{F}}[A(\boldsymbol{x}, f)]. \tag{6}$$

However, in practice, the defender typically lacks prior knowledge of the classifiers that will process the stored or transmitted sample $\boldsymbol{x}$, as well as the potential downstream tasks involved. Moreover, as discussed above, PR methods that rely on surrogate models and optimization or training procedures suffer from several limitations. Therefore, Equation 6 is clearly difficult to achieve in real world.

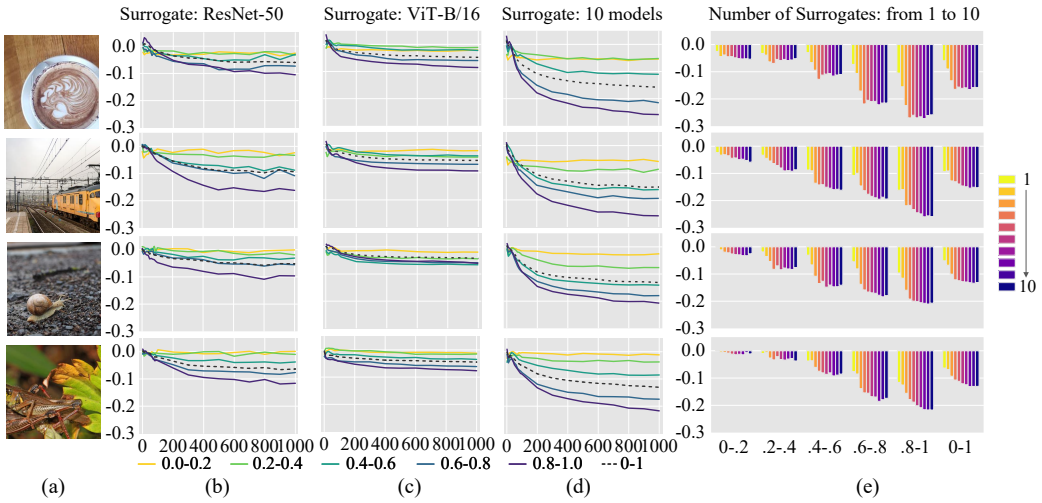


Figure 2: Cosine similarities between image Laplacian responses and their corresponding adversarial perturbations (generated by MI-FGSM). (a) Original images. (b) Perturbations generated on ResNet-50. X-axis denotes the attack iteration and y-axis denotes the cosine similarity. Different colored curves represent the cosine similarities within different texture intensity intervals. For example, $0.6 - 0.8$ indicates the cosine similarity calculated from texture elements ranked within the $60\% - 80\%$ intensity range after ascending sorting. (c) Perturbations generated on ViT-B/16. (d) Perturbations generated from 10 different surrogate models are averaged before computing the cosine similarity. (e) The x-axis denotes texture intensity intervals, and bars with different colors indicate the number of surrogate models. The iteration number is 1000.

Although the ideal protective variation $\boldsymbol{\delta}_{\mathrm{r}}^{*}$ is inaccessible in deployment, we can instead analyze its critical components. By taking the negation of such components as the approximation $\hat{\boldsymbol{\delta}}_{\mathrm{r}}$, we can partially compensate for the shared vulnerability and thereby improve resistance against TAs. Therefore, we shift our focus to analyzing the components underlying shared vulnerability.

We seek potential clues from existing studies. First, extensive research Yin et al. (2019); Wang et al. (2020); Caro et al. (2020); Pei et al. (2025) has shown that the energy of adversarial perturbations is concentrated in the mid-to-high frequency regions. Second, robustified samples generated by previous PR methods Moon et al. (2022); Frosio & Kautz (2023) consistently exhibit stronger textures and enhanced contrast compared with the original images. Furthermore, recent image attackability measure OTI Liang et al. (2026) has revealed that samples with stronger textures tend to be more resistant to adversarial attacks. Collectively, these findings from spectral analysis, empirical visualization of PR methods, and sample-centric robustness analysis all point to a close relationship between image high-frequency responses and adversarial perturbations. This naturally raises the question of whether high-frequency responses are closely associated with shared vulnerability, or even constitute the critical component of the shared vulnerability.

To this end, we design a dedicated experiment to investigate the quantitative relationship between high-frequency responses and shared vulnerability. Specifically, we convolve the image with the 8-neighborhood Laplacian kernel $K_{\mathscr{L}}$ to extract the high-frequency response:

$$\boldsymbol{x}_{\mathscr{L}} = K_{\mathscr{L}} * \boldsymbol{x}. \tag{7}$$

We generate ℓ_∞-bounded white-box adversarial perturbations (MI-FGSM) for ten classifiers, including R50, IncV3, D121, EB2, ConvN-B, ViT-B/16, BeiT-B/16, Swin-B, CaF, and PoolF. We then average these perturbations to obtain $\bar{\boldsymbol{\delta}}_{\mathrm{adv}}$ (without averaging for a single classifier), which serves as an empirical approximation of the shared vulnerability. Finally, the cosine similarity η is:

$$\eta = \frac{\boldsymbol{x}_{\mathscr{L}} \cdot \bar{\boldsymbol{\delta}}_{\mathrm{adv}}}{\|\boldsymbol{x}_{\mathscr{L}}\| \cdot \|\bar{\boldsymbol{\delta}}_{\mathrm{adv}}\|}. \tag{8}$$

The illustrated numerical results are presented in Figure 2.

Since an image $\boldsymbol{x}$ is a high-dimensional tensor (of size $3 \times 224 \times 224$ in our experiments), if high-frequency responses were unrelated to shared vulnerability, the expected η would be close to zero due to the concentration effect in high-dimensional spaces. As shown in Figure 2, this is indeed the case when adversarial perturbations are insufficiently optimized (the low-iteration regime in Figure 2(b–d)) or generated from only a single surrogate model (Figure 2(e)), where η remains close to zero. These results indicate that perturbations obtained before convergence or from a few surrogates exhibit little correlation with the Laplacian response. Although it has long been hypothesized that model vulnerability is related to high-frequency image components, previous studies have struggled to establish a statistically significant quantitative similarity, consistent with these observations.

However, as the optimization proceeds and the number of surrogate models increases, cosine similarity η between the mean of converged adversarial perturbations and the Laplacian response steadily rises to approximately 0.2. This finding is encouraging, as it provides empirical evidence that high-frequency responses constitute a critical component of the shared vulnerability, while also revealing a concrete quantitative similarity between them. Furthermore, we compute the cosine similarity over regions with different Laplacian-response magnitudes and observe that regions with stronger responses exhibit higher similarity. Specifically, the cosine similarity approaches 0.3 when evaluated on the top 20% elements with the strongest Laplacian response.

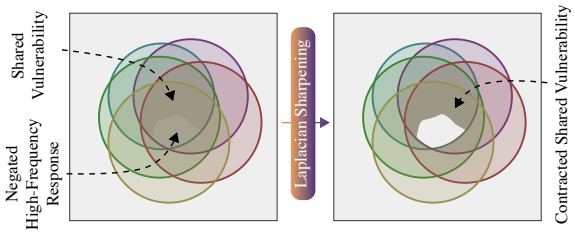


Figure 3: Illustration of Laplacian sharpening in contracting shared vulnerability by enhancing high-frequency response. Different circles represent the vulnerabilities of different models to the same image.

3.3 Fast Preemptive Robustification

Therefore, we demonstrate that high-frequency responses constitute an important component of shared vulnerability. By enhancing the high-frequency responses of images, this vulnerable component can be effectively compensated, thereby shrinking the shared vulnerability exploited by TAs and rendering them ineffective, as illustrated in Figure 3.

To this end, we introduce Laplacian sharpening as a novel PR method to defend against TAs, where the strength is controlled by the sharpening coefficient $\alpha(\alpha > 0)$. This method can be formulated as follows:

$$\boldsymbol{x}_r = \mathrm{Clip}_{[0,1]}\big(\boldsymbol{x} + \alpha \cdot (K_{\mathscr{L}} * \boldsymbol{x})\big). \tag{9}$$

Although the proposed method is extremely simple, requiring only a single channel-wise convolution with a 3×3 kernel, the experiments in Section 4 demonstrate its effectiveness. In addition, this method inherits the favorable properties of Laplacian sharpening. With a small α, the robustified images are imperceptible to humans, as shown in Figure 4. Even when the enhancement is slightly noticeable, Laplacian sharpening itself is a technique designed to improve perceptual quality. Unlike optimization-based or generator-based approaches, it does not compromise the visual quality.

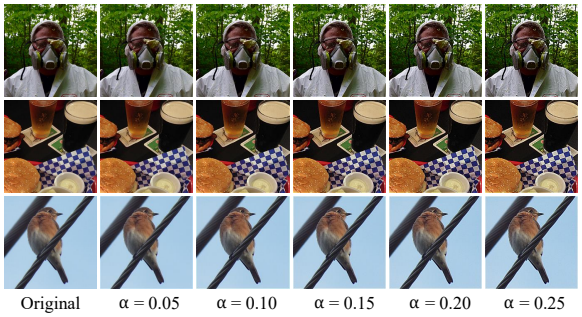


Figure 4: Laplacian sharpened images with varying α.

We name this method **F**ast **P**reemptive **R**obustification (**FPR**), which is the first surrogate-free, optimization-free, training-free, and human-interpretable PR method.

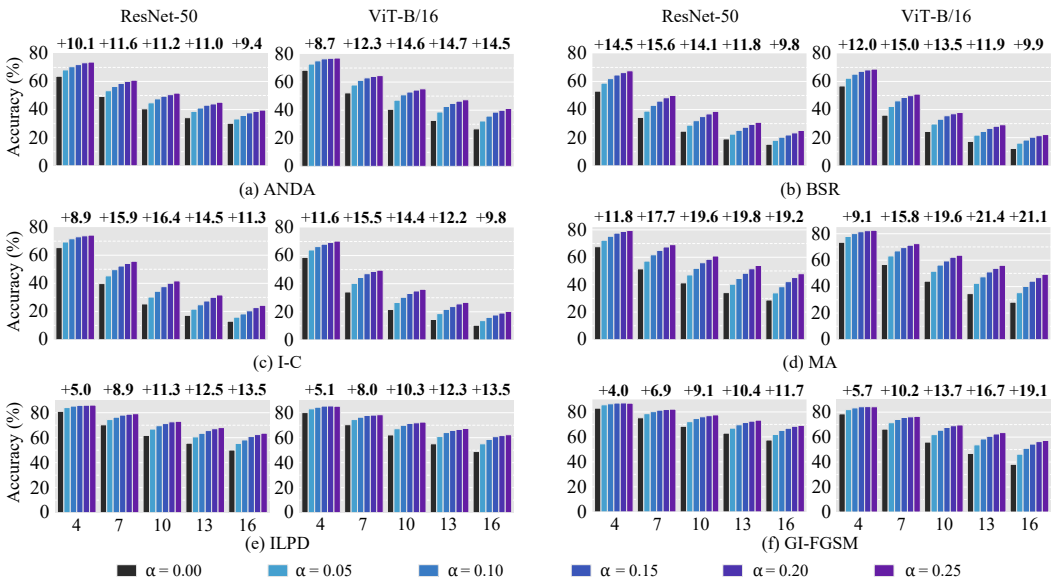


Figure 5: Accuracy (%) under various non-targeted transferable attacks and ϵ_{adv}. Gains from $\alpha = 0.00$ (w/o FPR) to $\alpha = 0.25$ are annotated at the top. Similar annotations apply hereafter.

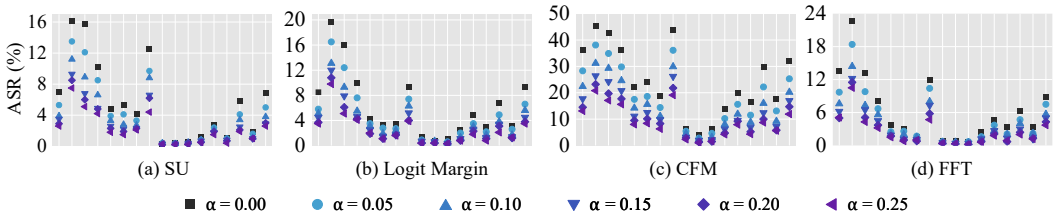


Figure 6: ASRs (%) of targeted transferable attacks on robustified samples by FPR with varying α.

4 Experiments and Results

4.1 Setup

Benchmarks. We conduct a comprehensive evaluation across a broad suite of representative transferable attack benchmarks, covering diverse attack mechanisms and optimization objectives. Specifically, ANDA Fang et al. (2024) and GI-FGSM Wang et al. (2024a) represent gradient-based attacks, BSR Wang et al. (2024b) and I-C Liang & Pun (2025) represent transformation-based attacks, ILPD Li et al. (2023) represents advanced objective attack, and MA Ma et al. (2024) represents model-related attacks. The targeted attacks include SU Wei et al. (2023), Logit Margin Weng et al. (2023), CFM Byun et al. (2023), and FFT Zeng et al. (2024). To assess the robustness of FPR against attacks generated from multiple surrogate models, we include three ensemble-based attacks, namely Ens Liu et al. (2017), AdaEA Chen et al. (2023a), and CWA Chen et al. (2024).

Datasets and Environment. Following previous work, we use the NIPS 2017 Adversarial Competition dataset [1]. All experiments are conducted on a single NVIDIA A100 GPU.

Models. The models are: (1) ResNet-50 (R50), (2) ConvNeXt-B (ConvN), (3) WideResNet-50 (WR50), (4) DenseNet-161 (D161), (5) RegNet-X (RegN), (6) EfficientNet-B2, (7) GoogLeNet (GgN), (8) Xception-71 (Xcept), (9) Inception-ResNet-V2 (IR2), (10) ViT-B/16 (VB16), (11) Convformer (ConvF), (12) Visformer (VisF), (13) Poolformer (PoolF), (14) ViT-B/8 (VB8), (15) Swin Transformer-B (Swin), (16) PiT-B (PiT), (17) XCiT-S (XCiT), and (18) Caformer-M36 (CaF).

[1] https://github.com/anlthms/nips-2017.git

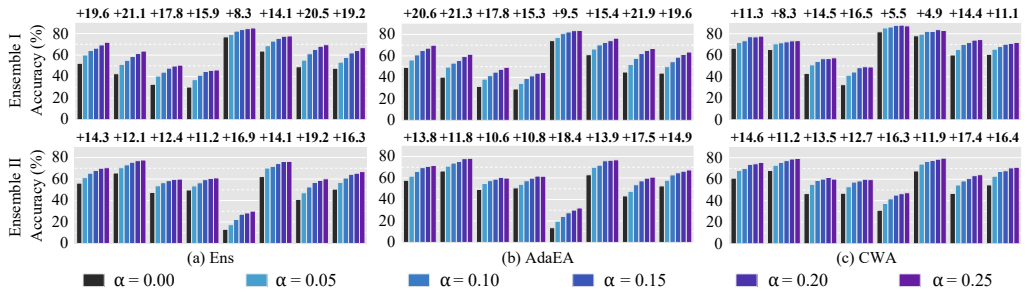


Figure 7: Accuracy (%) under various ensemble-based attacks, evaluated on robustified images by FPR with different α.

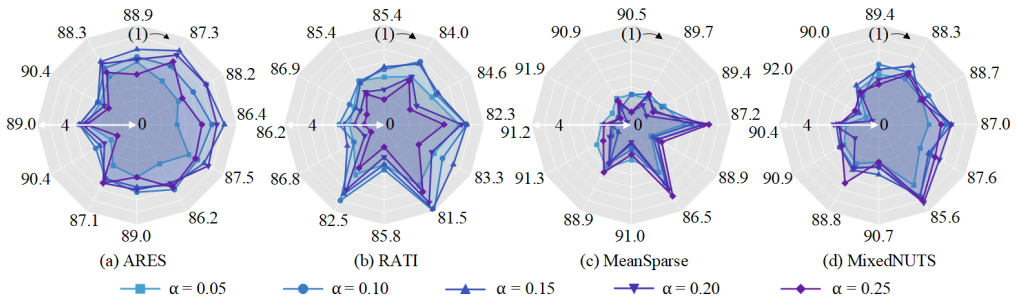


Figure 8: Gains (%) of adversarially trained models integrated with our FPR. The axes represent: ANDA (1–2), BSR (3–4), I-C (5–6), MA (7–8), ILPD (9–10), and GI-FGSM (11–12). Odd indices use R50 as the surrogate, while even indices use VB16.

4.2 To Non-Targeted Transferable Attacks

First, we evaluate FPR with varying sharpening coefficients α against diverse non-targeted transferable attacks under different perturbation budgets $\epsilon_{\text{adv}} \in \{4, 7, 10, 13, 16\}/255$. Moreover, we set the number of iterations to $T = 10$ and the step size to ϵ_{adv}/T. The surrogate models include R50 and VB16. The target models consist of the remaining 17 models listed in Section 4.1. Figure 5 demonstrates that FPR is highly effective against non-targeted transferable attacks. The robustness gains are consistently observed across different attack types, perturbation budgets, and surrogate models, demonstrating the generality of FPR. On average, FPR reduces ASR by 12.7%.

4.3 To Targeted Transferable Attacks

In this experiment, we evaluate our FPR against targeted transferable attacks. The target labels are *1* and *500*. The surrogate model is R50. The target models include (2)-(5) and (10)-(13) to which targeted transferable attacks are effective. The budget ϵ_{adv} is $16/255$ with the iterations $T = 200$ and the step size ϵ_{adv}/T. Results in Figure 6 indicate that our FPR provides strong robustification against transferable targeted attacks. On average, FPR reduces ASR from 10.7% to 4.1%. The robustness gains are consistently observed across different targeted attack objectives, with larger sharpening coefficients α generally yielding stronger improvements.

4.4 To Ensemble-based Transferable Attacks

Beyond evaluating FPR against single-surrogate attacks, we are also interested in the effectiveness against ensemble-based transferable attacks. In this experiment, ensemble I = {R50, EB0, IncV3} and ensemble II = {VS16, VS32, BeiTB16} are used as surrogates to evaluate across various ensemble-based attacks. The target models are the same as those in Section 4.3. The budget $\epsilon_{\text{adv}} = 10/255$ with $T = 10$. As shown in Figure 7, FPR consistently reduces the ASR across both ensembles, demonstrating its effectiveness against attacks that exploit shared vulnerabilities across multiple surrogate models. On average, FPR reduces ASR by 14.7%.

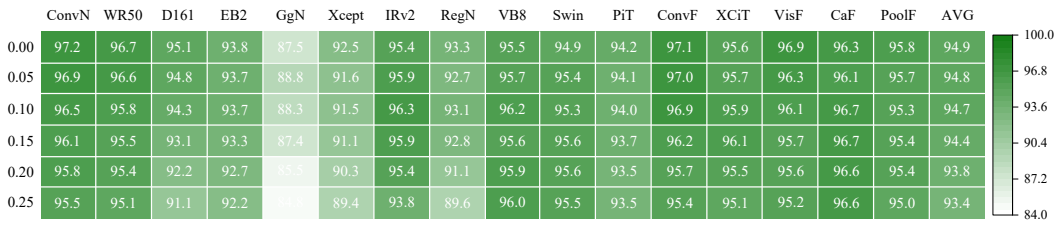


Figure 9: Impact of FPR with varying α on standard accuracy (%).

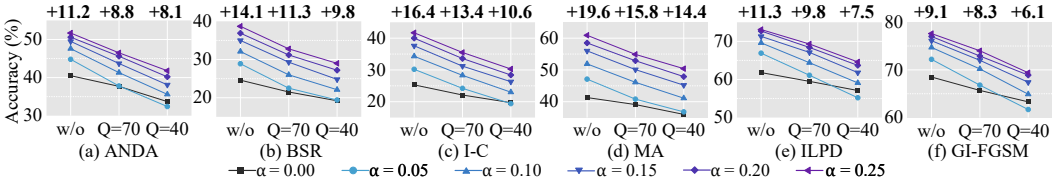


Figure 10: Impact of JPEG compression on FPR.

4.5 Integrating with Adversarial Training

Above experiments evaluate FPR as a standalone preemptive robustification method, without incorporating other defense mechanisms. We next investigate whether FPR can complement existing adversarial training methods by jointly applying FPR to adversarially trained models. Specifically, we evaluate four advanced methods: ARES (Swin-B) Liu et al. (2025a), RATI Singh et al. (2023) (ViT-S + ConvStem), MeanSparse Amini et al. (2024) (ConvNeXt-L), and MixedNUTS Bai et al. (2024) (ConvNeXtV2-L + Swin-L). Adversarial examples are generated following the same settings as in 4.2. Results in Figure 8 indicate that FPR can further enhance the robustness of adversarially trained models, providing complementary robustness gains on top of existing defenses.

4.6 Impact on Standard Accuracy

This experiment evaluates the impact of FPR on benign samples. The results in Figure 9 demonstrate that FPR achieves a favorable trade-off between adversarial robustness and standard accuracy. On average, when $\alpha = 0.10$, FPR incurs only a 0.2% accuracy fluctuation. Even with a $\alpha = 0.25$, the accuracy fluctuation remains merely 1.5%. Compared with other advanced defenses Wang et al. (2026); Mirsadeghi (2026), this negligible impact highlights the advantage of FPR.

4.7 Robustness against JPEG Compression

We further evaluate the robustness of FPR-robustified samples under JPEG compression, a common post-processing operation during storage and transmission. Specifically, we first apply JPEG compression to the FPR-robustified samples, then generate adversarial examples from the compressed samples and evaluate their robustness. Compression quality factors Q are set to 40 and 70. Results presented in Figure 10 demonstrate that FPR provides stable gains under JPEG compression.

5 Conclusions and Limitations

This work reveals a strong cosine similarity between image high-frequency responses and shared model vulnerability, providing quantitative evidence for their intrinsic relationship for the first time. Building on this finding, we propose FPR, the first surrogate-free, optimization-free, training-free, and human-interpretable PR method, which effectively mitigates transferable attacks through conventional Laplacian sharpening. Despite requiring only a single channel-wise 3×3 convolution, FPR achieves a favorable robustness-accuracy trade-off and remains effective under post-processing transformations, demonstrating its practical applicability. This work provides a new perspective on preemptive robustification and we hope it will inspire more effective and sophisticated methods.

AI USE STATEMENT

We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, and interpreting results. Generating synthetic data sets is not applicable to this work. Additionally, we used generative AI tools for editing a research paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, dataset release practices, potentially harmful insights, methodologies and applications, potential conflicts of interest or sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance issues, and research integrity issues.

REPRODUCIBILITY STATEMENT

We will release the complete source code upon the acceptance of the paper.

REFERENCES

Sajjad Amini, Mohammadreza Teymoorianfard, Shiqing Ma, and Amir Houmansadr. Meansparse: Post-training robustness enhancement through mean-centered feature sparsification. *arXiv preprint arXiv:2406.05927*, 2024.

Yatong Bai, Mo Zhou, Vishal Patel, and Somayeh Sojoudi. Mixednuts: Training-free accuracy-robustness balance via nonlinearly mixed classifiers. *Transactions on machine learning research*, 2024.

Junyoung Byun, Myung-Joon Kwon, Seungju Cho, Yoonji Kim, and Changick Kim. Introducing competition to boost the transferability of targeted adversarial examples through clean feature mixup. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24648–24657, 2023.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Josue Ortega Caro, Yilong Ju, Ryan Pyle, Sourav Dey, Wieland Brendel, Fabio Anselmi, and Ankit Patel. Local convolutions cause an implicit bias towards high frequency adversarial examples. *arXiv preprint arXiv:2006.11440*, 2020.

Bin Chen, Jiali Yin, Shukai Chen, Bohao Chen, and Ximeng Liu. An adaptive model ensemble adversarial attack for boosting adversarial transferability. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4489–4498, 2023a.

Erh-Chung Chen, Pin-Yu Chen, I Chung, Che-Rung Lee, et al. Data-driven lipschitz continuity: A cost-effective approach to improve adversarial robustness. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 698–707, 2026.

Huanran Chen, Yichi Zhang, Yinpeng Dong, Xiao Yang, Hang Su, and Jun Zhu. Rethinking model ensemble in transfer-based adversarial attacks. *arXiv preprint arXiv:2303.09105*, 2023b.

Huanran Chen, Yichi Zhang, Yinpeng Dong, Xiao Yang, Hang Su, and Jun Zhu. Rethinking model ensemble in transfer-based adversarial attacks. In *The Twelfth International Conference on Learning Representations*, 2024.

Gilad Cohen and Raja Giryes. Simple post-training robustness using test time augmentations and random forest. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3996–4006, 2024.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

Zhengwei Fang, Rui Wang, Tao Huang, and Liping Jing. Strong transferable adversarial attacks via ensembled asymptotically normal distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24841–24850, 2024.

Samar Fares, Klea Ziu, Toluwani Aremu, Nikita Durasov, Martin Takáč, Pascal Fua, Ivan Laptev, and Karthik Nandakumar. Mirrorcheck: Efficient adversarial defense for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 496–506, 2026.

Iuri Frosio and Jan Kautz. The best defense is a good offense: Adversarial augmentation against adversarial attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4067–4076, 2023.

Fuquan Gan and Yan Wo. Boosting the transferability of adversarial examples through gradient aggregation. *IEEE Transactions on Information Forensics and Security*, 2025.

Rafael C Gonzalez. *Digital image processing*. Pearson education india, 2009.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Yu Guo, Weiquan Liu, Qingshan Xu, Shijun Zheng, Shujun Huang, Yu Zang, Siqi Shen, Chenglu Wen, and Cheng Wang. Boosting adversarial transferability through augmentation in hypothesis space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19175–19185, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Lirong He, Qingzhong Ai, Yuqing Lei, Lili Pan, Yazhou Ren, and Zenglin Xu. Edge enhancement improves adversarial robustness in image classification. *Neurocomputing*, 518:122–132, 2023a.

Xilin He, Qinliang Lin, Cheng Luo, Weicheng Xie, Siyang Song, Feng Liu, and Linlin Shen. Shift from texture-bias to shape-bias: Edge deformation-based augmentation for robust object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1526–1535, 2023b.

Edgar Heinert, Matthias Rottmann, Kira Maag, and Karsten Kahl. Reducing texture bias of deep neural networks via edge enhancing diffusion. *arXiv preprint arXiv:2402.09530*, 2024.

Wenzheng Jiang, Ke Liang, Xuankun Rong, Jingxuan Zhou, Zhengyi Zhong, Guancheng Wan, and Ji Wang. Purmm: Attention-guided test-time backdoor purification in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35562–35570, 2026.

Doris Kaltenecker, Izabela Horvath, Rami Al-Maskari, Ying Chen, Zeynep Ilgin Kolabas, Luciano Hoeher, Mihail Todorov, David-Paul Minde, Saketh Kapoor, Sena Gül Turhan, et al. A deep-learning framework reveals whole-body perturbations at cell level. *Nature*, pp. 1–11, 2026.

Xiangyin Kong, Xiaoyu Jiang, Zhihuan Song, and Zhiqiang Ge. Data id extraction networks for unsupervised class-and classifier-free detection of adversarial examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Chun Tong Lei, Hon Ming Yam, Zhongliang Guo, Yifei Qian, and Chun Pong Lau. Instant adversarial purification with adversarial consistency distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24331–24340, 2025.

Airan Li, Xinzhi Wu, Longquan Wang, Gang Wu, Jiankang Li, Zhao Hu, Xinyuan Wang, and Takao Mori. Composable neural emulators accelerate thermoelectric generator design. *Nature*, 652 (8110):643–649, 2026.

Qizhang Li, Yiwen Guo, Wangmeng Zuo, and Hao Chen. Improving adversarial transferability via intermediate-level perturbation decay. *Advances in Neural Information Processing Systems*, 36: 32900–32912, 2023.

Teng Li, Xingjun Ma, and Yu-Gang Jiang. Aim: Additional image guided generation of transferable adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4941–4949, 2025.

Jiaming Liang and Chi-Man Pun. Ic attack: In-place and cross-pixel augmentations for highly transferable transformation-based attacks. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5345–5354, 2025.

Jiaming Liang, Haowei Liu, and Chi-Man Pun. Oti: A model-free and visually interpretable measure of image attackability, 2026. URL `https://arxiv.org/abs/2601.17536`.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

Chang Liu, Yinpeng Dong, Wenzhao Xiang, Xiao Yang, Hang Su, Jun Zhu, Yuefeng Chen, Yuan He, Hui Xue, and Shibao Zheng. A comprehensive study on robustness of image classification models: Benchmarking and rethinking. *International Journal of Computer Vision*, 133(2):567–589, 2025a.

Hui Liu, Bo Zhao, Jiabao Guo, Kehuan Zhang, and Peng Liu. A lightweight unsupervised adversarial detector based on autoencoder and isolation forest. *Pattern Recognition*, 147:110127, 2024.

Jiani Liu, Zhiyuan Wang, Zeliang Zhang, Chao Huang, Susan Liang, Yunlong Tang, and Chenliang Xu. Harnessing the computation redundancy in vits to boost adversarial transferability. *Advances in Neural Information Processing Systems*, 38:76855–76876, 2026.

Renpu Liu, Hao Wu, Jiawei Zhang, Xin Cheng, Xiangyang Luo, Bin Ma, and Jinwei Wang. Pixel2feature attack (p2fa): Rethinking the perturbed space to enhance adversarial transferability. In *Forty-second International Conference on Machine Learning*, 2025b.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.

Avery Ma, Amir-massoud Farahmand, Yangchen Pan, Philip Torr, and Jindong Gu. Improving adversarial transferability via model alignment. In *European Conference on Computer Vision*, pp. 74–92. Springer, 2024.

Seyed Mohammad Hadi Mirsadeghi. Active machine learning approach to adversarial training improves trade-off between natural accuracy and adversarial robustness. *Scientific Reports*, 2026.

Seungyong Moon, Gaon An, and Hyun Oh Song. Preemptive image robustification for protecting users against man-in-the-middle adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7823–7830, 2022.

Aamir Mustafa, Salman H Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 29: 1711–1724, 2019.

Krishna Kanth Nakka and Alexandre Alahi. Nat: Learning to attack neurons for enhanced adversarial transferability. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7593–7604. IEEE, 2025.

Ruicheng Niu, Ziyuan Zhu, Chaofei Li, and Dan Meng. Search robust and adaptable architecture. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5155–5159. IEEE, 2024.

Gaozheng Pei, Ke Ma, Yingfei Sun, Qianqian Xu, and Qingming Huang. Diffusion-based adversarial purification from the perspective of the frequency domain. In *International Conference on Machine Learning*, pp. 48679–48698. PMLR, 2025.

Jinjia Peng, Zeze Tao, Huibing Wang, Meng Wang, and Yang Wang. Boosting adversarial transferability via residual perturbation attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1261–1270, 2025.

ShengYun Peng, Weilin Xu, Cory Cornelius, Matthew Hull, Kevin Li, Rahul Duggal, Mansi Phute, Jason Martin, and Duen Horng Chau. Robust principles: Architectural design principles for adversarially robust cnns. *arXiv preprint arXiv:2308.16258*, 2023.

Debdeep Sanyal, Manodeep Ray, and Murari Mandal. Antidote: Bi-level adversarial training for tamper-resistant llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32893–32901, 2026.

John GM Schavemaker, Marcel JT Reinders, Jan J Gerbrands, and Eric Backer. Image sharpening by morphological filtering. *Pattern Recognition*, 33(6):997–1012, 2000.

Divya Shanmugam, Helen Lu, Swami Sankaranarayanan, and John Guttag. Test-time augmentation improves efficiency in conformal prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20622–20631, 2025.

Naman Deep Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. *Advances in Neural Information Processing Systems*, 36:13931–13955, 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Bowen Tang, Zheng Wang, Yi Bin, Qi Dou, Yang Yang, and Heng Tao Shen. Ensemble diversity facilitates adversarial transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24377–24386, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Donghua Wang, Wen Yao, Tingsong Jiang, Xiaohu Zheng, and Junqi Wu. Improving the transferability of adversarial examples by feature augmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8684–8694, 2020.

Jiafeng Wang, Zhaoyu Chen, Kaixun Jiang, Dingkang Yang, Lingyi Hong, Pinxue Guo, Haijing Guo, and Wenqiang Zhang. Boosting the transferability of adversarial attacks with global momentum initialization. *Expert Systems with Applications*, 255:124757, 2024a.

Kunyu Wang, Xuanran He, Wenxuan Wang, and Xiaosen Wang. Boosting adversarial transferability by block shuffle and rotation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24336–24346, 2024b.

Yanyun Wang, Qingqing Ye, Li Liu, Zi Liang, and Haibo Hu. Robust alignment: Harmonizing clean accuracy and adversarial robustness in adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 768–778, 2026.

Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Enhancing the self-universality for transferable targeted attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12281–12290, 2023.

Juanjuan Weng, Zhiming Luo, Shaozi Li, Nicu Sebe, and Zhun Zhong. Logit margin matters: Improving transferable targeted adversarial attack by logit calibration. *IEEE Transactions on Information Forensics and Security*, 18:3561–3574, 2023.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2730–2739, 2019.

Honghui Xu, Chuangjie Fang, Yibin Wang, Jie Wu, and Jianwei Zheng. Laboring on less labors: Rpca paradigm for pan-sharpening. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11393–11402, 2025.

Zhaoyuan Yang, Zhiwei Xu, Jing Zhang, Richard Hartley, and Peter Tu. Adversarial purification with the manifold hypothesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16379–16387, 2024.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019.

Hui Zeng, Biwei Chen, and Anjie Peng. Enhancing targeted transferability via feature space fine-tuning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4475–4479. IEEE, 2024.

Buyue Zhang and Jan P Allebach. Adaptive bilateral filter for sharpness enhancement and noise removal. *IEEE transactions on Image Processing*, 17(5):664–678, 2008.

Yi Zhang, Yuhang Chen, Zhen Chen, Wenjie Ruan, Xiaowei Huang, Siddartha Khastgir, and Xingyu Zhao. Adversarial training for probabilistic robustness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1675–1685, 2025.

Desheng Zheng, Wuping Ke, Xiaoyu Li, Yaoxin Duan, Guangqiang Yin, and Fan Min. Enhancing the transferability of adversarial attacks via multi-feature attention. *IEEE Transactions on Information Forensics and Security*, 20:1462–1474, 2025.

Monty-Maximilian Zühlke and Daniel Kudenko. Adversarial robustness of neural networks from the perspective of lipschitz calculus: A survey. *ACM Computing Surveys*, 57(6):1–41, 2025.