SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

December 2025, Vol. 68, Iss. 12, 222102:1–222102:15
https://doi.org/10.1007/s11432-024-4542-8

Propagation rectified attack: on improving adversarial transferability

Xuxiang SUN†, Hongyu PENG†, Gong CHENG* & Junwei HAN

School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

Received 25 September 2024/Revised 8 December 2024/Accepted 30 May 2025/Published online 11 September 2025

Abstract In the context of enhancing adversarial transferability along the line of surrogate refinement, this paper investigates transferable black-box attacks and proposes propagation rectified attack (PRA), which rectifies both the forward and backward propagation of the surrogate. Specifically, on rectifying the forward propagation, we develop multi-scale feature rectification (MSFR), which applies the feature rectifications to different levels of features, encouraging the forward propagation to be in the proper status of adversarial optimization, and highlighting the necessity and benefits of multi-scale decays for enhancing transferability, which has been ignored by existing studies. Additionally, for the backward propagation, existing studies only pursue the smoothness of the alternative activation derivative. Instead, we derive a more feasible and comprehensive conclusion. First, the derivative of the activation should be non-negative and monotonic, maintaining the gradient integrity. Besides, its second derivative should have a certain degree of magnitude near zero. Based on these findings, we further propose adaptive activation rectification (AAR), which takes the specificity of the features from each layer into account, thereby building a more effective activation alternative. Our evaluations are performed on two widely adopted datasets: ImageNet (with average gains of +13.85% over ten classical CNN models and +15.38% over six non-conventional-CNN models) and CIFAR-10 (with average gains of +5.5%). Codes will be released at https://github.com/phyyyy/PRA.

Keywords computer vision, image recognition, adversarial attack, transferability, surrogate refinement

Citation Sun X X, Peng H Y, Cheng G, et al. Propagation rectified attack: on improving adversarial transferability. Sci China Inf Sci, 2025, 68(12): 222102, https://doi.org/10.1007/s11432-024-4542-8

1 Introduction

Deep neural networks (DNNs) have been widely applied in various areas such as image understanding [1–8], underscoring the urgent concerns regarding their security. Among these concerns, adversarial attacks have emerged as a significant challenge, exposing the profound vulnerability of DNNs to such threats [9–11]. To date, these attacks have posed serious risks to almost all the methods based on DNNs for existing vision tasks such as image classification [12–18] and object tracking [19–21], threatening the reliability and security of these systems. In this regard, researchers have dedicated significant efforts to studying adversarial robustness [9, 10, 22, 23]. Generally speaking, black-box attacks, which do not require access to the internal information of the victim model, exhibit a more severe threat to information security.

In the black-box setting, transferable black-box attacks are particularly concerning compared to query-based methods [24, 25], especially when the opportunity for numerous queries is limited. Therefore, studying transferable attacks is essential for understanding the potential vulnerabilities of DNNs and developing effective defenses to protect sensitive information. Within this research domain, modifying the propagation of the surrogate model has been shown to be an effective strategy for enhancing the transferability of such attacks. However, previous efforts [26–29] usually focused on rectifying either the forward or backward propagation alone. Instead, this paper introduces a bidirectional approach, propagation rectified attack (PRA), which simultaneously rectifies the forward and backward propagation.

Specifically, during forward propagation, applying perturbation decay to a single intermediate layer is proven to be a promising way [27]. However, simply applying perturbation decay to one layer risks the possibility that the decayed perturbations may be amplified by subsequent layers [30], potentially

* Corresponding author (email: gcheng@nwpu.edu.cn)
† These authors contributed equally to this work.

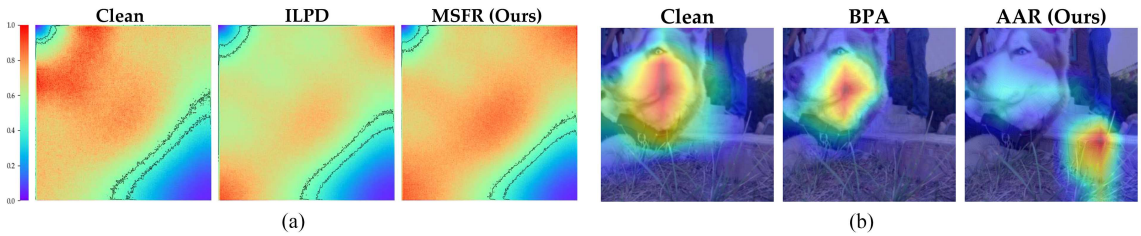

Figure 1 (Color online) Qualitative results. (a) Spectrum attention maps [31] in discrete cosine transform space, with the specific inputs indicated by the top caption. In summary, a broader range of differences means more levels of information distortions. (b) Attention maps of different inputs.

allowing the perturbations to escape the constraints of the decay operation. Thus, encouraging the entire forward propagation to be in a proper adversarial optimization state[1)] could better address this limitation. Moreover, since different levels of features contain varying information, limiting the decay to only the intermediate-level features ignores the usage of the other levels of features, making it somewhat one-sided. To this end, we develop multi-scale feature rectification (MSFR), which covers a broader attention range, involving the features at different levels. Based on the above analysis, this comprehensive approach facilitates the perturbations containing more levels of information, achieving a higher level of distortions. As shown in Figure 1(a) [31], MSFR achieves more intense drifting on attention regions than ILPD, in both the range (referring to the contours in the upper left and lower right) and intensity (referring to the overall differences). This means that MSFR exhibits broader and stronger information distortions than ILPD, thereby ensuring higher transferability.

Meanwhile, on the backward propagation rectification, existing studies usually pursue replacing ReLU with the specific alternative, of which the derivative is smooth. However, are there other necessary characteristics of the ideal alternative activation? To answer this question, we conduct extensive explorations and conclude that the smoothness alone is insufficient. Instead, the derivative of the alternative activation should first be non-negative and monotonic, and it also should maintain the gradient integrity[2)]. Accordingly, Softplus-β might be a suitable alternative activation. Nevertheless, we further reveal an unprecedented insight, i.e., the second derivative of the activation should have a certain degree of magnitude near zero (see Subsection 3.4 for detailed explanation). Meanwhile, considering that the characteristics across different layers usually vary a lot, using a fixed β and simply aligning them to the same setting may not be entirely appropriate. Thus, we propose adaptive activation rectification (AAR) for the backward propagation rectification, where AAR adaptively adjusts the β of each Softplus-β according to its input features. Consequently, as affirmed by Figure 1(b), AAR can effectively shift the attention of the victim model further compared to other methods, thereby enhancing the transferability.

Our contributions are: (1) we propose PRA, a bidirectional method that incorporates both the forward and backward propagation rectifications; (2) in our PRA, we propose MSFR that applies the perturbation decays to different levels of features, encouraging the forward propagation to maintain a proper adversarial optimization state and overcoming the issue of incomplete utilization of features when applying decay to a single intermediate layer; (3) meanwhile, we clarify the incomplete definition of the ideal substitution activation used during backward propagation in previous studies, and further propose AAR that considers the specificity of the features from each layer; (4) we perform extensive experiments over numerous models on the ImageNet and CIFAR-10 datasets, and these results demonstrate the significant superiority of our PRA.

2 Related work

2.1 Adversarial attacks

In general, adversarial attacks can be broadly classified into white-box and black-box attacks. In short, white-box attacks assume that the attacker can access the gradients of the victim model. In contrast, the black-box setting implies that the attacker lacks any knowledge except for the training data of the victim

1) In this paper, "adversarial optimization state" refers to the state where an obstacle opposing the optimization objective is introduced, and we achieve this state by decaying the perturbed features.

2) In this paper, "gradient integrity" means that no truncation is performed on the gradients during backward propagation.

model. However, the practical applicability of white-box attacks could be limited, as direct access to the internal information of the victim model is often unavailable in real-world scenarios. Thus, the study on black-box attacks is more urgent, to some extent. To date, black-box attacks typically include query-based and transfer-based attacks. Among them, query-based attacks [24, 25] involve interacting with the victim by conducting a series of queries and obtaining the outputs to infer the decision boundaries. In contrast, transfer-based attacks craft adversarial examples on a local surrogate model.

2.2 Transfer-based attacks

In the field of transfer-based attacks, existing research can be roughly categorized into several parallel perspectives, including input transformation, gradient stabilization, advanced objective, and surrogate refinement. To be short, input transformation [32–34] involves applying feasible transformations to the input data, thereby enhancing the transferability of adversarial examples. Gradient stabilization is also widely studied in transfer-based attacks [35–37], focusing on the design of more stable gradient update strategies. Besides, the scope of advanced objective involves designing more effective objective functions that can provide further performance gains in terms of transferability. For example, some of the approaches in this area design feature-level objective functions [26, 38], or leverage integrated gradients [39, 40] to build new objectives. Surrogate refinement focuses on refining the surrogate model to enhance the effectiveness of the attack. These approaches can be categorized into forward or backward refinements. For forward refinement, intermediate-level attack++ (ILA++) [41] enhances transferability by mapping intermediate-level feature discrepancies to prediction losses using linear regression. Intermediate-level perturbation decay (ILPD) [27] improves the transferability by applying perturbation decay to a single layer. For backward refinement, skip gradient method (SGM) [29] encourages prioritizing the gradients propagated through the skip connections over residual modules in ResNet-like neural networks. Linear backward propagation (LinBP) [28] performs the backward propagation in a more linear manner. It maintains the forward propagation normally but restores a proportion of the gradients from the inactivated neurons while retaining the gradients from the activated neurons. Backward propagation attack (BPA) [42] further proposes to replace the conventional activation, ReLU, with SiLU and incorporates the softmax with temperature to calculate the derivative of max-pooling.

As can be found, the substantial differences between our PRA and these studies lie in: (1) our PRA is a bidirectional approach that incorporates the rectifications for both forward and backward propagation, rather than focusing on one of them; (2) the MSFR of PRA is a comprehensive strategy that applies decays to different levels of features, thereby leveraging a wider range of features and achieving more extensive information distortions; (3) the AAR of PRA confirms more concrete and feasible properties of the ideal activation alternative, i.e., being non-negative and monotonic, maintaining the gradient integrity, and having a second derivative with a certain magnitude near zero. Additionally, it also adaptively accounts for the specificity of the features from each layer.

3 Methodology

3.1 Problem formulation

The basic goal of adversarial attacks is to find the adversarial example $\boldsymbol{x}^{\text{adv}}$ that causes the model f to misclassify the input. In general, the perturbation $\boldsymbol{\delta}$ in $\boldsymbol{x}^{\text{adv}}$ is typically constrained by ℓ_p norm to ensure its imperceptibility: $\|\boldsymbol{\delta}\|_p \leqslant \epsilon$. Formally, the above description can be expressed as

$$\boldsymbol{x}^{\text{adv}} = \boldsymbol{x} + \boldsymbol{\delta} \quad \text{s.t.} \quad f(\boldsymbol{x}^{\text{adv}}) \neq \boldsymbol{y}, \tag{1}$$

where $\boldsymbol{x}$ is the clean input, $\boldsymbol{y}$ is the one-hot ground truth.

In the black-box setting, a surrogate model f_s is used to generate adversarial examples to attack the victim model f_t. In this context, the perturbation $\boldsymbol{\delta}$ is usually optimized to maximize the loss function $\mathcal{L}$ of the surrogate model, as stated by

$$\boldsymbol{\delta} = \arg\max_{\|\boldsymbol{\delta}\|_p \leqslant \epsilon} \mathcal{L}(f_s(\boldsymbol{x} + \boldsymbol{\delta}), \boldsymbol{y}). \tag{2}$$

To achieve this, a commonly adopted attack framework is the gradient-sign based iterative approach, where the adversarial example at the $i+1$ iteration is formulated as follows:

$$\boldsymbol{x}_{i+1}^{\text{adv}} = \Pi_0^{255}(\boldsymbol{x} + \boldsymbol{\delta}_{i+1}), \quad \boldsymbol{\delta}_{i+1} = \Pi_{-\epsilon}^{\epsilon}(\boldsymbol{\delta}_i + \alpha \cdot \text{sign}(\nabla_{\boldsymbol{x}} \mathcal{L}(f_s(\boldsymbol{x} + \boldsymbol{\delta}_i), \boldsymbol{y}))), \tag{3}$$

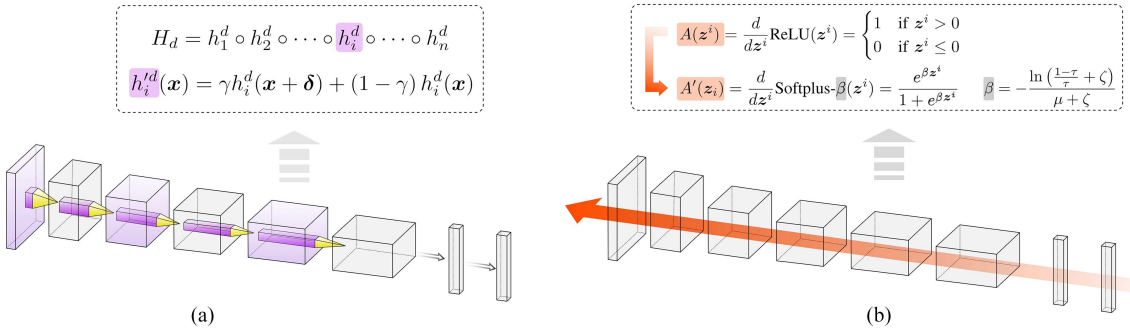

Figure 2 (Color online) Sketches of the key components in our PRA. (a) Illustration of our MSFR, which periodically applies perturbation decays to the layers highlighted by purple blocks. Its operation is formulated in the dotted box. (b) Illustration of our AAR, which rectifies the derivative of the activation during the backward propagation. Specifically, we use the proposed $A'(\boldsymbol{z}_i)$ to replace $A(\boldsymbol{z}_i)$.

Algorithm 1 Algorithm of our PRA.

Definitions: The adversarial example at the i-th iteration $\boldsymbol{x}_i^{\text{adv}}$, clean example $\boldsymbol{x}$, the number of attack iterations T.
Definitions: $\boldsymbol{x}_1^{\text{adv}} = \boldsymbol{x}$.
for $i \leftarrow 1$ to T **do**
  // Keeping the activation the same as the default one (ReLU) in the surrogate;
  Feeding $\boldsymbol{x}_i^{\text{adv}}$ and $\boldsymbol{x}$ to the surrogate;
  Applying MSFR via Eqs. (6) and (7) to the shallow, intermediate, and deep layers;
  Calculating current loss via Eq. (4);
  // Applying AAR via replacing all the activations with our adaptive Softplus-β;
  Acquiring $\boldsymbol{x}_{i+1}^{\text{adv}}$ via Eq. (3);
end
Adversarial examples: $\boldsymbol{x}^{\text{adv}} = \boldsymbol{x}_T^{\text{adv}}$.

where $\Pi_a^b(\cdot)$ clips its input to the interval $[a, b]$, α is the step size, and sign$(\cdot)$ represents the sign function. Generally, for the loss function $\mathcal{L}$, we use the cross-entropy loss as the objective, as shown in

$$\mathcal{L}_{\text{CE}}(f(\boldsymbol{x}+\boldsymbol{\delta}), \boldsymbol{y}) = -\sum_{i=1}^{C} \boldsymbol{y}_i \log f(\boldsymbol{x}+\boldsymbol{\delta})_i, \tag{4}$$

where C is the number of classes, $\boldsymbol{y}_i$ is the i-th entry of $\boldsymbol{y}$, and $f(\boldsymbol{x}+\boldsymbol{\delta})_i$ is the probabilities for class i.

3.2 Method overview

In summary, our PRA involves two novel approaches, multi-scale feature rectification, and adaptive activation rectification, as illustrated in Figure 2. For ease of implementation, we provide the overall pseudo-code, outlined in Algorithm 1. Next, we will describe the concrete forms of our MSFR and AAR.

3.3 Multi-scale feature rectification (MSFR)

In this paper, we denote the surrogate model as $f_s(\boldsymbol{x}) = G(H(\boldsymbol{x}))$, where G stands for the classification head and H represents the feature extractor. Then, the architecture of H is represented by

$$H = H_s \circ H_m \circ H_d, \tag{5}$$

where H_s, H_m, and H_d represent shallow, intermediate, and deep levels of feature extractors, respectively.

The motivation behind intermediate-level attacks lies in the observation that different models may share similar representations within H_m. Thus, manipulating the features from the layer of H_m could enhance the transferability. Currently, the most advanced method in the field of intermediate-level attacks is ILPD [27], which applies the decay to the features from a specific intermediate layer. In fact, this can be seen as a form of adversarial idea when optimizing adversarial examples [43].

However, an intuitive and critical idea is: whether the transferability could further benefit from a form of multi-scale perturbation decays? To answer this question, we make the following hypothesis. First, since the adversarial perturbations can be amplified as the network deepens [30], simply applying perturbation decay to a single layer, like ILPD [27], may risk a potential case where the decayed perturbations

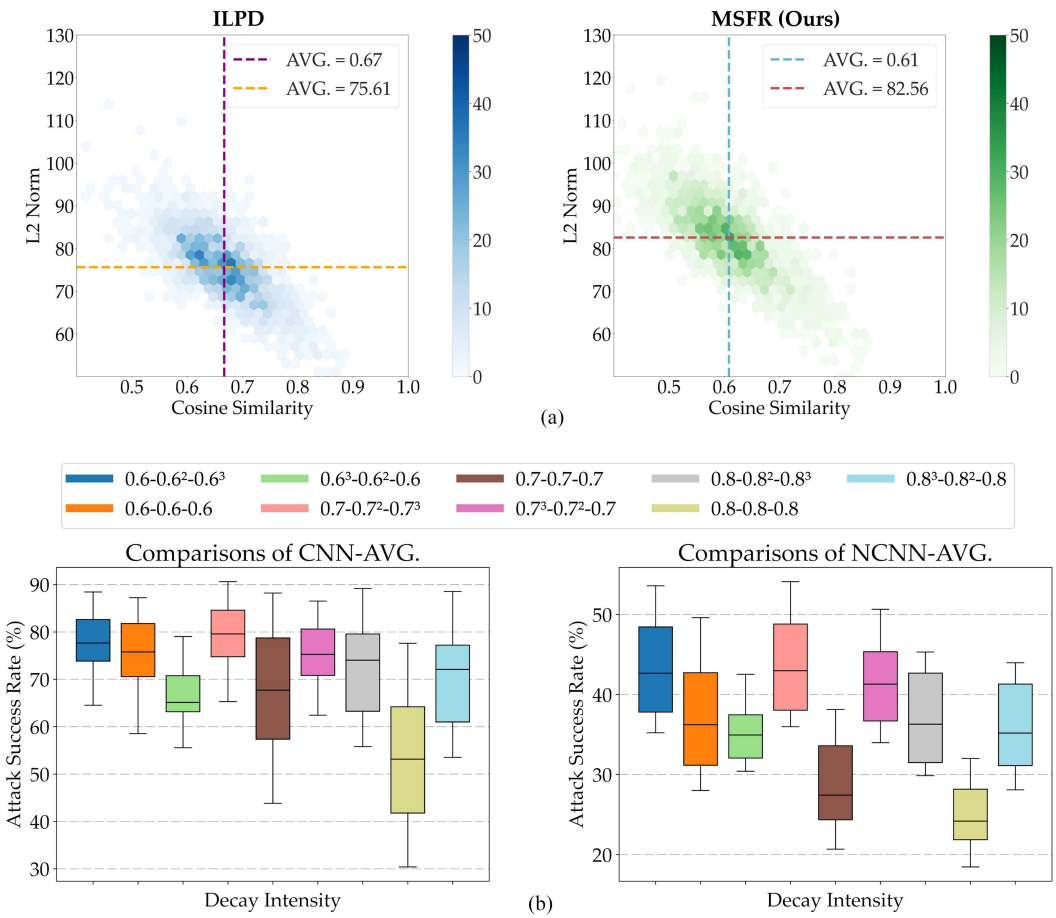

Figure 3 (Color online) Comparison results on ImageNet. (a) Comparisons regarding prediction distortions; (b) comparisons of different form of decay intensities. Each subplot in (a) reports the joint distribution of the angle offset (shown in the x-axis) and the distortion magnitude (shown in the y-axis). The means of these indicators are marked in legends, where smaller mean of cosine similarity and larger mean of perturbation magnitude indicate the better method. For subplot (b), the x-axis represents the form of decay intensities (shown in legend).

are amplified by subsequent layers, which might cause the perturbations to escape the constraint of decay operation. Thus, periodically applying the decay is necessary, thereby encouraging the forward propagation to be in the status of adversarial optimization [43]. As indicated by Figure 3(a), we can achieve higher intensity of distortions with larger angle offset and distortion magnitude. Besides, considering the fact that features at different levels usually contain unique levels of information. Therefore, utilizing different levels of features can facilitate the perturbations contain more levels of information, achieving a higher level of distortions. As proved by Figure 1(a), our MSFR drifts the spectrum attention regions with a higher intensity than ILPD, demonstrating that our MSFR contains a broader range of information distortions, ensuring the promising potential of such multi-scale approach.

Based on this insight, we conduct the corresponding ablation studies regarding the decay positions and periods to verify our hypothesis. Taking H_s as an example, where it can be decomposed as

$$H_s = h_1^s \circ h_2^s \circ \cdots \circ h_i^s \circ \cdots \circ h_n^s. \tag{6}$$

Here, h_i^s represents the i-th layer of H_s. Within our approach, we replace a specific layer h_i^k of H_k with $h_i'^k$, where $k \in \{s, m, d\}$ denotes the specific feature extractor. Taking $k = s$ as an example, $h_i'^s$ is defined in (7), where γ represents the decay intensity.

$$h_i'^s(\boldsymbol{x}) = \gamma h_i^s(\boldsymbol{x} + \boldsymbol{\delta}) + (1 - \gamma)\, h_i^s(\boldsymbol{x}) \quad \text{s.t.} \quad \|\boldsymbol{\delta}\|_p \leqslant \epsilon. \tag{7}$$

Theoretically, when the decay intensity increases, a greater proportion of the perturbed features are preserved, resulting in a weakened adversarial effect during forward propagation. Conversely, when the decay intensity decreases, the preservation of the perturbed features is reduced, leading to a higher level of adversarial status during forward propagation.

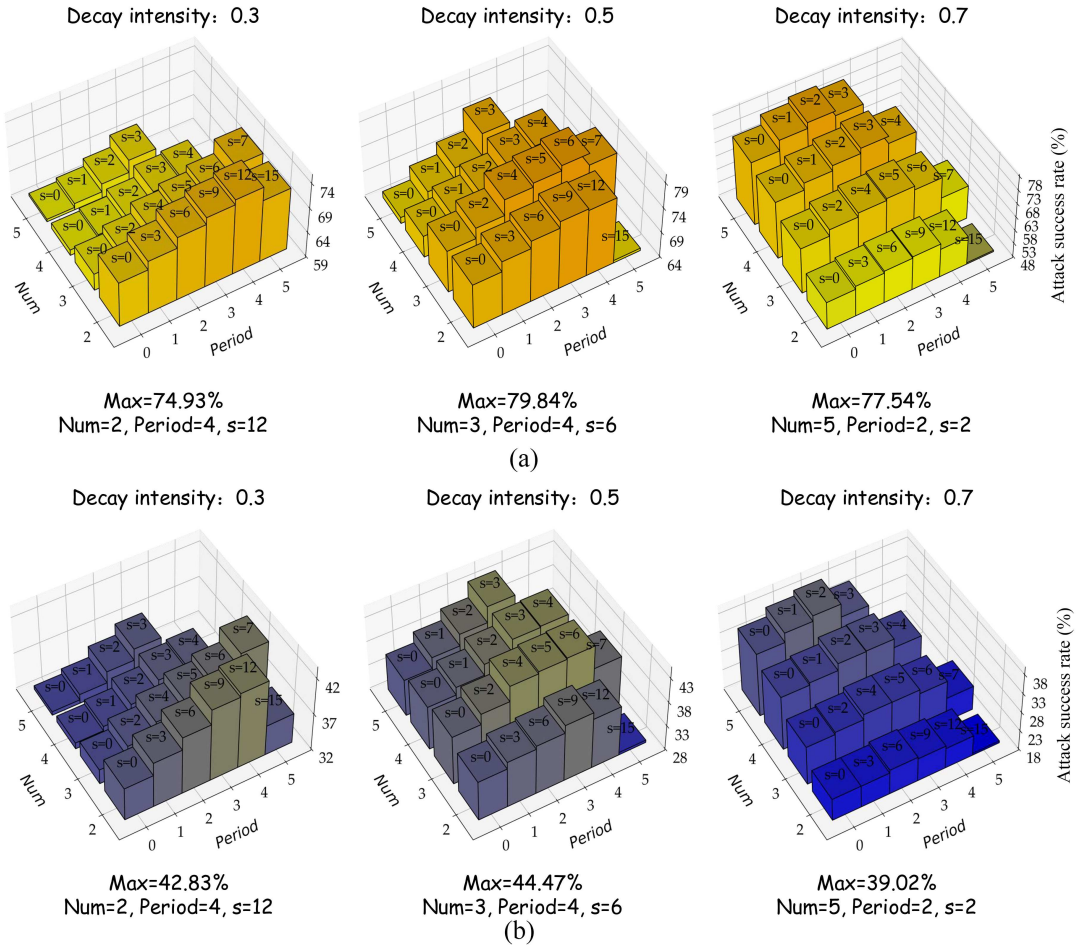

Figure 4 (Color online) Results on ImageNet regarding the application positions of decays. (a) Results of CNN-AVG.; (b) results of NCCN-AVG. The x-axis (Period) shows the decay period. Since the values of 'Period' are usually discontinuous, to format the results, we label the specific value of 'Period' on the top of each bar, which is denoted as 's'. The y-axis (Num) denotes the number of decays. The z-axis shows the attack success rate (%). The annotation on the bottom of each subplot indicates the maxima and the corresponding parameters.

To verify that the decays to the features at different levels can further enhance the transferability, we conduct the experiments and summarize the results in Figure 4[3)]. Here, our decays are performed at some specific layers, ranging from conv1 to layer-4-2, covering 17 layers in total, as specified in the official code[4)]. In this notation, layer-x-y stands for the y-th block of layer-x. Taking the second subplot in the top row of Figure 4 as an example, with the parameters `num=3`, `period=4`, and `s=6`, the three decays are applied to conv1, layer-2-3, and layer-4-0, respectively. In Figure 4, we can conclude that applying decay several times, with proper decay intensity and period, indeed can yield remarkable performance gains. Besides, higher decay strength (lower decay intensity) usually needs fewer decays. Moreover, a relatively larger decay period seems to always provide the optimal performance. We hypothesize that the underlying reason for this phenomenon lies in the insurmountable barriers introduced to the optimization of perturbations by sequentially applying perturbation decay to all the modules. As the network depth increases, the disruptive effects of the perturbations are progressively diminished, rendering the loss function incapable of providing effective gradients to guide perturbation updates. Consequently, this greedy approach (applying decay to all the modules) is less favorable compared to the periodic strategy.

Next, we discuss how the variations in decay intensities affect the transferability. Generally, the forms of decay intensities can be categorized into three types: progressively increasing, constant, and progressively decreasing. We believe that progressively increasing the decay intensities while ensuring a certain level

3) For simplicity, "CNN-AVG." refers to the average attack success rate (ASR, shown in %) over ten classical CNN models employed for our evaluations, and "NCNN-AVG." refers to the average ASR over six non-conventional-CNN models employed for our evaluations.

4) https://huggingface.co/timm/resnet50.tv_in1k.

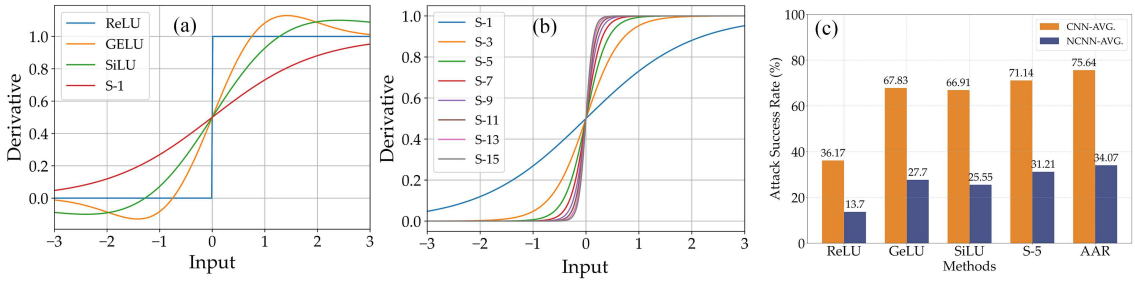

Figure 5 (Color online) Derivative curves and performances of different activations. (a) Derivatives of some activations; (b) derivatives of S-β with different β; (c) performances of different activations. For simplicity, we report the performance of S-β with the optimal choice, $\beta = 5$.

of initial intensity is a more effective way to enhance the transferability of adversarial examples. This hypothesis could be verified by the results in Figure 3(b), where progressively increasing the decay intensities exhibits the highest transferability. In contrast, using a stronger initial decay intensity or maintaining a constant decay intensity would not provide significant performance gains.

In fact, the progressively increasing decay intensities act as an information-filtering mechanism. It allows low-level details to be conveyed from the shallow layers while emphasizing high-level, integrated information at deeper layers. Yet, stronger decay in a shallow layer may cause excessive information loss. From an adversarial perspective, this implies that imposing an overly strong obstacle can prevent the model from effectively overcoming it, thus failing to maintain stable and high-quality gradients.

3.4 Adaptive activation rectification (AAR)

For conventional backward propagation, the gradient at each layer is defined as

$$\frac{\partial L}{\partial \boldsymbol{z}^i} = A(\boldsymbol{z}^i) \odot \left((\boldsymbol{W}^{i+1})^{\mathrm{T}} \frac{\partial L}{\partial \boldsymbol{z}^{i+1}} \right), \quad A(\boldsymbol{z}^i) = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{z}^i} \mathrm{ReLU}(\boldsymbol{z}^i) = \begin{cases} 1, & \text{if } \boldsymbol{z}^i > 0, \\ 0, & \text{if } \boldsymbol{z}^i \leqslant 0, \end{cases} \tag{8}$$

where $\boldsymbol{z}^i$ is the input to the i-th layer, $\frac{\partial L}{\partial \boldsymbol{z}^i}$ represents the gradient at the i-th layer, $A(\cdot)$ is the derivative of the activation, $\odot$ denotes the Hadamard product, and $\boldsymbol{W}^{i+1}$ is the weight matrix of layer $i+1$.

As shown in (8), the derivative of ReLU could truncate the gradient. However, the gradient truncated by ReLU might be precisely the critical information needed for crafting more transferable adversarial examples. As a result, this truncation suppresses the gradient flow of some inactive neurons at the current layer, leading to a form of sparse gradient, which could raise the great risk of dropping the required information for updating the gradients in the following layers. In contrast, a more complete gradient flow could allow the attacker to capture the vulnerabilities from various directions, ensuring that the attacker exploits potential weaknesses of the model during the backward propagation, thereby generating more transferable adversarial examples. Therefore, maintaining the gradient integrity is crucial for creating perturbations that can transfer across different models.

Along this line, LinBP [28] modifies the backward propagation by allowing a certain proportion of negative gradient flow, where it proposes to restore the gradients passed from the inactivated neurons by a proportion while retaining those gradients from the activated neurons. Subsequently, BPA [42] replaces the ReLU activation with the more continuous and smoother SiLU activation during the backward propagation. As observed, employing a more continuous and smoother activation during backward propagation may play a critical role in enhancing transferability. Accordingly, we plot the curves of the derivatives of different activations. As shown in Figure 5(a), an intuitive issue arises. Although the derivatives of these activations are more continuous and smoother than those of ReLU, they share a common problem: they are non-monotonic and can include negative values.

During backward propagation, the non-monotonicity of the derivative might cause the gradient map to lack the structure information of the beginning image. Specifically, if the derivative is non-monotonic, the gradient map could be inconsistent. Moreover, considering that for standard models, almost all of them employ the activation with non-negative property in its derivative. In this context, the presence of negative gradients might result in abrupt changes in the optimization direction. Therefore, we believe that a more optimal alternate would be one with monotonic, non-negative derivative.

Table 1 Performances of Softplus-β (S-β). We report the results (attack success rate, shown in %) of other activations and mark the best in bold and the second-best in underline.

Method	S-1	S-3	S-5	S-7	S-9	S-11	S-13	S-15	SiLU	AAR
CNN-AVG.	48.84	69.87	<u>71.14</u>	67.67	60.52	56.31	53.43	51.20	66.91	**75.64**
NCNN-AVG.	17.62	29.41	<u>31.21</u>	26.99	24.06	22.23	20.83	19.64	25.55	**34.07**

Although the motivation of Zhang et al. [44] is orthogonal to us, they have pointed out that the Softplus function could be an alternative to ReLU to improve the smoothness. Coincidentally, Softplus happens to have the two characteristics we emphasize above. To verify our hypothesis, we perform preliminary explorations regarding the choice of activation during backward propagation. Here, we extend Softplus to Softplus-β (denoted by S-β) for a more general choice. Its derivative is formulated in

$$A'(\boldsymbol{z}_i) = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{z}^i}\text{Softplus-}\beta(\boldsymbol{z}^i) = \frac{\mathrm{e}^{\beta \boldsymbol{z}^i}}{1+\mathrm{e}^{\beta \boldsymbol{z}^i}}, \tag{9}$$

where $A'(\cdot)$ is the derivative of S-β, and β is a manually pre-defined and fixed hyperparameter.

However, we find that not all the choices of β for S-β could yield better performances, despite its advantageous properties of preserving gradient integrity and being monotonic and non-negative. In fact, its performance is even inferior to that of SiLU when $\beta = 1$ (see Table 1 for reference). Comparing the derivatives of S-1 and SiLU, an intuitive reason for this discrepancy could be attributed to the fact that the second derivative of S-1 owns a relatively smaller degree of magnitude near zero than that of SiLU. In other words, the absolute value of the second derivative of the function near the zero point should lie within an appropriate range-neither too small nor too large, and a relatively higher magnitude of the second derivative near zero might play a key role in enhancing transferability. Such a characteristic ensures that the function can respond quickly to small input changes. As pointed out by Table 1, when we enlarge β, its performance could be enhanced to some extent. Nevertheless, increasing β greedily is also an undesirable choice, as larger β could cause S-β to gradually approach the behavior of ReLU.

Despite this, the optimal choice of β in S-β still needs intensive manual efforts to be decided. Besides, this fixed value may not adapt well to the varying characteristics of different layers. Specifically, since the architecture and parameters differ across layers, the same β value may fail to accommodate the range of feature magnitudes across layers, leading to suboptimal gradient adjustments. For example, different layers may require varying sensitivity to second-order derivatives near zero or different levels of reactivation for values previously truncated by ReLU. To this end, we propose AAR, which adjusts the β of each layer based on its minimum input.

This dynamic adjustment strategy helps the attacker more accurately identify and amplify potential weaknesses in each layer, realizing more aggressive and transferable attacks. Within this context, to ensure that all the gradients could pass through the network, we set a small threshold, thereby maintaining the integrity of the gradient, as shown in

$$\tau = \frac{1}{1+\mathrm{e}^{-\beta(\mu+\zeta)}}, \quad \beta = -\frac{\ln\left(\frac{1-\tau}{\tau}+\zeta\right)}{\mu+\zeta}. \tag{10}$$

Here, τ is the critical threshold, ζ is a small positive value, e.g., 1×10^{-10}, and μ is the minimum input.

As we can see in Figure 5(c), the preliminary results clearly demonstrate that our AAR outperforms ReLU, GeLU, SiLU, and the best value of S-β. Specifically, AAR not only ensures that the derivative of the activation is non-negative and monotonic but also maintains the gradient integrity. Moreover, its second derivative near zero has a certain degree of magnitude. Last but not least, it takes into account the diverse characteristics of the features from different layers, thereby enhancing the transferability.

4 Experiments

4.1 Experimental settings

(1) Datasets. We evaluate on two commonly utilized datasets: CIFAR-10 [45] and ImageNet [46]. For the CIFAR-10 dataset, we employ the entire test dataset for evaluations. For the ImageNet dataset, we randomly sample 5000 examples from the validation set, as suggested by [27, 28, 41, 42].

Table 2 Results in ASR (%) on ImageNet dataset using ResNet-50 as the surrogate. The perturbation budget $\epsilon = 8$ and we run 100 attack iterations. The best results are shown in bold, the same as the other tables.

Method	LinBP [28]	SGM [29]	IAA [60]	BPA [42]	Admix [34]	FIA [26]	TAIG [39]	NAA [40]	GRA [37]	PGN [36]	ILA++ [41]	ILPD [27]	Ours
Res-101 [5]	68.80	68.33	77.47	81.69	71.69	53.08	70.86	73.74	74.11	76.91	84.74	82.58	**94.40**
VGG-16 [7]	82.59	75.99	91.25	88.90	79.71	71.90	83.52	81.65	83.08	86.31	87.82	84.89	**96.56**
Den-121 [8]	85.30	80.59	90.67	91.63	89.65	73.16	90.95	85.08	87.60	89.10	92.24	89.55	**97.75**
Inc-v4 [4]	48.12	44.13	55.46	57.35	57.31	45.70	68.49	66.46	62.09	62.70	68.01	69.85	**88.93**
Inc-v3 [2]	50.14	47.92	59.90	56.67	60.40	48.92	71.31	68.29	63.63	66.43	68.36	70.65	**87.80**
IncRes [4]	35.86	35.86	41.58	41.04	48.19	36.06	57.46	57.81	53.58	54.06	52.90	61.36	**79.35**
WRN-50 [47]	67.11	68.26	76.16	79.29	67.31	48.62	64.97	71.96	72.02	74.65	83.43	80.07	**92.47**
Mob-v2 [49]	70.91	72.60	83.18	80.83	78.19	63.40	81.81	79.63	80.75	83.64	86.47	85.61	**96.26**
SERes-101 [52]	51.74	59.48	56.01	70.18	61.25	43.49	66.33	69.22	67.19	69.97	76.12	76.12	**90.65**
PNASNet [53]	43.41	44.62	51.10	55.05	57.99	41.47	68.83	65.35	67.22	68.76	68.76	73.61	**88.63**
CNN-AVG.	60.40	59.78	68.28	70.26	67.17	52.58	72.45	71.92	71.13	73.25	76.89	77.43	**91.28**
ViT-B [54]	15.69	22.03	19.95	19.17	23.31	14.77	31.75	30.57	30.14	30.32	28.14	32.05	**45.79**
Swin-B [55]	18.63	23.46	20.05	23.55	20.05	15.62	26.06	34.53	25.37	24.72	35.71	35.37	**48.53**
Deit3 [56]	33.07	36.05	41.53	37.41	34.60	28.63	37.77	47.02	40.63	42.86	51.08	50.76	**69.58**
Beit-B [57]	16.48	20.49	19.93	19.58	22.32	14.56	28.48	32.04	30.01	28.82	30.87	33.45	**48.06**
MlpMix-B [58]	24.50	31.64	30.91	25.84	30.73	21.32	37.20	39.03	38.79	39.22	38.30	42.43	**58.28**
ConvNeXt-B [59]	30.54	33.59	35.28	44.35	34.82	26.86	41.84	46.81	42.95	43.77	52.19	53.54	**69.63**
NCNN-AVG.	23.15	27.88	27.94	28.32	27.64	20.29	33.85	38.33	34.65	34.95	39.38	41.27	**56.65**

(2) Victim models. On CIFAR-10 dataset, the involved victims are VGG-19-BN (VGG-19) [7], Wide-ResNet-28-10 (WRN) [47], ResNet-18 (Res-18) [5], ResNeXt-29 (ResNeXt) [48], DenseNet-BC (Den-BC) [8], Inception-V3 (Inc-v3) [2] and MobileNet-V2 (Mob-v2) [49]. Additionally, we utilize two advanced models: a 272-layer PyramidNet [50] (Pyrm-Net) and GDAS [51]. On ImageNet dataset, we leverage ten classical victim models, including ResNet-101 (Res-101) [5], VGG-16 [7], DenseNet-121 (Den-121) [8], Inception-V4 (Inc-v4) [4], Inception-V3 (Inc-v3) [2], Inception-ResNet-V2 (IncRes) [4], Wide-ResNet-50 (WRN-50) [47], MobileNet-V2 [49], SE-ResNeXt-101-32x4d (SERes-101) [52], and PNASNet [53]. Besides, we introduce six victim models with non-conventional-CNN architectures, encompassing ViT-B [54], Swin-B [55], DeiT3 [56], BEiT-B [57], MLP-Mixer-B (MlpMix-B) [58], and ConvNeXt-B [59].

(3) Attack settings. Our attack operates under the black-box scenario within the ℓ_∞ norm constraint. For CIFAR-10, we generate adversarial examples using VGG-19 as the surrogate model, with the perturbation budget of $\epsilon = 4/255$. For ImageNet, ResNet-50 is served as the surrogate, with the perturbation budget of $\epsilon = 8/255$. We run 100 iterations for all attack methods on both datasets, using a step size of $1/255$ for fair comparisons. For all the peer methods included in our comparisons, we follow their suggested hyperparameters and official codes. The evaluation metric is the ASR.

(4) Hyperparameters and settings. Unless otherwise specified, the experimental hyperparameter settings in this paper are as follows. For our PRA with ResNet-50 as the surrogate model, the decay positions in MSFR are set to conv1, layer-2-3, and layer-3-5, corresponding to the shallow, intermediate, and deep layers, respectively. When using VGG-19 as the surrogate, the decay positions in MSFR are layer-0, layer-10, and layer-23. The decay intensities are set to 0.7, 0.7^2, and 0.7^3, respectively. For our AAR, the critical threshold τ in (10) for our adaptive Softplus-β is 1×10^{-9}.

4.2 Peer comparisons

We first comprehensively compare our PRA on ImageNet dataset with several advanced approaches, including Admix [34] that improves from the perspective of input transformation. We also compare with GRA [37] and PGN [36], which are developed from the perspective of gradient stabilization. Additionally, we introduce some methods with advanced objectives such as FIA [26], TAIG [39], and NAA [40]. Furthermore, we compare with several surrogate refinement methods, including LinBP [28], SGM [29], the novel and powerful IAA [60], BPA [42], ILA++ [41], and the most related transferable attack, ILPD [27]. The experimental results in Table 2 [2, 4, 5, 7, 8, 26–29, 34, 36, 37, 39–42, 47, 49, 52–60] indicate that the transferability of the adversarial examples generated by our PRA significantly outperforms all the other methods under the same conditions, achieving the ASR of 91.28% regarding CNN-AVG. and the ASR of 56.65% regarding NCNN-AVG., with absolute gains of +13.85% and +15.38% over the second-best method, ILPD, respectively.

For the results on CIFAR-10 dataset, as detailed in Table 3 [2,5,7,8,26–28,31,34,36,37,39–41,47–51], we evaluate our PRA against several approaches, including Admix [34], SSM [31], GRA [37], PGN [36], LinBP [28], FIA [26], TAIG [39], NAA [40], ILA++ [41], and ILPD [27]. The results show that PRA also achieves remarkable results on CIFAR-10, with the average ASR of 59.03%, surpassing the second-best method PGN by +5.5%.

Table 3 Results in ASR (%) on CIFAR-10 dataset, with VGG-19 as the surrogate. We set $\epsilon = 4$ and run 100 iterations.

Method	LinBP [28]	Admix [34]	SSM [31]	FIA [26]	TAIG [39]	NAA [40]	GRA [37]	PGN [36]	ILA++ [41]	ILPD [27]	Ours
VGG-19 [7]	64.93	51.67	71.05	53.80	49.23	58.27	67.88	71.58	67.22	68.32	**74.29**
WRN [47]	56.20	43.99	58.31	42.43	38.21	50.15	57.24	61.57	57.91	60.16	**67.52**
Res-18 [5]	29.13	20.12	27.60	19.12	16.96	24.78	27.99	31.29	31.02	30.09	**37.38**
ResNeXt [48]	58.10	47.52	58.20	45.80	42.03	53.12	57.88	61.47	58.63	61.04	**69.31**
Den-BC [8]	55.06	44.93	54.51	43.88	39.51	51.74	54.68	59.20	55.93	58.67	**66.13**
Inc-v3 [2]	58.93	46.52	65.14	47.16	44.73	52.23	61.19	65.02	60.62	63.11	**70.97**
Mob-v2 [49]	63.40	52.83	69.71	57.24	52.76	58.44	67.07	70.85	66.34	67.96	**73.78**
Pyrm-Net [50]	15.88	11.32	13.77	10.53	8.29	15.96	12.79	14.58	15.53	15.47	**20.35**
GDAS [51]	40.80	30.99	41.90	34.96	29.08	38.26	42.81	46.21	42.48	46.93	**51.53**
AVG.	49.16	38.88	51.13	39.44	35.64	44.77	49.95	53.53	50.63	52.42	**59.03**

Table 4 Ablation results (attack success rate, shown in %) on the contributions of our MSFR.

Method	ILA++ [41]	ILPD [27]	MSFR
CNN-AVG.	76.89	77.43	**80.51**
NCNN-AVG.	39.38	41.27	**42.48**

Table 5 Ablation results (attack success rate, shown in %) on the optimal positions of our MSFR. Period-x-y denotes the intervals among the three decays, where x represents the interval between the first two decays, and y represents the interval between the last two decays.

Method	Period-5-5	Period-6-6	Period-5-6	Period-6-5
CNN-AVG.	90.49	90.19	91.01	**91.28**
NCNN-AVG.	56.52	55.86	56.55	**56.65**

4.3 Ablation studies

4.3.1 *A closer look at MSFR*

Here, we conduct ablation studies to evaluate the effectiveness of our MSFR on forward propagation rectification. First, we compare MSFR with other existing methods to demonstrate its significant performance when used independently. Further, we investigate the optimal decay positions.

(1) Methodology ablation study. As shown in Table 4 [27,41], we conduct comparisons with several peer attacks, which aim at refining the forward propagation, such as ILA++ [41], and the latest ILPD [27], over ten classical CNN models and six non-conventional-CNN models. Based on these results, we can draw the following conclusions. Comparing ILPD with ILA++, it is clear that using an adversarial strategy to decay the perturbation effectively rectifies forward propagation. Moreover, comparing MSFR with ILPD shows that our MSFR approach, which employs multi-scale rectification, exhibits greater transferability than applying the perturbation decay solely to a single intermediate layer. In particular, this allows the attacker to explore a wider range of potential pathways when generating adversarial examples, which might be missed with single-layer decay. Overall, our MSFR achieves remarkable results among peers.

(2) Hyperparameter ablation study. Previously, in Subsection 3.3, we conduct preliminary explorations on the decay positions, and forms of intensities of our MSFR. In ILA++ [41] and ILPD [27], a common experimental conclusion is that adjusting the forward propagation later than layer-4-0 will have a negative impact on the final performance improvement. To further explore whether this view is still valid for our MFSR, we conduct further ablation experiments. Although the final differences were not substantial, the optimal positions in shallow, intermediate, and deep layers shown in Table 5 are conv1, layer-2-3, and layer-3-5, respectively. Among them, Period-6-6 involves layer-4-0, with relatively the worst results. This implies that the decay intervals should be slightly larger and should not beyond layer-4-0. By avoiding the application of decay to layers deeper than layer-4-0, we can maintain the effectiveness and ensure that the decay does not adversely affect the deeper layers of the network. Thus, we recommend conv1, layer-2-3, and layer-3-5 as the default choice.

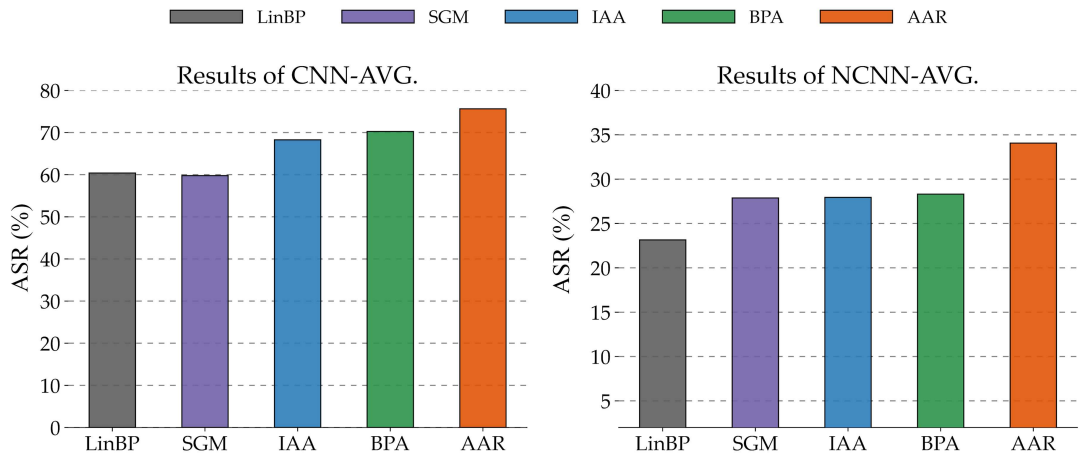

Figure 6 (Color online) Results in ASR (y-axis, %) of our AAR and other peer methods.

Table 6 Ablation results (attack success rate, shown in %) on the critical threshold τ.

Method	1×10^{-5}	1×10^{-6}	1×10^{-7}	1×10^{-8}	1×10^{-9}	1×10^{-10}	1×10^{-11}	1×10^{-12}	1×10^{-13}
CNN-AVG.	89.10	90.29	90.99	91.13	**91.28**	91.10	90.86	90.68	90.38
NCNN-AVG.	52.09	54.52	55.99	56.37	**56.65**	56.52	55.61	55.58	54.88

4.3.2 *A closer look at AAR*

We further perform ablation experiments to evaluate the effectiveness of our AAR, which is specifically designed for backward propagation rectification. Firstly, we compare AAR with other existing methods tailored for backward propagation rectification to demonstrate its significant performance when used independently. Then, we perform ablations on the critical threshold τ of our AAR.

(1) Methodology ablation study. Here, we conduct comparative experiments with other methods on backward propagation rectification. Our comparisons include LinBP [28], SGM [29], and the novel and powerful IAA [60] and BPA [42]. As shown in Figure 6, our AAR significantly outperforms all the other peer methods over both classical CNN and non-conventional-CNN models. This fully demonstrates the superior performance of our AAR. Moreover, it supports our view that, during the backward propagation, it is essential not only to consider the non-negativity, monotonicity, and gradient integrity, but also to encourage the second derivative to own a certain degree of magnitude near zero. Moreover, considering the diverse characteristics across different layers is critical for further performance gains. This ensures the excellent transferability of the generated adversarial examples.

(2) Hyperparameter ablation study. As described in Subsection 3.4, when generating adaptive β for different layers, we set a small threshold to allow all the gradients to pass through the model, thereby maintaining the integrity of the gradients. To this end, our ablation experiments on the threshold τ, shown in Table 6, indicate that a smaller threshold will prematurely truncate the gradient, while a larger threshold will fail to ensure the magnitude degree of the second derivative near zero. Ultimately, we verify that the optimal threshold is 1×10^{-9}, which we select as the critical threshold τ for our AAR.

4.4 Extended evaluations

(1) Scalability in terms of the architecture. We first evaluate the performance of our PRA when applied to surrogate models with different architectures. For ease of comparison, based on the findings in Table 2, we select the second-best methods, ILPD, and NAA, as performance benchmarks. In terms of surrogate model architectures, we choose VGG-19, Den-121, and ResNeXt-50 (RXT-50) for our experiments. The parameters of PRA in these models are similar to those in Table 2, except for the location of the application of MSFR[5)]. As shown in Table 7 [7,8,27,40,48], our PRA consistently outperforms the other methods across all the surrogate models, particularly when paired with Den-121. This demonstrates the scalability of our PRA in terms of surrogate model architectures.

5) The specific settings are available in our code.

Table 7 Results in ASR (%) on ImageNet with different surrogates. The evaluation settings are the same as Table 2.

Surrogate	VGG-19 [7]			Den-121 [8]			RXT-50 [48]		
Method	NAA [40]	ILPD [27]	Ours	NAA [40]	ILPD [27]	Ours	NAA [40]	ILPD [27]	Ours
CNN-AVG.	57.57	67.68	**72.16**	65.07	74.99	**90.22**	70.16	78.52	**90.25**
NCNN-AVG.	31.52	35.30	**39.07**	36.12	39.58	**60.58**	37.73	44.14	**60.17**

Table 8 Results (attack success rate, shown in %) over typical defenses on ImageNet. We set $\epsilon = 8$, total iteration as 100, and ResNet-50 as the surrogate.

Method	LinBP [28]	SGM [29]	IAA [60]	BPA [42]	Admix [34]	FIA [26]	TAIG [39]	NAA [40]	GRA [37]	PGN [36]	ILA++ [41]	ILPD [27]	Ours
HGD [61]	46.03	43.23	55.30	54.04	51.63	43.03	52.77	64.71	58.04	58.84	53.70	69.65	**85.99**
JPEG [62]	27.59	26.10	30.05	32.97	40.69	32.42	47.34	49.96	49.10	49.17	34.41	54.31	**71.12**
NRP [63]	28.50	26.59	29.49	30.09	35.49	30.81	41.86	39.23	38.67	38.95	30.65	37.94	**42.71**
SIN [64]	44.23	40.69	56.33	48.51	56.42	51.68	59.22	61.36	59.31	61.17	48.98	64.43	**82.77**
R-Inc-V3	25.83	24.75	29.89	31.17	34.28	26.64	41.31	47.67	42.60	42.26	33.27	50.91	**69.57**
CNS [65]	4.80	5.49	5.23	4.62	7.93	6.63	8.11	8.54	8.68	8.42	5.23	6.45	**8.81**
AVG.	29.50	27.81	34.38	33.57	37.74	31.87	41.77	45.25	42.73	43.14	34.37	47.28	**60.16**

Table 9 Results in ASR (%) on ImageNet. We set $\epsilon = 12$ with 100 iterations using ResNet-50 as the surrogate.

Method	LinBP [28]	SGM [29]	IAA [60]	BPA [42]	Admix [34]	FIA [26]	TAIG [39]	NAA [40]	GRA [37]	PGN [36]	ILA++ [41]	ILPD [27]	Ours
Res-101 [5]	65.76	83.88	93.90	93.52	78.59	81.33	83.94	88.55	86.75	88.99	88.04	93.34	**98.38**
VGG-16 [7]	77.14	89.92	98.57	96.56	83.94	90.68	94.20	91.76	92.33	93.55	91.98	95.20	**99.14**
Den-121 [8]	84.30	91.83	97.82	97.75	92.30	92.30	95.78	94.14	94.69	95.50	93.33	96.32	**99.32**
Inc-v4 [4]	50.33	60.04	79.44	78.20	61.20	70.83	80.38	82.65	78.35	77.32	73.82	86.47	**96.38**
Inc-v3 [2]	54.94	62.70	79.27	76.54	62.98	72.53	82.21	82.86	79.63	76.47	73.74	86.08	**95.27**
IncRes [4]	42.90	50.99	60.40	62.24	50.92	59.58	69.94	74.30	71.71	69.53	60.53	80.91	**92.77**
WRN-50 [47]	62.51	84.13	92.60	94.10	76.60	77.48	80.29	86.57	85.76	87.20	87.70	90.96	**98.37**
Mob-v2 [49]	77.31	85.74	96.25	94.35	83.71	87.84	90.74	91.59	91.20	91.46	89.49	94.29	**98.95**
SERes-101 [52]	55.38	75.85	78.95	87.74	70.61	70.86	78.27	82.93	81.42	83.31	79.03	90.97	**96.97**
PNASNet [53]	51.77	61.07	74.05	76.79	65.02	68.09	81.47	81.47	81.67	81.00	73.58	88.09	**96.59**
CNN-AVG.	62.23	74.62	85.13	85.78	72.59	77.15	83.72	85.68	84.35	84.43	81.12	90.26	**97.21**
ViT-B [54]	26.68	33.37	33.07	32.66	27.88	30.63	44.38	48.38	44.72	41.67	32.66	54.27	**64.96**
Swin-B [55]	23.13	35.01	33.45	37.39	23.82	26.87	32.95	50.09	35.37	33.39	38.71	56.68	**65.37**
Deit3 [56]	35.48	52.20	63.35	59.76	39.90	52.80	52.89	66.61	58.89	56.42	56.27	72.71	**85.06**
Beit-B [57]	23.88	32.70	33.89	34.87	26.91	28.40	42.03	47.73	43.62	42.12	34.51	56.24	**69.37**
MlpMix-B [58]	33.48	48.01	50.95	47.04	34.33	43.13	50.89	57.48	55.04	51.50	44.96	64.08	**77.52**
ConvNeXt-B [59]	34.41	47.28	56.11	64.35	41.84	48.33	58.51	64.31	60.21	59.63	56.11	73.73	**84.90**
NCNN-AVG.	29.51	41.43	45.14	46.01	32.45	38.36	46.94	55.77	49.64	47.46	43.87	62.95	**74.53**

(2) Transferability against defense methods. Here, we select several representative defense methods, including HGD [61], JPEG [62], NRP [63], SIN [64], robust Inception-v3 (R-Inc-V3) from Timm[6] repository, and the recent proposed CNS [65]. As shown in Table 8 [26–29, 34, 36, 37, 39–42, 60–65], our PRA outperforms all the other attacks, which fully proves that the adversarial examples generated by PRA not only transfer well to normally trained victim models, but also maintain their superiority when facing defense methods.

(3) Transferability under higher ℓ_∞ norm constraint. When we set the ℓ_∞ norm $\epsilon = 12/255$, as shown in Table 9 [2, 4, 5, 7, 8, 26–29, 34, 36, 37, 39–42, 47, 49, 52–60], the PRA achieves an impressive average ASR of 97.21% over ten classical CNN models, outperforming the second-best method by +6.95%. Surprisingly, the average ASR achieved by our PRA is almost close to 99% on classical CNN models, far exceeding ILPD and all other peer methods. Meanwhile, over six non-conventional-CNN models, we still could achieve an unprecedented average ASR of 74.53%, surpassing the second-best method by +11.58%. Through the above analysis, it appears that our PRA can still maintain its superior performance even under relatively loose visual constraints.

(4) Transferability with fewer iterations. We then reduce the number of attack iterations from 100 to 20. The experimental results are shown in Table 10 [2,4,5,7,8,26–29,34,36,37,39–42,47,49,52–60], where we can find that even with just 20 iterations, our PRA still could achieve the average ASR of 87.78% over ten classical CNN models and the average ASR of 54.33% over six non-conventional-CNN models. These results are significantly higher than the second-best method ILPD by +15.60% and +18.33%, respectively. Compared to the results under 100 attack iterations, our performances regarding CNN-AVG. and NCNN-AVG. only decrease by 3.50% and 2.32%, respectively, whereas the ILPD sees larger decreases of 5.25% and 5.27%. This demonstrates that the adversarial examples of our PRA maintain strong transferability and experience smaller performance drops with fewer iterations, illustrating the

6) https://github.com/huggingface/pytorch-image-models.

Table 10 Results in ASR (%) on ImageNet. We set $\epsilon = 8$, the surrogate is ResNet-50, and we run 20 iterations here.

Method	LinBP [28]	SGM [29]	IAA [60]	BPA [42]	Admix [34]	FIA [26]	TAIG [39]	NAA [40]	GRA [37]	PGN [36]	ILA++ [41]	ILPD [27]	Ours
Res-101 [5]	45.27	64.65	66.22	77.21	57.81	48.48	64.63	72.06	73.74	69.76	74.60	77.16	**90.72**
VGG-16 [7]	55.30	72.90	77.71	86.60	66.24	66.38	79.66	78.71	81.58	79.07	80.09	80.14	**93.91**
Den-121 [8]	67.67	78.20	76.47	90.13	78.47	68.53	86.18	83.31	87.40	86.17	85.43	85.56	**95.85**
Inc-v4 [4]	34.59	43.24	51.49	56.46	42.76	38.73	60.22	62.91	64.21	58.61	56.73	62.77	**83.39**
Inc-v3 [2]	40.82	47.85	52.30	56.03	46.20	43.76	64.78	65.42	66.21	60.76	58.39	65.78	**83.00**
IncRes [4]	28.15	36.67	35.84	41.45	35.51	30.06	49.63	54.40	56.44	47.92	43.29	56.17	**75.39**
WRN-50 [47]	42.12	64.87	63.81	74.95	52.82	44.10	56.75	68.70	69.82	66.94	73.07	72.90	**88.76**
Mob-v2 [49]	58.44	71.09	70.26	79.71	65.64	57.56	78.14	76.81	80.49	77.66	77.35	80.81	**94.35**
SERes-101 [52]	40.30	59.17	48.26	67.40	51.14	39.44	56.29	66.25	69.72	65.23	63.87	72.63	**87.62**
PNASNet [53]	35.52	45.95	44.75	55.45	45.15	37.06	60.00	62.54	69.90	63.14	57.39	67.89	**84.82**
CNN-AVG.	44.82	58.46	58.71	68.54	54.17	47.41	65.63	69.11	71.95	67.53	67.02	72.18	**87.78**
ViT-B [54]	16.36	21.78	15.93	18.93	17.51	14.15	25.52	28.68	31.73	26.54	20.21	28.98	**45.79**
Swin-B [55]	17.44	26.39	20.52	25.94	16.88	13.82	19.95	32.91	28.01	24.96	27.08	32.44	**47.99**
Deit3 [56]	22.41	34.18	26.92	34.82	23.33	25.68	32.47	42.86	39.72	35.20	40.12	43.40	**64.10**
Beit-B [57]	14.99	21.00	14.46	20.30	15.16	12.95	22.99	29.59	30.55	26.01	22.27	27.80	**46.63**
MlpMix-B [58]	21.32	30.18	20.75	25.78	20.65	18.14	31.40	35.61	38.85	32.74	29.51	35.25	**56.51**
ConvNeXt-B [59]	21.42	33.70	27.44	42.13	25.69	23.58	36.40	44.76	43.48	38.15	40.84	48.10	**64.95**
NCNN-AVG.	18.99	27.87	21.00	27.98	19.87	18.05	28.12	35.74	35.39	30.60	30.01	36.00	**54.33**

robustness and efficiency of our approach.

5 Conclusion

In this paper, we studied the refinements on the surrogate model for transferable attacks and proposed a bidirectional method, PRA, which incorporates both the forward and backward propagation rectifications. Specifically, during forward propagation, we developed MSFR that applies feature rectifications to different levels of features and encourages the entire forward propagation to maintain a proper adversarial optimization state, mitigating the issue of incomplete feature utilization when applying decay to a single intermediate layer. MSFR highlights the necessity and benefits of multi-scale decays for enhancing transferability, indicating its promising effects in causing larger distortions. On backward propagation rectification, we concluded a more feasible insight regarding the characteristics of the alternative activation, i.e., maintaining the gradient integrity, encouraging the non-negativity and monotonicity of the activation derivatives, and ensuring a certain level of magnitude of its second derivative near zero. Moreover, we further proposed AAR to adaptively consider the specificity of the features. Extensive evaluations strongly demonstrate our remarkable superiority. We hope this research will inspire further interest in the study of propagation rectifications within the field of adversarial attacks.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant No. 62376223), Fundamental Research Funds for the Central Universities, and Cultivation Foundation for Excellent Doctoral Dissertation of the School of Automation, Northwestern Polytechnical University.

References

1 Yu Y, Ji Z, Han J, et al. Episode-based prototype generating network for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 14035–14044

2 Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2818–2826

3 Peng C L, Wang B, Liu D C, et al. Pyramid-resolution person restoration for cross-resolution person re-identification. Sci China Inf Sci, 2024, 67: 169101

4 Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-V4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2017. 4278–4284

5 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778

6 Cheng G, Yuan X, Yao X, et al. Towards large-scale small object detection: survey and benchmarks. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 13467–13488

7 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556

8 Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 4700–4708

9 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. 2014. ArXiv:1412.6572

10 Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. 2013. ArXiv:1312.6199

11 Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 9185–9193

12 Ji Z, Yan J T, Wang Q, et al. Triple discriminator generative adversarial network for zero-shot image classification. Sci China Inf Sci, 2021, 64: 120101

13 Sun X, Cheng G, Li H, et al. Task-specific importance-awareness matters: on targeted attacks against object detection. IEEE Trans Circ Syst Video Technol, 2024, 34: 11619–11629

14 He X, Zhu M R, Wang N N, et al. BiTGAN: bilateral generative adversarial networks for Chinese ink wash painting style transfer. Sci China Inf Sci, 2023, 66: 119104

15 Wei X S, Xu S L, Chen H, et al. Prototype-based classifier learning for long-tailed visual recognition. Sci China Inf Sci, 2022, 65: 160105

16 Sun X, Cheng G, Li H, et al. Exploring effective data for surrogate training towards black-box attack. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022. 15355–15364

17 Wei X S, Cui Q, Yang L, et al. RPC: a large-scale and fine-grained retail product checkout dataset. Sci China Inf Sci, 2022, 65: 197101

18 Sun X, Cheng G, Li H, et al. STDatav2: accessing efficient black-box stealing for adversarial attacks. IEEE Trans Pattern Anal Mach Intell, 2025, 47: 2429–2445

19 Yuan X, Cheng G, Yan K, et al. Small object detection via coarse-to-fine proposal generation and imitation learning. In: Proceedings of the IEEE International Conference on Computer Vision, 2023. 6317–6327

20 Lai P, Cheng G, Zhang M, et al. NCSiam: reliable matching via neighborhood consensus for siamese-based object tracking. IEEE Trans Image Process, 2023, 32: 6168–6182

21 Lai P, Zhang M, Cheng G, et al. Target-aware transformer for satellite video object tracking. IEEE Trans Geosci Remote Sens, 2024, 62: 1–10

22 Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. In: Proceedings of the International Conference on Learning Representations, 2018. 1–10

23 Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: attacks and defenses. In: Proceedings of the International Conference on Learning Representations, 2018. 1–14

24 Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: reliable attacks against black-box machine learning models. 2017. ArXiv:1712.04248

25 Yan Z, Guo Y, Liang J, et al. Policy-driven attack: learning to query for hard-label black-box adversarial examples. In: Proceedings of the International Conference on Learning Representations, 2021. 1–11

26 Wang Z, Guo H, Zhang Z, et al. Feature importance-aware transferable adversarial attacks. In: Proceedings of the IEEE International Conference on Computer Vision, 2021. 7639–7648

27 Li Q, Guo Y, Zuo W, et al. Improving adversarial transferability via intermediate-level perturbation decay. In: Proceedings of the Advances in Neural Information Processing Systems, 2024. 1–12

28 Guo Y, Li Q, Chen H. Backpropagating linearly improves transferability of adversarial examples. In: Proceedings of the Advances in Neural Information Processing Systems, 2020. 85–95

29 Wu D, Wang Y, Xia S T, et al. Skip connections matter: on the transferability of adversarial examples generated with ResNets. In: Proceedings of the International Conference on Learning Representations, 2020. 1–12

30 Inkawhich N, Wen W, Li H H, et al. Feature space perturbations yield more transferable adversarial examples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 7066–7074

31 Long Y, Zhang Q, Zeng B, et al. Frequency domain model augmentation for adversarial attack. In: Proceedings of European Conference on Computer Vision, 2022. 549–566

32 Dong Y, Pang T, Su H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 4312–4321

33 Xie C, Zhang Z, Zhou Y, et al. Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 2730–2739

34 Wang X, He X, Wang J, et al. Admix: enhancing the transferability of adversarial attacks. In: Proceedings of the IEEE International Conference on Computer Vision, 2021. 16158–16167

35 Lin J, Song C, He K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks. In: Proceedings of the International Conference on Learning Representations, 2020. 1–11

36 Ge Z, Liu H, Wang X, et al. Boosting adversarial transferability by achieving flat local maxima. In: Proceedings of the Advances in Neural Information Processing Systems, 2023. 70141–70161

37 Zhu H, Ren Y, Sui X, et al. Boosting adversarial transferability via gradient relevance attack. In: Proceedings of the IEEE International Conference on Computer Vision, 2023. 4741–4750

38 Huang Q, Katsman I, He H, et al. Enhancing adversarial example transferability with an intermediate level attack. In: Proceedings of the IEEE International Conference on Computer Vision, 2019. 4733–4742

39 Huang Y, Kong A W K. Transferable adversarial attack based on integrated gradients. 2022. ArXiv:2205.13152

40 Zhang J, Wu W, Huang J T, et al. Improving adversarial transferability via neuron attribution-based attacks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022. 14993–15002

41 Guo Y, Li Q, Zuo W, et al. An intermediate-level attack framework on the basis of linear regression. IEEE Trans Pattern

Anal Mach Intell, 2022, 45: 2726–2735
42 Wang X, Tong K, He K. Rethinking the backward propagation for adversarial transferability. In: Proceedings of the Advances in Neural Information Processing Systems, 2023. 1905–1922
43 Sun X, Cheng G, Li H, et al. On single-model transferable targeted attacks: a closer look at decision-level optimization. IEEE Trans Image Process, 2023, 32: 2972–2984
44 Zhang C, Benz P, Cho G, et al. Backpropagating smoothly improves transferability of adversarial examples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2021. 1–5
45 Krizhevsky A, Hinton G. Learning Multiple Layers of Features from Tiny Images. Technical Report, 2009
46 Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis, 2015, 115: 211–252
47 Zagoruyko S, Komodakis N. Wide residual networks. In: Proceedings of British Machine Vision Conference, 2016. 1–15
48 Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1492–1500
49 Sandler M, Howard A, Zhu M, et al. Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 4510–4520
50 Han D, Kim J, Kim J. Deep pyramidal residual networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 5927–5935
51 Dong X, Yang Y. Searching for a robust neural architecture in four GPU hours. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 1761–1770
52 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 7132–7141
53 Liu C, Zoph B, Neumann M, et al. Progressive neural architecture search. In: Proceedings of European Conference on Computer Vision, 2018. 19–34
54 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations, 2020. 1–12
55 Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE International Conference on Computer Vision, 2021. 10012–10022
56 Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention. In: Proceedings of International Conference on Machine Learning, 2021. 10347–10357
57 Bao H, Dong L, Piao S, et al. BEIT: BERT pre-training of image transformers. In: Proceedings of the International Conference on Learning Representations, 2021. 1–13
58 Tolstikhin I O, Houlsby N, Kolesnikov A, et al. MLP-mixer: an all-MLP architecture for vision. In: Proceedings of the Advances in Neural Information Processing Systems, 2021. 24261–24272
59 Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022. 11976–11986
60 Zhu Y, Sun J, Li Z. Rethinking adversarial transferability from a data distribution perspective. In: Proceedings of the International Conference on Learning Representations, 2021. 1–13
61 Liao F, Liang M, Dong Y, et al. Defense against adversarial attacks using high-level representation guided denoiser. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1778–1787
62 Guo C, Rana M, Cisse M, et al. Countering adversarial images using input transformations. In: Proceedings of the International Conference on Learning Representations, 2018. 1–12
63 Naseer M, Khan S, Hayat M, et al. A self-supervised approach for adversarial robustness. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 262–271
64 Geirhos R, Rubisch P, Michaelis C, et al. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: Proceedings of the International Conference on Learning Representations, 2018. 1–12
65 Singh N D, Croce F, Hein M. Revisiting adversarial training for imagenet: architectures, training and generalization across threat models. In: Proceedings of the Advances in Neural Information Processing Systems, 2024. 13931–13955