Boosting Adversarial Transferability via Spatial-Perceptual Reconfiguration

Chunqi Li, Zijian Ying, Qianmu Li*, Zhichao Lian
Nanjing University of Science and Technology, Nanjing, China
lichunqi@njust.edu.cn, zjying@njust.edu.cn, qianmu@njust.edu.cn, lzcts@163.com

***Abstract*—Transferable adversarial attacks aim to craft small perturbations on a surrogate model to mislead unseen target models. Existing input transformation-based attacks enrich gradient information but often produce perturbations with limited spatial coverage, reducing overlap with discriminative regions of unseen models. To address this, we systematically study how input transformations affect gradient distributions and introduce the Mean Distance to Centroid (MDC) to quantify spatial dispersion of high-magnitude gradients. We categorize transformations into Spatial Reconfiguration (SR), which redistributes existing gradients via geometric operations, and Perceptual Reconfiguration (PR), which induces new high-magnitude gradient regions without substantially altering spatial locations. Guided by this analysis, we propose the Spatial-Perceptual Reconfiguration (SPR) attack, combining SR and PR to generate diverse transformed inputs and aggregate gradients with broader coverage. Experiments on benchmark datasets show that SPR consistently outperforms seven baseline methods in cross-model transferability.**

***Index Terms*—Adversarial attacks, adversarial transferability, input transformation, Spatial-Perceptual Reconfiguration.**

I. Introduction

Despite the progress of deep neural networks (DNNs) in computer vision, their vulnerability to adversarial examples poses critical challenges to security-sensitive applications [1]. By adding imperceptible perturbations to input images, adversarial attacks can mislead DNNs into making incorrect predictions [1], [2]. Prior studies have shown that DNNs tend to rely heavily on a small number of discriminative regions, which often correspond to areas with high-magnitude gradient responses [3], [4]. Consequently, gradient information has become a fundamental basis for identifying critical regions and constructing effective adversarial perturbations.

In practical scenarios, attackers typically operate under black-box settings, where the architecture of the target model are unknown. As a result, transfer-based attacks are widely adopted, in which adversarial examples are generated on a surrogate model and then evaluated on unseen target models [5]. However, due to architectural differences and diverse feature extraction mechanisms, distinct or only partially overlapping discriminative regions [6]. Perturbations optimized on a single surrogate therefore tend to overfit its specific gradient patterns, failing to sufficiently cover the discriminative regions of other models, which significantly limits transferability.

To improve transferability, input transformation-based attack methods have been proposed to diversify gradients during adversarial optimization. For example, Block Shuffle and Rotation (BSR) [7] attempts to align attention distributions across models through structural perturbations such as block shuffling and rotation. While such strategies can partially enhance cross-model consistency, they often adopt transformation combinations in an ad-hoc manner, without explicitly modeling their impact on the spatial coverage of discriminative regions. As a result, the aggregated gradients may remain spatially concentrated or imbalanced, particularly when different models attend to complementary or spatially shifted regions, causing the resulting perturbations to inadequately cover discriminative regions of unseen models and thus limiting transferability across heterogeneous architectures.

Motivated by this observation, we propose a new perspective: rather than focusing solely on attention alignment, we aim to explicitly expand the spatial coverage of adversarial perturbations to encompass a broader set of potential discriminative regions. In black-box settings, although the discriminative regions of the target model are inaccessible, the gradient information from a surrogate model can serve as a useful proxy. By leveraging more comprehensive high-magnitude gradient distributions revealed by the surrogate, the crafted perturbations are more likely to overlap with the discriminative regions of unseen target models, thereby improving transferability.

To this end, we systematically investigate the effects of different input transformations on gradient distributions. Based on how various transformations influence gradient patterns, we categorize common input transformations into two complementary types: Spatial Reconfiguration (SR) and Perceptual Reconfiguration (PR). We further introduce the Mean Distance to Centroid (MDC) as a quantitative metric to measure the spatial dispersion of high-magnitude gradient regions under input transformations. Building on these insights, we propose the Spatial-Perceptual Reconfiguration (SPR) attack strategy, which combines SR and PR transformations to enhance the transferability of adversarial examples across heterogeneous models. The main contributions of this paper are summarized as follows:

- We introduce the Mean Distance to Centroid (MDC) metric to quantitatively assess how input transformations

Supported by the Special Project for Service-Oriented High-Quality Development of Jiangsu by Universities Directly under the Central Government: "Research on Targeted Multimodal Risk Assessment Technology for Large Model APIs and Equipment Industrialization" (Grant No. 27). * Corresponding author.

affect the spatial coverage of high-magnitude gradient regions.
- We categorize input transformations into spatial and perceptual reconfiguration types and propose the SPR attack strategy based on their complementary effects.
- Extensive experiments on benchmark datasets demonstrate that SPR significantly outperforms existing methods in cross-model adversarial transferability.

The code associated with this paper is publicly available at: https://github.com/FirecrackerPuppy/SPR

II. Related Work

A. Gradient-based Adversarial Attacks

Early gradient-based attacks include the Fast Gradient Sign Method (FGSM) [1] and its iterative variant, Iterative Fast Gradient Sign Method (I-FGSM) [8]. FGSM generates adversarial perturbations through a single gradient update, while I-FGSM applies multi-step iterative optimization to achieve stronger attacks. Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [9] builds upon I-FGSM by introducing a momentum term, which helps stabilize gradient updates and improves the transferability of perturbations across models. Subsequent input transformation-based adversarial attacks generally follow this framework: gradients are computed on a series of transformed inputs and integrated to update the momentum, thereby improving both generalization and cross-model transferability.

B. Input Transformation-based Adversarial Attacks

Input transformation-based attacks aim to enhance transferability by applying diverse transformations to inputs before gradient computation. Classic methods include DIM [10], which applies domain-invariant transformations to improve robustness; SIM [11], which distorts input images to confuse model decision boundaries; and TIM [12], which amplifies adversarial effects across various model architectures. Admix [13] combines adversarial perturbations with original images, SIA [14] focuses on image-specific augmentations for improved transferability, and BSR [7] shuffles image blocks to disrupt model attention. L2T [15] further improves transferability by dynamically learning input transformations during the attack process. More recently, Guo et al. [16] proposed the SGP method to construct multi-scale representations, compute gradients at each scale, and average them to generate more transferable perturbations. These studies show that carefully designed input transformations, including multi-scale transformations and learned approaches, can substantially enhance attack transferability.

III. Methodology

A. Preliminaries

The objective of adversarial attacks is to introduce imperceptible perturbations to clean inputs in order to mislead a DNN. Formally, given a model $f(\cdot;\theta)$ with parameters θ and a clean image x with label y, an adversarial example x^{adv} satisfies $\|x^{adv} - x\|_p \leq \epsilon$, $f(x^{adv};\theta) \neq y$, where ϵ is the perturbation budget and $\|\cdot\|_p$ denotes the ℓ_p-norm. Following prior work, we focus on the ℓ_∞-norm constraint. The adversarial generation problem can then be formulated as

$$x^{adv} = \arg \max_{\|x^{adv}-x\|_\infty \leq \epsilon} J(x^{adv}, y; \theta), \tag{1}$$

where $J(\cdot)$ represents the loss function. A classical method for generating adversarial examples is FGSM [1]:

$$x^{adv} = x + \epsilon \cdot \text{sign}(\nabla_x J(x, y; \theta)). \tag{2}$$

Its iterative version, I-FGSM [8], improves attack effectiveness by applying multiple small updates:

$$x_t^{adv} = x_{t-1}^{adv} + \alpha \cdot \text{sign}\big(\nabla_{x_{t-1}^{adv}} J(x_{t-1}^{adv}, y; \theta)\big), \quad x_0^{adv} = x, \tag{3}$$

where α denotes the step size. The momentum-based MI-FGSM [9] updates as

$$g_t = \mu \cdot g_{t-1} + \frac{\nabla_{x_{t-1}^{adv}} J(x_{t-1}^{adv}, y; \theta)}{\big\|\nabla_{x_{t-1}^{adv}} J(x_{t-1}^{adv}, y; \theta)\big\|_1}, \quad g_0 = 0, \tag{4}$$

$$x_t^{adv} = x_{t-1}^{adv} + \alpha \cdot \text{sign}(g_t), \tag{5}$$

where μ is the momentum decay factor. This momentum-based approach stabilizes the gradient direction across iterations and significantly improves attack transferability.

B. Gradient Distribution under Transformations

Previous works have shown that input transformations can significantly affect the distribution of gradients in deep neural networks [7], [10], [12]. To evaluate their impact on the spatial characteristics of gradient distributions, we introduce the Mean Distance to Centroid (MDC) metric, which quantifies the spatial distribution of high-magnitude gradient locations, thereby providing insights into their coverage of the model's discriminative regions.

Specifically, we identify spatial locations whose gradient magnitude exceeds a threshold τ:

$$S = \{(i, j) \mid |g(i, j)| \geq \tau\}, \tag{6}$$

where $|g(i, j)|$ denotes the absolute gradient magnitude at location (i, j). Following prior work on gradient-based saliency analysis [3], [4], we set τ to a fixed threshold of 0.1 [17], which provides a consistent criterion across images. This threshold balances the inclusion of sufficiently strong gradient responses with the preservation of their spatial distribution.

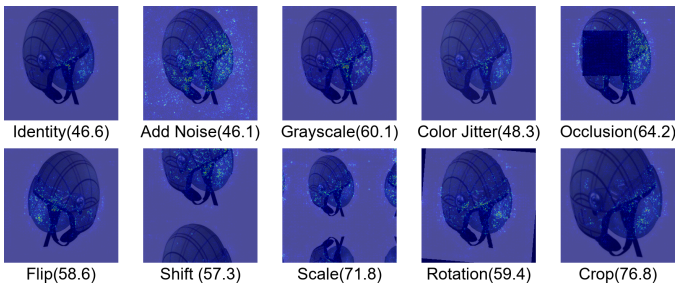

Fig. 1. Different transformations and their average corresponding MDC values.

We empirically observe that the relative trends of MDC across different transformations remain consistent under a wide range of τ values, indicating that our analysis is not sensitive to the exact choice of τ. The centroid (geometric center) of these high-magnitude gradient locations is then computed as

$$\mathbf{c} = \frac{1}{|S|} \sum_{(i,j)\in S} \begin{bmatrix} i \\ j \end{bmatrix}, \tag{7}$$

where $|S|$ denotes the number of high-magnitude gradient locations. The Mean Distance to Centroid (MDC) is subsequently defined as

$$\text{MDC} = \frac{1}{|S|} \sum_{(i,j)\in S} \left\| \begin{bmatrix} i \\ j \end{bmatrix} - \mathbf{c} \right\|_2, \tag{8}$$

which represents the average Euclidean distance from each high-magnitude gradient location to their collective centroid. A greater MDC indicates that salient gradient regions are more spatially dispersed, suggesting broader coverage of discriminative features, while a lower MDC implies that gradients are concentrated around a central region. Figure 1 illustrates the MDC values obtained under different input transformations, demonstrating their varying effects on gradient spatial distribution. We formally categorize input transformations into two complementary types based on their effects on gradient distribution:

Spatial Reconfiguration (SR) transformations primarily modify the spatial layout or structure of an image through geometric operations, such as flipping, shifting, scaling, rotating, and cropping. These operations cause existing high-magnitude gradient regions to diffuse along with the image's displacement or scaling. As a result, SR transformations generally increase the MDC compared to the original image, reflecting a wider spatial distribution of salient gradient regions.

Perceptual Reconfiguration (PR) transformations mainly modify perceptual attributes of an image, such as color, brightness, or local texture, without significantly altering spatial positions. Typical examples include adding noise, grayscale conversion, color jittering, and occlusion. These transformations change the model's perception of the input and induce new high-magnitude gradient regions that are absent in the original image. As a result, PR transformations exhibit different effects on MDC. Adding noise induces a large number of new gradient responses over wide areas; however, since these gradients are densely distributed, it leads to only marginal changes in MDC, which remains close to the baseline. Color jitter introduces only a limited number of new regions, leading to a modest MDC increase. In contrast, occlusion and grayscale produce more spatially dispersed gradients, yielding significantly higher MDC values. Overall, PR transformations enhance feature-level diversity, while MDC captures both the amount and spatial dispersion of newly activated gradient regions.

C. Spatial-Perceptual Reconfiguration

Gradient analysis reveals the complementary roles of SR and PR transformations: SR redistributes existing gradients to broaden spatial diversity, while PR elicits additional discriminative gradient regions to enrich feature diversity. To leverage these complementary effects, we propose the **Spatial-Perceptual Reconfiguration (SPR)** attack strategy. Specifically, SPR reconfigures the spatial distribution of high-magnitude gradient regions through SR transformations, ensuring broader coverage of original discriminative areas, and applies PR transformations to weaken or obscure original salient features and activate new ones, producing transformed images with more extensive high-magnitude gradient coverage. Formally, given an input image x, along with a pool of SR transformations $\mathcal{T}_{\text{sr}}$ and a pool of PR transformations $\mathcal{T}_{\text{pr}}$, SPR can be expressed as

$$\text{SPR}(x) = T_{\text{pr}} \circ T_{\text{sr}}, \quad T_{\text{sr}} \in \mathcal{T}_{\text{sr}}, T_{\text{pr}} \in \mathcal{T}_{\text{pr}}, \tag{9}$$

where $\circ$ denotes function composition. The overall workflow is summarized in Algorithm 1.

Algorithm 1 SPR Attack

Input: Classifier $f(\cdot)$, input image x with label y, loss $J(\cdot)$, SR transformations T_{sr}, PR transformations T_{pr}, number of transformed images N, perturbation bound ϵ, step size $\alpha = \epsilon/T$, iterations T, momentum μ
Output: Adversarial example x^{adv}
Initialize $g_0 = 0$, $x_0^{adv} = x$
for $t = 1$ **to** T **do**
  for $i = 1$ **to** N **do**
    Apply SR transformations: $x_i^{sr} = T_{sr}(x_{t-1}^{adv})$
    Apply PR transformations: $x_i^{spr} = T_{pr}(x_i^{sr})$
    Compute gradient: $g_i = \nabla_x J(f(x_i^{spr}), y)$
  end for
  Aggregate gradients: $\bar{g}_t = \frac{1}{N}\sum_{i=1}^{N} g_i$
  Update momentum: $g_t = \mu \cdot g_{t-1} + \frac{\bar{g}_t}{\|\bar{g}_t\|_1}$
  Update adversarial example:
    $x_t^{adv} = \text{clip}(x_{t-1}^{adv} + \alpha \cdot \text{sign}(g_t), 0, 1)$
end for
return x_T^{adv}

To validate the effectiveness of SPR, we instantiate it using two representative transformations from each pool, selected according to their impact on gradient distribution as characterized by MDC. For SR transformations, we prioritize those that achieve the largest MDC values, as higher MDC directly corresponds to broader spatial dispersion of existing high-magnitude gradient regions. Accordingly, scaling (S_1) is selected for its ability to produce the greatest MDC, effectively relocating salient gradients over a wider area, while cropping (S_2) is further applied to expand their spatial coverage.

For PR transformations, the selection follows a complementary criterion. Occlusion (P_2) is chosen because it yields the largest MDC among PR transformations, indicating the most spatially dispersed activation of new gradients. In contrast, adding noise (P_1) does not significantly increase MDC; however, it activates the largest number of new high-magnitude gradients in regions that are inactive in the original image. Although these gradients are densely distributed and thus

lead to a relatively modest MDC, they substantially enrich the set of discriminative regions exposed during optimization. Together, occlusion and noise provide complementary benefits in activating new gradient regions for PR. These can be composed as

$$T'_{\text{sr}} = S_1 \circ S_2, \quad T'_{\text{pr}} = P_1 \circ P_2, \tag{10}$$

the instantiated attack is then:

$$\text{SPR}'(x) = T'_{\text{pr}} \circ T'_{\text{sr}}(x). \tag{11}$$

By instantiating the attack using transformations selected based on MDC, broader coverage of high-magnitude gradient regions is ensured, providing more effective gradient redistribution than purely random methods. All subsequent experiments are conducted using this SPR$'$ instantiation.

IV. Experiments

A. Experimental Setup

Dataset. Following prior work, we evaluate all methods on 1,000 images randomly selected from the ILSVRC2012 validation set [18], as provided by Lin et al. [11].

Models. We consider ten ImageNet-pretrained models from both CNN and Transformer families. The CNNs include ResNet-50 (R50) [19], ResNet-101 (R101) [19], Inception-v3 (Inc-v3) [20], Inception-v4 (Inc-v4) [21], ConvNeXt-B (CN-B) [22], DenseNet-121 (D121) [23], and ResNeXt-101 (RX101) [24]. The Transformer-based models include Swin-T (Sw-T) [25], ViT-T [26], and ConViT-B (CV-B) [27].

Baselines. We compare SPR with state-of-the-art input transformation-based attacks, including DIM [10], TIM [12], SIM [11], Admix [13], SIA [14], BSR [7], and L2T [15]. All methods are implemented within the MI-FGSM framework.

Parameter Settings. We follow the standard MI-FGSM configuration used in prior work [7], [14], [15], setting the perturbation budget to $\epsilon = 16$, the number of iterations to $T = 10$, the step size to $\alpha = \epsilon/T$, and the momentum decay factor to $\mu = 1$. All attacks are constrained under the ℓ_∞ norm and optimized using the cross-entropy loss. For baseline methods, we adopt the recommended settings from their original papers. Specifically, DIM uses a transformation probability of $p = 0.7$, TIM applies a 7×7 Gaussian kernel, and SIM employs $m = 5$ scale copies. Admix mixes three images with a mixing ratio of $\eta = 0.2$. BSR divides images into 2×2 blocks with a maximum rotation angle of 24° and computes gradients over $N = 20$ transformations. For SIA, images are split into $s = 3$ blocks, with gradients computed over $N = 20$ transformed samples. L2T strictly follows the original implementation [15].

B. Evaluation on Single Model

We assess the transferability of adversarial examples generated on a single surrogate model against multiple black-box target models, with the results summarized in Table I.

Baseline methods achieve high attack success rates (ASR) on target models sharing the same architecture as the surrogate but suffer notable performance drops on different architectures.

TABLE I
ATTACK SUCCESS RATES (%) OF SPR AND SEVEN ATTACKS ON TEN MODELS. LIGHT GRAY COLUMNS INDICATE WHITE-BOX ATTACKS ON THE SURROGATE MODEL, AND THE DARK GRAY ROWS DENOTE OURS.

Model	Attack	R50	R101	Inc-v3	Inc-v4	CN-B	D121	RX101	Sw-T	ViT-T	CV-B	AVE
R50	DIM	99.8	66.7	61.6	57.3	53.2	73.8	49.0	49.4	46.3	31.4	58.9
	TIM	99.5	45.6	34.0	28.5	21.6	51.0	26.5	24.6	41.7	16.4	38.9
	SIM	99.9	72.9	63.9	57.1	59.2	79.0	56.6	55.5	49.0	36.8	63.0
	Admix	99.9	86.5	76.5	71.1	73.3	89.3	70.5	69.9	58.8	48.1	74.4
	SIA	**100.0**	90.4	85.7	83.6	81.9	93.9	75.0	81.6	69.9	55.7	81.8
	BSR	98.8	88.2	88.1	85.3	73.3	93.3	71.6	76.4	75.0	53.2	80.3
	L2T	98.5	89.4	87.5	86.9	81.3	92.1	78.7	80.3	78.5	64.5	83.8
	SPR	99.6	**91.0**	**89.4**	**89.7**	**86.4**	**94.4**	**83.0**	**84.0**	**80.3**	**69.8**	**86.8**
Inc-v3	DIM	57.2	58.0	99.7	69.2	35.3	61.7	38.8	35.9	39.9	23.5	51.9
	TIM	31.3	38.4	99.9	40.4	16.4	41.6	23.3	18.5	40.5	13.8	36.4
	SIM	56.2	58.8	**100.0**	71.3	34.0	66.8	40.5	37.0	43.8	24.8	53.3
	Admix	73.5	75.4	**100.0**	87.1	46.2	81.1	55.5	48.4	56.0	31.4	65.5
	SIA	88.2	87.2	**100.0**	97.5	64.4	90.5	68.6	69.5	67.0	39.8	77.3
	BSR	83.1	80.1	99.9	91.8	47.4	85.6	59.5	57.5	66.8	33.3	70.5
	L2T	90.5	91.1	99.8	96.4	72.2	92.9	76.8	75.3	78.7	51.6	82.5
	SPR	**92.3**	**92.7**	99.9	**97.9**	**76.5**	**93.7**	**81.4**	**78.4**	**80.7**	**56.9**	**85.0**
ViT-T	DIM	52.5	64.9	52.3	47.4	38.1	68.7	39.2	65.9	**100.0**	75.0	35.5
	TIM	25.6	39.6	27.5	23.0	11.5	44.2	20.0	27.3	**100.0**	36.6	60.4
	SIM	40.5	54.9	42.1	36.2	24.2	59.4	30.8	54.4	**100.0**	70.3	51.3
	Admix	48.8	61.1	48.7	41.6	29.3	67.5	35.1	63.7	**100.0**	79.6	57.5
	SIA	80.1	89.2	78.1	73.6	63.7	92.2	64.8	90.6	**100.0**	**94.0**	82.6
	BSR	78.7	86.5	78.0	72.4	48.9	89.6	61.9	76.2	**100.0**	74.6	76.7
	L2T	83.1	90.6	84.2	80.6	66.9	93.0	70.6	88.7	**100.0**	91.1	84.9
	SPR	**89.6**	**93.1**	**88.0**	**85.2**	**75.9**	**94.7**	**78.7**	**91.5**	**100.0**	93.3	**89.0**
Sw-T	DIM	73.5	67.3	69.4	65.1	82.7	75.3	47.5	99.8	71.0	68.0	72.0
	TIM	40.0	44.1	39.2	34.2	41.8	49.7	26.7	**100.0**	61.5	41.1	47.8
	SIM	52.2	45.5	44.2	38.3	73.8	58.4	30.8	**100.0**	61.7	50.9	55.6
	Admix	58.5	51.1	50.4	44.1	78.8	63.0	35.6	**100.0**	64.7	57.0	60.3
	SIA	91.8	86.4	86.9	86.9	97.5	90.7	66.3	**100.0**	88.2	85.7	88.0
	BSR	94.4	91.2	92.4	92.5	93.8	93.8	73.1	99.6	90.0	81.1	90.2
	L2T	92.5	90.8	91.6	90.8	95.7	92.7	77.3	**100.0**	92.7	90.0	91.4
	SPR	**96.2**	**95.5**	**95.6**	**95.7**	**97.3**	**96.4**	**87.1**	99.6	**95.8**	**94.1**	**95.3**

In contrast, SPR consistently achieves the highest average ASR, improving both intra- and cross-architecture transferability. For example, using Swin-T as the surrogate, SPR attains an ASR of 95.3%, outperforming L2T (91.4%), with gains of 4.1% on ConViT-B and 9.8% on ResNeXt-101. These results demonstrate that SPR generates more transferable perturbations that generalize well across diverse architectures.

C. Evaluation on Defensive Methods

We further evaluate SPR against two adversarially trained models (Inc-v3$_{adv}$ and IncRes-v2$_{adv}$) and four preprocessing defenses, including JPEG [28], R&P [29], FD [30], and PD [31]. As shown in Table II, SPR consistently outperforms all baselines, achieving an average improvement of 3.3% over L2T. Notably, on IncRes-v2$_{adv}$, SPR exceeds L2T by 8.2% while maintaining around 85% ASR against other defenses, demonstrating strong robustness against defensive mechanisms.

TABLE II
AVERAGE ATTACK SUCCESS RATES (%) OF SPR AND FOUR ATTACKS AGAINST SIX DEFENSES, USING INC-V3 AS THE SURROGATE.

Attack	Inc-v3$_{\text{adv}}$	IncRes-v2$_{\text{adv}}$	JPEG	RP	FD	PD	AVE
Admix	51.5	26.1	65.1	65.3	66.4	65.6	56.7
SIA	55.3	28.6	75.4	75.3	76.8	77.2	64.8
BSR	45.2	23.1	70.0	72.1	72.0	71.0	58.9
L2T	67.4	45.9	81.9	83.3	82.9	82.7	74.0
SPR	**69.1**	**54.1**	**84.8**	**85.6**	**85.6**	**84.6**	**77.3**

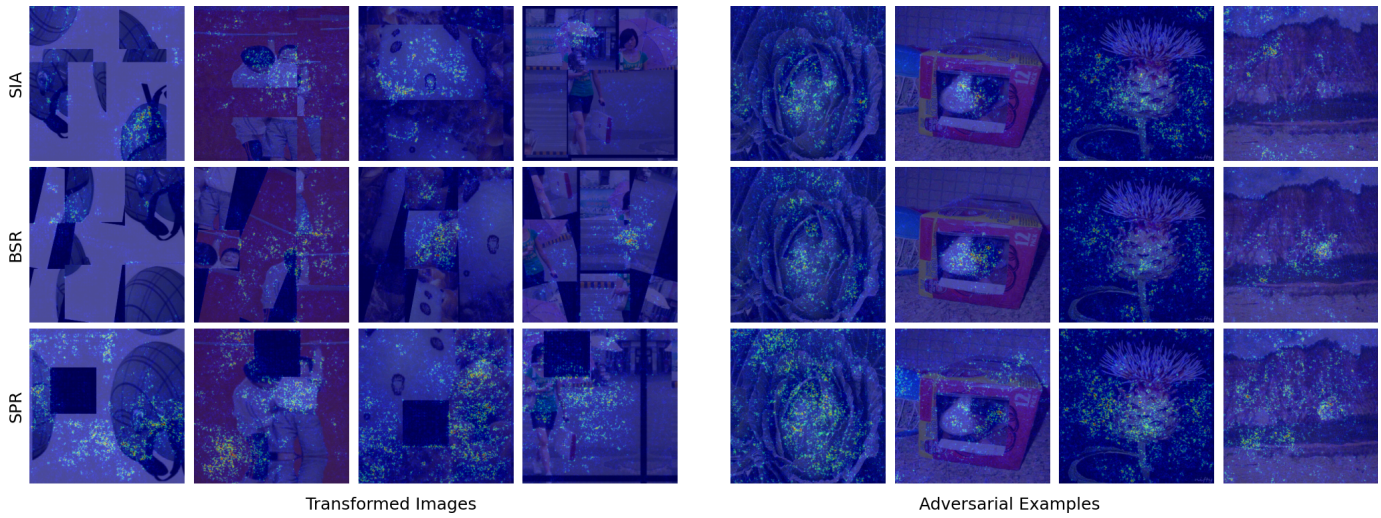

Fig. 2. Transformed images and adversarial examples generated by SPR, BSR, and SIA on the surrogate model R50.

D. Ablation Study

Effectiveness of SR and PR. Table III reports the ASR of different SPR configurations. The full SPR achieves the highest average ASR of 85.3%, outperforming variants using only SR or only PR, confirming their complementary effects.

Moreover, both SR and PR individually outperform MI-FGSM, indicating that each module contributes positively to transferability.

Number of transformed images N. We investigate the effect of the number of transformed images N, which determines the number of transformed inputs used for gradient aggregation during each attack iteration. As illustrated in Fig. 3, increasing N from 1 to 20 leads to a substantial improvement in the average ASR from 59.8% to 86.0% across ten target models, indicating that a larger set of transformed samples promotes more comprehensive gradient coverage and improves transferability. Further increasing N yields only marginal gains, revealing diminishing returns. Notably, SPR with $N = 1$ already surpasses MI-FGSM, highlighting its strong effectiveness even with limited transformations. Based on the trade-off between computational overhead and performance, we set $N = 20$ in all subsequent experiments.

E. Gradient Distribution Validation

Broader Gradient Coverage of SPR. To assess SPR's effectiveness in expanding gradient coverage, we compare the average MDC of transformed and adversarial examples generated by SIA, BSR, and SPR. As shown in Table IV and Fig. 2, SPR consistently achieves the highest MDC on both the surrogate model (R50) and black-box models (Sw-T, Inc-v4). A higher MDC indicates that high-magnitude gradient regions are more spatially dispersed, implying that the resulting perturbations are distributed over a broader set of potential discriminative regions. This is particularly desirable in black-box transfer attacks, where the exact discriminative regions of the target model are unknown and may differ substantially from those of the surrogate. By covering a wider spatial range of high-magnitude gradient regions, SPR increases the likelihood that the crafted perturbations overlap with discriminative regions of unseen target models.

Relationship Between Gradient Distribution and Transferability. To further examine the relationship between gradient dispersion and attack transferability, we generate adversarial examples from transformed inputs ranked by their MDC values, selecting the top-5 and bottom-5 cases under identical attack settings. As reported in Table V, adversarial examples with higher MDC consistently achieve higher ASR across all evaluated models. This result confirms that broader and more dispersed high-magnitude gradient distributions are strongly correlated with improved transferability.

TABLE III
ATTACK SUCCESS RATES (%) OF MI-FGSM, SPR, SR, AND PR ON MULTIPLE MODELS, USING R50 AS THE SURROGATE.

Attack	R101	Inc-v3	Inc-v4	CN-B	D121	RX101	Sw-T	ViT-T	CV-B	AVE
SPR	**91.0**	**89.4**	**89.7**	**86.4**	**94.4**	**83.0**	**84.0**	**80.3**	**69.8**	**85.3**
SR	88.7	87.7	89.0	83.4	90.3	79.4	80.1	71.5	61.6	81.3
PR	81.7	70.8	65.3	70.0	86.6	63.1	65.4	56.2	41.7	66.8
MI-FGSM	48.0	35.7	30.7	32.0	54.5	29.7	32.0	30.2	16.3	34.3

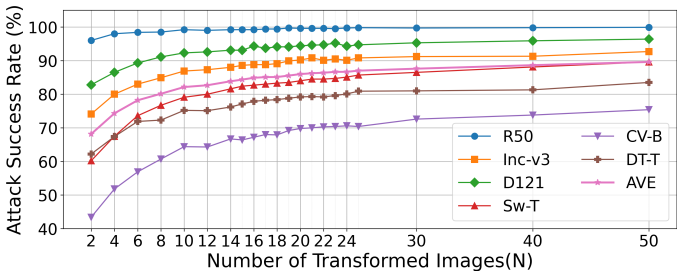

Fig. 3. Attack success rates (%) of SPR under different numbers of transformed images N.

TABLE IV
AVERAGE MDC OF TRANSFORMED INPUTS (TRANS.) AND ADVERSARIAL EXAMPLES (ADV.) GENERATED BY DIFFERENT ATTACKS, USING R50 AS THE SURROGATE MODEL.

Attack	R50 (Trans.)	R50 (Adv.)	Sw-T (Adv.)	Inc-v4 (Adv.)
SIA	69.1	75.6	75.9	66.0
BSR	66.6	74.9	73.6	68.3
SPR	**80.2**	**76.2**	**77.7**	**69.7**

TABLE V
ATTACK SUCCESS RATES (%) OF SPR ADVERSARIAL EXAMPLES GENERATED FROM TRANSFORMED IMAGES WITH DIFFERENT MDC VALUES, USING R50 AS THE SURROGATE AND TEN IMAGES PER ITERATION.

Attack	R50	R101	Inc-v3	Inc-v4	CN-B	D121	RX101	Sw-T	ViT-T	CV-B	AVE
Bottom-5	97.0	80.6	79.6	78.2	72.7	87.5	67.0	69.2	67.9	52.5	75.2
Top-5	**98.0**	**83.5**	**82.6**	**81.0**	**74.9**	**88.8**	**73.7**	**72.5**	**70.0**	**57.2**	**78.2**

V. CONCLUSION

In this paper, we propose a novel input transformation-based attack, **Spatial-Perceptual Reconfiguration (SPR)**, to enhance the transferability of adversarial examples. By generating a diverse set of transformed images through spatial and perceptual reconfiguration, SPR produces broader high-magnitude gradient responses, which are subsequently aggregated into adversarial perturbations with wider spatial coverage. Extensive experiments across both convolutional neural networks (CNNs) and Vision Transformers (ViTs) demonstrate that SPR consistently outperforms state-of-the-art attacks in cross-model transferability. Furthermore, our results provide valuable insights into the spatial distribution and generalization of adversarial perturbations, highlighting the importance of high-magnitude gradient coverage for effective transferable attacks.

REFERENCES

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint* arXiv:1312.6199, 2013.

[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.

[4] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint* arXiv:1312.6034, 2013.

[5] W. Ma, Y. Li, X. Jia, and W. Xu, "Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023.

[6] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 12116–12128, 2021.

[7] K. Wang, X. He, W. Wang, and X. Wang, "Boosting adversarial transferability by block shuffle and rotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[8] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC, 2018.

[9] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.

[10] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.

[11] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[12] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.

[13] X. Wang, X. He, J. Wang, and K. He, "Admix: Enhancing the transferability of adversarial attacks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021.

[14] X. Wang, Z. Zhang, and J. Zhang, "Structure invariant transformation for better adversarial transferability," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023.

[15] R. Zhu, Z. Zhang, S. Liang, Z. Liu, and C. Xu, "Learning to transform dynamically for better adversarial transferability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[16] Z. Guo, C. Wan, Y. Zheng, H. Kuang, and X. Lu, "Boosting adversarial transferability against defenses via multi-scale transformation," in *Proc. ICIC*. Springer, 2025, pp. 502–513.

[17] M. Luperto, V. Stakanov, G. Boracchi, N. Basilico, and F. Amigoni, "Biasing frontier–based exploration with saliency areas," in *Proc. IEEE Eur. Conf. Mobile Robots (ECMR)*, 2025, pp. 1–7.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[21] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017.

[22] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.

[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

[24] S. Xie, R. Girshick, P. Doll'ar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[27] S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "ConViT: Improving vision transformers with soft convolutional inductive biases," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.

[28] Z. Liu, Q. Liu, T. Liu, N. Xu, X. Lin, Y. Wang, and W. Wen, "JPEG compression for adversarial robustness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.

[29] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[30] Z. Liu, Q. Liu, T. Liu, N. Xu, X. Lin, Y. Wang, and W. Wen, "Feature distillation: DNN-oriented JPEG compression against adversarial examples," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.

[31] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer, "Deflecting adversarial attacks with pixel deflection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.