STAR: Stylized and Transferable Adversarial Robustness

Qing Luo[1 [0009−0003−2560−8610]], Yili Wang[3 [0009−0005−7615−9071]], Zhi Chen[3 [0009−0005−6693−6850]], Xiangwen Xiao[3 [0009−0009−1479−9304]], Cheng Zeng[3 [0009−0006−8953−5619]], and Xiaoyi Zhou[3 [0000−0003−3777−9479]]

Hainan University, Haikou, China

Abstract. Deep Neural Networks (DNNs) remain vulnerable to adversarial examples. Conventional L_P-norm attacks achieve high white-box success but transfer poorly due to reliance on model-specific minima, while unrestricted attacks (e.g., style transfer) improve naturalness but overfit to surrogate model preferences, limiting reliability. We identify surrogate overfitting as the key bottleneck and propose STAR, a bilevel framework combining style transfer with Reverse Adversarial Perturbation (RAP). STAR explicitly seeks flat regions of the loss landscape: inner-loop ascent explores robust adversarial directions, while outer-loop minimization embeds transfer-optimized perturbations into holistic style transformations via a composite loss balancing fidelity and attack strength. Experiments show STAR achieves 99% white-box and 95% black-box success on ImageNet (24% better than prior style attacks), 91% on CIFAR-100, and 84–100% under defenses. It also preserves visual quality (SSIM=0.73, LPIPS=0.24). Ablation confirms flat-region optimization is critical—its removal reduces transferability by 28%. Code will be released publicly upon publication. Deep Neural Networks (DNNs) remain vulnerable to adversarial examples. Conventional L_P-norm attacks achieve high white-box success but transfer poorly due to reliance on model-specific minima, while unrestricted attacks (e.g., style transfer) improve naturalness but overfit to surrogate model preferences, limiting reliability. We identify surrogate overfitting as the key bottleneck and propose STAR, a bilevel framework combining style transfer with Reverse Adversarial Perturbation (RAP). STAR explicitly seeks flat regions of the loss landscape: inner-loop ascent explores robust adversarial directions, while outer-loop minimization embeds transfer-optimized perturbations into holistic style transformations via a composite loss balancing fidelity and attack strength. Experiments show STAR achieves 99% white-box and 95% black-box success on ImageNet (24% better than prior style attacks), 91% on CIFAR-100, and 84–100% under defenses. It also preserves visual quality (SSIM=0.73, LPIPS=0.24). Ablation confirms flat-region optimization is critical—its removal reduces transferability by 28%. Code will be released publicly upon publication.

Keywords: Adversarial Examples · Transferability · Style Transfer · artistic attack · Black-Box Attacks

1 Introduction

Deep Neural Networks (DNNs) have achieved substantial progress in critical computer vision tasks such as image classification, object detection, and semantic segmentation, driving advancements in domains like autonomous driving and medical imaging[2–4, 6, 8]. However, their susceptibility to adversarial examples—inputs perturbed imperceptibly to cause misclassification—poses a significant security risk[12]. This vulnerability undermines trust in real-world deployments; for instance, manipulated traffic signs could mislead autonomous vehicles[13]. Understanding and mitigating these threats is thus crucial for both security hardening and revealing fundamental model weaknesses.

Adversarial attacks fall into two dominant paradigms: L_P-norm constrained and unrestricted methods. L_P-norm attacks (e.g., FGSM, PGD) achieve high white-box success by confining perturbations within small ϵ-balls, yet suffer from severely limited black-box transferability due to over-reliance on local pixel manipulations that disregard global semantics[11]. Conversely, unrestricted attacks (e.g., style transfer) improve visual naturalness by modifying semantic attributes[19, 35], but still exhibit remains limited transferability—often below 70% against diverse black-box models[27].

In this context, the term style refers to the set of high-level perceptual attributes that characterize the visual appearance of an image, such as texture patterns, color distributions, brushstroke characteristics, or other artistic features. This notion is widely adopted in neural style transfer literature, where style is typically captured by feature correlations (Gram matrices[20]) extracted from convolutional neural networks, in contrast to content, which encodes the spatial and semantic structure of objects within an image. By embedding perturbations into these stylistic attributes, style-transfer-based adversarial examples alter the visual presentation while preserving the underlying semantic content.

We identify a unified root cause for this dilemma: pronounced overfitting to the surrogate model. In L_P-norm attacks, perturbations overfit to model-specific gradient patterns and sharp loss minima; in unrestricted attacks, adversarial textures become entangled with the surrogate's idiosyncratic visual preferences[7]. This overfitting traps perturbations in model-specific regions of the loss landscape, fundamentally limiting cross-model generalization. Consequently, breaking this over-reliance on the surrogate model to discover a more "universal" solution residing in a "flat" region of the loss landscape remains a key challenge for enhancing the transferability of unrestricted attacks[26].

In this paper, we propose an unrestricted attack method based on style transfer that introduces a transferability enhancement technique capable of generating adversarial images with high transferability. As shown in Figure 1. Our method's core principle is to proactively seek model-agnostic adversarial features that reside in the flat regions of the loss landscape, as these areas represent more fundamental vulnerabilities likely to be shared across different models[26]. To achieve this, we introduce a bi-level optimization framework rooted in Reverse Adversarial Perturbation (RAP). The inner maximization step of RAP deliberately probes for resilient adversarial directions, effectively steering the outer

optimization away from sharp, model-specific minima towards these broad, flat plateaus. Within our framework, style transfer not only embeds these highly transferable perturbations into the image's global texture and style but also acts as a holistic regularizer, ensuring the resulting adversarial style is inherently robust and generalizable.

Our method is an iterative optimization process guided by a composite loss function. This function includes artistic loss terms that maintain visual quality and, more critically, optimizes adversarial loss through Reverse Adversarial Perturbation (RAP) mechanism to enhance transferability. Specifically, we directly integrate RAP's bilevel optimization strategy into style generation: the inner loop actively searches for robust adversarial directions via gradient ascent, while the outer loop guides the optimization process to avoid model-specific "sharp minima" and converge to more generalized "flat" regions.

Within this framework, style transfer plays a dual role: it functions not only as a feature transfer tool but also as a regularizer, capable of integrating optimized perturbations into the image's overall texture and style, and transferring them to the adversarial image. This ensures that the generated adversarial images exhibit the artistic style of the target image without altering critical semantic content, while maintaining a high attack success rate across various deep neural network (DNN) models, thereby effectively balancing attack efficacy and visual quality.

Transcending conventional approaches confined to geometry-specific sharp minima, our principle deliberately steers adversarial optimization toward broad loss manifolds housing cross-model generalizable weaknesses.

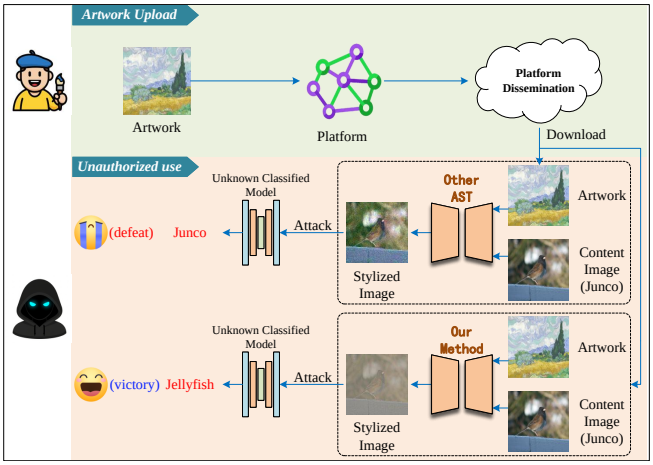

Fig. 1. Attack scenarios for STAR. Artists register and upload their art style images on art sharing platforms. Unauthorized users may download these art style images and use Artistic Style Transfer (AST) for style migration.Our method demonstrates robust performance when attacking unknown models.

Our main contributions are as follows:

- Extending flatness to style-transfer adversarial attacks. We are the first to extend the concept of flatness from classification robustness to style-transfer-based adversarial attacks, introducing the notion of style-loss flat regions—areas of the loss landscape where small perturbations induce minimal changes. Furthermore, we reformulate style selection from arbitrary or manual choices into a data-driven, model-aware optimization process.
- Unified STAR framework with conflict-aware optimization. We propose STAR, a unified bilevel optimization framework that systematically integrates transferability enhancement with style transfer. Unlike prior works that apply these techniques independently, STAR redefines style transfer as a structural regularizer, embedding optimized perturbations holistically into global image attributes, thereby improving both visual fidelity and adversarial generalization.
- Empirical validation of synergy. Comprehensive ablation studies confirm that STAR's effectiveness arises from the synergy between transferability enhancement and stylization rather than isolated effects. Removing the transferability component reduces transferability by 28%, while disabling style-loss terms simultaneously damages robustness and perceptual quality. These results demonstrate that STAR's improvements stem from its tightly coupled design, achieving significant gains unattainable by individual components alone.

2 Related Work

2.1 Style-Transfer Adversarial Examples

To address the limitations of imperceptible, norm-bounded perturbations, a paradigm of Unrestricted Adversarial Examples (UAEs) has emerged[17, 18, 27]. Instead of constraining perturbation magnitude, UAEs aim to generate more naturalistic attacks by modifying semantic or stylistic image attributes. This is accomplished through diverse strategies, such as using generative models to synthesize adversarial candidates[9], altering differentiable properties like color and texture[17], or manipulating features on a low-dimensional manifold of natural images[18]. While often more stealthy and robust to certain defenses, UAEs still face challenges in ensuring consistent semantic integrity and achieving high attack transferability[9].

Within the UAEs paradigm, **style-based attacks** have become a particularly promising research direction[10]. These methods integrate principles from artistic style transfer, embedding adversarial perturbations within a holistic style transformation. By transferring stylistic features—such as textures, color palettes, and brushstroke patterns[27, 35]—from a reference image, these attacks aim to disrupt a model's reliance on superficial, texture-based cues, inducing misclassification while appearing as a legitimate artistic rendering. Although current style-based attacks offer improved visual quality and higher transferability

than their Lp-norm counterparts, enhancing their attack efficacy across diverse, unknown models without introducing visual artifacts remains a key research challenge[14, 24].

2.2 Transferability

Transferability is a critical property of an adversarial example, measuring its effectiveness against models other than the one it was generated on surrogate model. High transferability is essential for practical black-box attacks, as it indicates that the perturbation captures a fundamental vulnerability shared across diverse models, rather than merely overfitting to the idiosyncrasies of a specific surrogate model.

To enhance adversarial transferability, extensive research has been conducted, primarily focusing on three strategic directions: optimizing the perturbation, transforming the input data, and analyzing the underlying factors[26].

Perturbation Optimization: This dominant approach refines the adversarial perturbation generation process itself to improve generalization. Key innovations include: finding perturbations in flat regions of the loss landscape to circumvent surrogate overfitting [26]; incorporating momentum to stabilize update directions and escape poor local optima [30]; and compute universal perturbations effective across a wide range of models[21] or leverage neighborhood attribution to improve transferability[25].

Input Data Transformation: This strategy applies diverse transformations to input images during attack generation, forcing the learned adversarial features to be more generalizable. Representative methods include: Admix, which diversifies gradient directions by computing gradients on images mixed with small portions of out-of-class samples while preserving the original label [15]; and leveraging style transfer to project images into varied stylistic domains, thereby enhancing the transferability of the resulting adversarial examples [24].

Foundational Analysis and Loss Landscape Modification: This line of work focuses on altering the underlying geometry of the loss landscape or theoretically reconciling fundamental trade-offs. Significant advancements encompass: techniques to flatten long-range representation-space loss landscapes, exemplified by FLoR's stochastic interpolation of normalized features [22]; theoretical analysis attributing the robustness-transferability trade-off to balancing model smoothness and gradient similarity, leading to regularization-based solutions [23]; and methods to enhance perturbation diversity within attention or feature spaces, such as ADA's latent-conditioned generator [7]

3 Methodology

3.1 Problem Analysis and Core Principle

Although advanced style-transfer attacks like AdvST[27] generate visually natural and imperceptible adversarial examples, their transferability against unknown models in black-box settings remains suboptimal. Our analysis reveals

this intrinsic limitation stems from pronounced overfitting to the surrogate (source) model during optimization.

From a geometric perspective of the loss landscape, existing methods tend to converge to sharp, narrow valleys (sharp minima) in the surrogate model's loss landscape. These regions correspond to highly vulnerable yet model-specific points on the decision boundary. Consequently, an adversarial sample crafted within such sharp minima of Surrogate Model A may no longer occupy an effective low-loss region in Target Model B's landscape, resulting in failed attacks.

To overcome this overfitting, our core principle proposes: instead of seeking any loss-maximizing point, actively pursue broad, flat regions (flat minima). This approach is grounded in a key insight—solutions in flat regions exhibit enhanced generalization and robustness[26]. Such areas typically correspond to intrinsic, transferable vulnerabilities in decision boundaries that are more likely to be shared across architectures. Adversarial samples generated in flat minima thus possess inherently high transfer potential.

3.2 Preliminaries

Reverse Adversarial Perturbation (RAP) [26] is an optimization technique designed to find adversarial samples residing in flat regions of the loss landscape, which exhibit superior transferability compared to those in sharp minima. Its core motivation stems from the insight presented in Section 3.1: adversarial samples residing in flat regions of the loss landscape exhibit superior transferability compared to those in sharp regions.

RAP operates as a Min-Max bilevel optimization process:

- **Inner Maximization:** For the current adversarial sample I_{adv}, RAP performs K steps of gradient *ascent* within a small neighborhood (defined by ϵ_n) to find a *worst-case* perturbation point n_{rap} that maximizes the classification loss:

$$n_{adv} \leftarrow n_{adv} + \alpha_n \cdot \text{sign}\left(\nabla_{n_{rap}} \mathcal{L}(M_s(I_{adv} + n_{adv}), y)\right) \tag{1}$$

  where α_n is the inner loop step size, M_s denotes the surrogate model, y denotes the ground-truth label for $I_{adv}(y \in Y)$
- **Outer Minimization:** The optimizer then performs standard gradient *descent* on the adversarial sample I_{adv}, starting from the perturbed point $I_{adv} + n_{adv}$ found in the inner loop:

$$I_{adv} \leftarrow \text{Clip}_{\mathcal{B}_\epsilon(I)}\Big[I_{adv} - \alpha \cdot \text{sign} \left(\nabla_{I_{adv}} \mathcal{L}(M_s(G(I_{adv} + n_{adv})), y)\right)\Big] \tag{2}$$

  Here, α is the outer loop step size, $G(\cdot)$ incorporates stochastic input transformations to improve generalization, and we enforce pixel-range clipping to keep image values within valid bounds. For instance, in our implementation $G(\cdot)$ is instantiated as a combination of random resizing[28] and scale transformations[29]. When we report norm-bounded results, we additionally

project onto the neighborhood $\mathcal{B}_\epsilon(x) = \{x' : ||x' - x||_p \leq \epsilon\}$, with ϵ=16/255. Concretely, after each outer update we first project I_{adv} onto $B_\epsilon(I_{adv})$ when using a norm constraint, and then clip pixel values to the valid range to ensure a legitimate image.

The key insight is that this "ascend-then-descend" process inherently discourages convergence to sharp minima. Starting descent from a local maximum within a sharp valley often causes escape, while descent within a flat region remains stable, naturally guiding the optimization towards more transferable adversarial solutions.

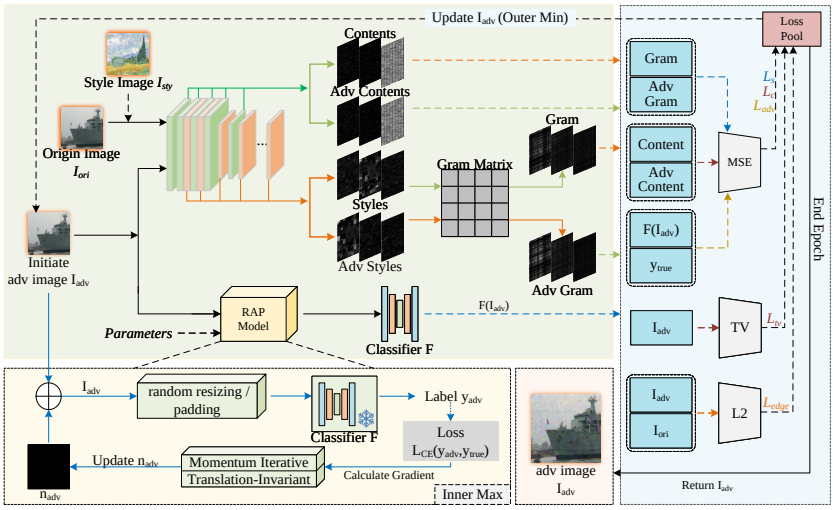

Fig. 2. STAR framework workflow. The pipeline synergizes artistic loss (ensuring visual quality) with Reverse Adversarial Perturbation (RAP) for transferability enhancement, synthesizing adversarial examples through gradient fusion. Notation: I_{ori}: original image, I_{sty}: style reference, I_{adv}: adversarial output.

3.3 The Proposed STAR Framework

The core concept of the STAR framework formulates the generation of stylized, highly transferable adversarial samples as a unified, end-to-end composite optimization problem. This approach seamlessly integrates artistic style transfer objectives with adversarial goals for enhanced transferability. Through synergistic optimization, STAR systematically uncovers vulnerable features shared across models while maintaining high visual quality. The overall workflow of our method is shown in Figure 2 and then present the complete optimization algorithm in Algorithm 1(Appendix A).

We formalize this optimization problem with the composite objective function L_{STAR}(Equation 3) comprising two key components: artistic fidelity loss L_{art}(Equation 4) ensuring visual quality, adversarial loss L_{adv}(Equation 9) inducing model misclassification while enhancing transferability.

The optimization seeks an adversarial image I_{adv} that minimizes the weighted sum of these objectives:

$$\min_{I_{adv}} \mathcal{L}_{STAR} = \mathcal{L}_{art} + \lambda \mathcal{L}_{adv} \tag{3}$$

The hyperparameter λ regulates the trade-off between artistic quality and attack strength.

Artistic Fidelity Component: Stylization Loss L_{art} To generate visually natural adversarial samples conforming to a specific artistic style, we employ a composite artistic loss L_{art} analogous to AdvST[27]. This loss function constitutes a weighted sum of four sub-loss terms that collectively enforce content integrity(Equation 5,6,7,8), stylistic coherence, structural clarity, and visual smoothness in the generated image.

$$L_{art} = \alpha L_c + \beta L_s + \gamma L_{edge} + \tau L_{tv} \tag{4}$$

The component losses are formally defined as follows:

- Content Loss(L_c): This measures the content discrepancy between the generated image I_{adv} and the original content image I_c. The loss is computed by comparing feature map activations at specific higher-level convolutional layers of the VGG network, ensuring the generated image preserves key semantic structures and objects of the original. It is defined as:

$$\mathcal{L}_c = \sum_{l_c} \frac{1}{2} \left\| \phi_{content}(I_{adv}) - \phi_{content}(I_{ori}) \right\|_2^2, \tag{5}$$

  where $\phi_{content}(\cdot)$ denotes the VGG feature extractor at the designated content layer l_c, I_{ori} is the original content image, I_{sty} is the style reference image
- Style Loss(L_s): This quantifies the stylistic similarity between the generated image I_{adv} and the style reference I_{sty}. Style is defined as feature correlations across different convolutional layers in the VGG network, typically measured through Gram matrices[20]. The style loss matches these correlations to reproduce artistic characteristics such as color schemes, brushstrokes, and textures. It is formally defined as:

$$\mathcal{L}_s = \sum_{l_s} \left\| G_m(\phi_{style}(I_{adv})) - G_m(\phi_{style}(I_{sty})) \right\|_F^2, \tag{6}$$

  where $G_m(\cdot)$ is Gram matrix operator, $\phi_{style}(\cdot)$ represents the feature extractor at designated style layers l_s, $||\cdot||_F$ indicates the Frobenius norm

– Edge Loss (L_{edge}): To preserve clear object contours and fine details in the original image, we introduce an edge loss[27]. This penalizes structural contour discrepancies between I_{adv} and I_{ori} using a spatially extended Laplacian operator S[33]:

$$L_{edge} = \sum_{c=1}^{3} \|S^n[I_{adv}] - S^n[I_{ori}]\|^2 \tag{7}$$

where S denotes the edge extraction operator, c indexes image channels, and $||\cdot||^2$ is the Euclidean norm.

– Total Variation Loss (L_{tv}): This loss serves as a smoothness regularization term that suppresses high-frequency noise and artifacts in the generated image, enhancing its visual naturalness[27]:

$$\begin{aligned} L_{tv} = \sum_{x,y} \big(& (I_{adv}(x, y+1) - I_{adv}(x, y))^2 \\ & + (I_{adv}(x+1, y) - I_{adv}(x, y))^2 \big) \end{aligned} \tag{8}$$

where $I(x, y)$ represents the pixel value of I at coordinates (x, y).

The detailed mathematical formulations for these components are adopted from the original AdvST paper[27].

Transferability-Promoting Adversarial Objective L_{adv} The distinguishing characteristic of the STAR framework resides in its optimization strategy for the adversarial objective, L_{adv}. Conventional stylized attack methodologies typically employ direct gradient ascent on the classification loss, a process that exhibits a tendency to converge towards sharp, model-specific minima in the loss landscape. This overfitting to the surrogate model's idiosyncratic vulnerabilities inherently curtails the transferability of the generated adversarial examples.

To overcome this limitation, we formulate the adversarial objective as a bilevel min-max optimization problem(Equation 9) rather than directly minimizing a simple adversarial loss. This design seamlessly integrates the Reverse Adversarial Perturbation (RAP) mechanism introduced in Section 3.2 into the adversarial loss computation(Equation 1,2). Formally, we first define the base adversarial loss as a standard classification loss $\mathcal{L}_{CE}(M_s(G(I_{adv})), y)$.

We then embed this loss within RAP's min-max framework, yielding the complete L_{adv} optimization objective:

$$\mathcal{L}_{adv} = \min_{I_{adv}} [\lambda \cdot \max \mathcal{L}_{CE}(M_s(G(I_{adv} + n_{adv})), y)] \tag{9}$$

By employing the "ascend-then-descend" bilevel optimization process of RAP, as detailed in Section 3.2, the STAR framework actively steers optimization trajectories away from surrogate model-specific sharp vulnerabilities. This strategy intrinsically penalizes sharp regions and promotes convergence within broad, flat plateaus, which correspond to vulnerable regions more likely to be shared across diverse models. Consequently, the generated adversarial perturbations intrinsically possess high transferability—the fundamental reason for STAR's superior black-box attack performance.

4 EXPERIMENTS

To rigorously validate the efficacy of the proposed STAR framework, this chapter designs and conducts a extensive set of experiments. The evaluation systematically measures STAR's performance across three critical dimensions: (1) Attack effectiveness and transferability, quantified via Attack Success Rate (ASR) on diverse black-box models; (2) Robustness against common defense mechanisms; and (3) Visual quality assessed using standard perceptual metrics. The section structure explicitly delineates experimental objectives and success criteria, establishing a rigorous foundation for subsequent results analysis.

4.1 Experimental Setup

Datasets To ensure experimental comprehensiveness and result comparability, we construct a composite dataset integrating both content images and style reference images.

Content Images: The experiment employs a 2,000-image content set comprising: (1) 1,000 high-resolution images carefully selected from ImageNet[37], and (2) 1,000 images sourced from CIFAR-100.

Style Images: We carefully curated 10 high-quality artworks exhibiting substantial stylistic diversity from the WikiArt dataset as style references. A critical control variable is the complete absence of thematic or categorical overlap between these style images and the content images.

Evaluation Models To rigorously test transferability, we employ a suite of pretrained image classification models exhibiting high architectural diversity. These models alternately serve as either surrogate models for generating adversarial samples or black-box target models for evaluating attack effectiveness during experiments. *ImageNet Models:* ResNet-50, Inception-v3, VGG-19, DenseNet-201, AlexNet, and MobileNet-v2. *CIFAR-100 Models:* ResNet-56, VGG-19, DenseNet-121, MobileNet-v2, and a critical non-convolutional architecture—Vision Transformer (ViT). *Perceptual Loss Computation:* Following neural style transfer conventions, we employ a fixed, pretrained VGG-19 network as the feature extractor for content and style loss calculations.

Baseline We comprehensively compared our method with seven state-of-the-art style transfer-based adversarial attack methods: CAdv[16], ColorFool[19], ACE[17], NCF[35], ACA[18], AdvST[27], and DiffAttack[36].The parameters for these unrestricted attacks follow the corresponding default settings.

Training Configuration To ensure full reproducibility, we disclose all implementation details, including hardware specifications and hyperparameter settings. *Hardware Platform:* All experiments were executed on a workstation equipped with dual NVIDIA GeForce RTX 3090 GPUs. *Hyperparameter Configuration:*

The weights for the components of the composite loss function L_{STAR} were determined after extensive empirical analysis and are specifically set as follows: content loss α=100, style loss β=1000, edge loss γ=5, total variation loss τ=5, and RAP loss λ=1000. The internal optimization loop of RAP runs for $T_{adv} = 40$ iterations, with a step size of α_n=2/255. The overall perturbation is constrained within an L_∞-norm ball with a radius of ϵ=16/255. The determination of these values is fully supported by the ablation studies in Section 4.5.

4.2 Attack performance evaluation

To thoroughly evaluate the effectiveness of our adversarial style transfer method, we assessed its performance in two scenarios: attacking standard benign models and attacking models equipped with adversarial defense mechanisms.

Attacking Benign Models Our experimental results show that the STAR framework generates adversarial samples with high attack success rates (ASR) and strong transferability when attacking various standard benign models. We quantified the ASR against different model architectures on two prominent benchmark datasets: ImageNet and CIFAR-100. Detailed experimental data are presented in Table 1 and Table 3 (Appendix B).

As shown in Tables, evaluations on ImageNet and CIFAR-100 using standard undefended models demonstrate that STAR significantly enhances adversarial sample transferability. With ResNet-50 as the surrogate model, STAR achieved a 99.0% white-box ASR and a 95% black-box transfer ASR on ImageNet, this outperforms baselines methods (e.g., 20% higher than AdvST and DiffAttack). STAR maintained high average ASRs (93% and 95%) when using InceptionV3 or DenseNet201 as surrogates, respectively, while most baselines exhibited notable degradation. On CIFAR-100, STAR also remains at the forefront with average ASRs of 83%-91%, showing consistent efficacy against non-convolutional architectures like ViT. These results validate the superiority of our method in cross-model transferability.

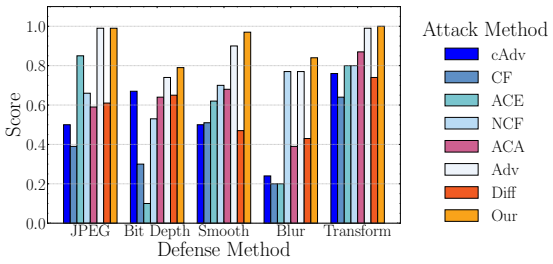

Fig. 3. Comparison of Attack Results Between Our Method and Other Methods Against Adversarial Defense Schemes.

Table 1. Attack Success Rate (ASR) of different methods on various models. All models were evaluated on the ImageNet dataset.

surrogate	Target	Attack							
		cAdv	CF	ACE	NCF	ACA	Adv	Diff	our
Resnet50	Res	0.96	0.90	0.99	0.90	0.88	0.97	0.99	**1.00**
	Inc	0.56	0.18	0.05	0.50	0.60	0.46	0.69	**0.84**
	VGG	0.46	0.24	0.15	0.71	0.60	0.65	0.75	**0.97**
	Den	0.21	0.18	0.15	0.48	0.61	0.48	0.50	**0.91**
	Alex	0.28	0.42	0.39	0.89	0.31	0.89	0.52	**1.00**
	Mob	0.56	0.44	0.51	0.71	0.69	0.82	0.77	**0.99**
Inceptionv3	Res	0.78	0.31	0.16	0.61	0.60	0.46	0.40	**0.90**
	Inc	0.91	0.92	0.93	0.84	0.79	0.89	0.82	**0.99**
	VGG	0.18	0.27	0.18	0.57	0.50	0.47	0.44	**0.92**
	Den	0.18	0.19	0.12	0.50	0.58	0.39	0.23	**0.87**
	Alex	0.09	0.11	0.41	0.83	0.51	0.30	0.26	**1.00**
	Mob	0.30	0.40	0.33	0.64	0.66	0.43	0.44	**0.90**
VGG19	Res	0.51	0.15	0.12	0.64	0.60	0.64	0.78	**0.98**
	Inc	0.49	0.45	0.10	0.49	0.50	0.60	0.70	**0.84**
	VGG	0.97	0.84	0.99	0.92	0.80	0.99	0.97	**1.00**
	Den	0.26	0.16	0.07	0.42	0.58	0.43	0.56	**0.80**
	Alex	0.31	0.62	0.40	0.92	0.54	0.76	0.47	**1.00**
	Mob	0.39	0.34	0.44	0.71	0.64	0.83	**0.81**	**0.81**
Densenet201	Res	0.20	0.38	0.18	0.62	0.52	0.76	0.79	**0.97**
	Inc	0.13	0.28	0.06	0.50	0.65	0.79	0.74	**0.88**
	VGG	0.36	0.44	0.15	0.64	0.48	0.75	0.97	**0.99**
	Den	0.91	0.76	0.97	0.80	0.88	**0.99**	0.98	0.97
	Alex	0.39	0.62	0.45	0.87	0.54	0.90	0.73	**1.00**
	Mob	0.41	0.44	0.61	0.67	0.53	**0.90**	0.70	0.87
MobilenetV2	Res	0.50	0.25	0.17	0.66	0.62	0.56	0.76	**1.00**
	Inc	0.49	0.09	0.05	0.51	0.56	0.49	0.68	**0.75**
	VGG	0.52	0.39	0.15	0.72	0.55	0.37	0.76	**0.93**
	Den	0.33	0.15	0.11	0.47	0.59	0.29	0.49	**0.90**
	Alex	0.53	0.64	0.20	0.92	0.44	0.37	0.48	**1.00**
	Mob	0.96	0.93	0.99	0.92	0.93	0.97	0.98	**1.00**

Bold: best results; underline: second best

Attacking Defended Models To evaluate robustness against defenses, we tested adversarial samples generated by STAR under common image processing techniques (JPEG compression[31], bit-depth reduction and median smoothing[32], Gaussian blur, random transforms). As shown in Figure 3 (ResNet-50 surrogate), STAR consistently maintained high attack success rates (ASR) after these transformations. In contrast, baseline methods exhibited significant performance degradation. This demonstrates STAR's enhanced robustness against practical preprocessing defenses, complementing its superior transferability.

4.3 Interpretability Analysis: Grad-CAM Visualization

To analyze the decision-making process of models under our attack, we employ Gradient-weighted Class Activation Mapping (Grad-CAM)[20] to visualize feature attribution. Grad-CAM generates heatmaps that highlight the image regions most influential to a model's prediction.

As illustrated in Figure 4, a model correctly processing a benign image focuses its attention on the key semantic features of the primary object. In con-

trast, when presented with an adversarial example generated by our method, the model's focus significantly shifts. The heatmaps reveal that the model's attention is diverted from the object to irrelevant background regions, leading to the incorrect classification. This demonstrates that our attack effectively manipulates the model's feature attribution, causing it to base its prediction on non-salient information.

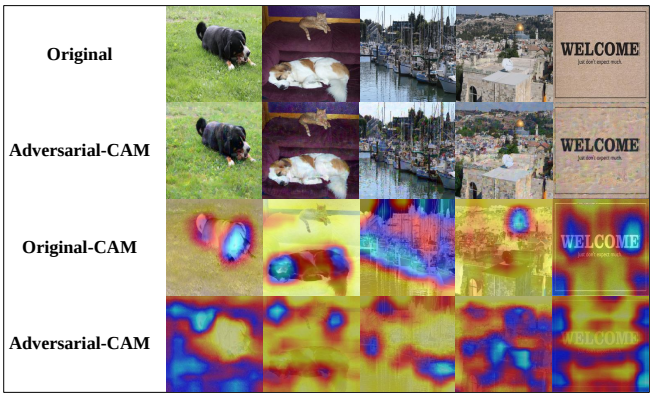

Fig. 4. Grad-CAM Visualization Results.The first row shows the original images.

4.4 Image Quality Analysis

To comprehensively evaluate the visual quality of the generated images, we employed several standard metrics: Structural Similarity Index (SSIM)[5], Learned Perceptual Image Patch Similarity (LPIPS)[38], and Neural Image Assessment (NIMA)[34]. As shown in Figure 5, the STAR method performs chieves SSIM=0.73 and LPIPS=0.24, outperforming baselines by 10% across all metrics while maintaining high attack efficacy: its higher NIMA scores indicate superior aesthetic quality of the generated images; lower LPIPS values reflect better human perceptual similarity; and SSIM scores are also competitive. This demonstrates that STAR doesn't over-optimize for a single metric but rather achieves an effective balance of perceptual similarity under the premise of high structural fidelity and aesthetic quality. This validates the framework's ability to generate visually realistic adversarial samples while ensuring strong attack performance.

4.5 Ablation Studies

To precisely verify the contribution of each core component within the STAR framework, we conducted comprehensive ablation studies. The results indicate that every module in the framework plays an critical and complementary role in balancing attack efficacy, transferability, and visual quality. The quantitative

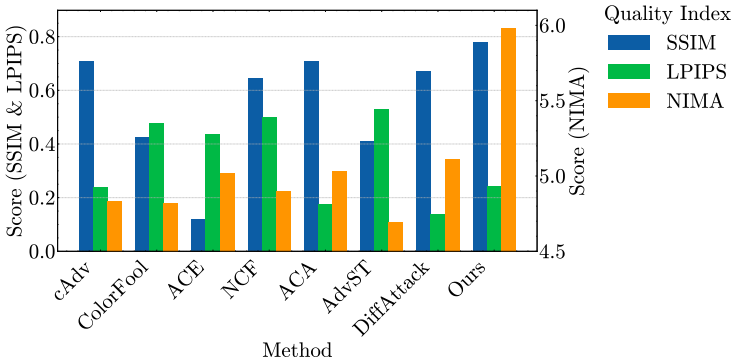

Fig. 5. Comparison of Image Quality between Our STAR and Other Methods

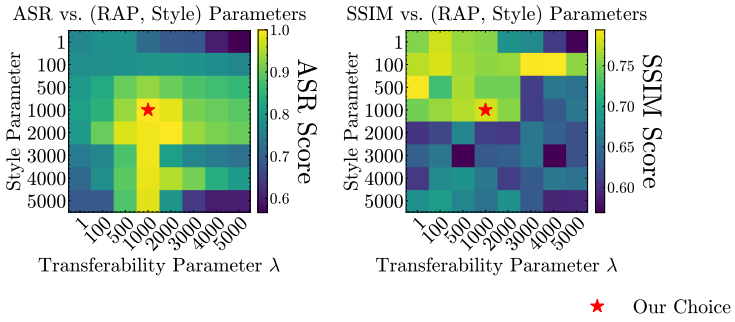

Fig. 6. Impact of different style and Transferability parameters λ on Attack Success Rate (ASR) and SSIM.

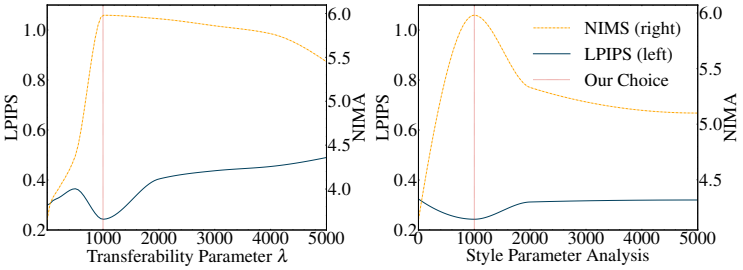

Fig. 7. Impact of different style and Transferability parameters λ on LPIPS and NIMA.

results of removing key modules are presented in Table 2, while a visual analysis of how adjusting critical parameters impacts performance is shown in Figure 6 and Figure 7.

First, the transferability-promoting loss (L_{adv}) was confirmed as the primary driver of transferability. Its removal led to a decrease in Attack Success

Rate (ASR) to 0.71. More critically, it resulted in a significant decline in visual quality (SSIM dropped to 0.51, LPIPS rose to 0.45) and a sharp decay in black-box attack performance, thus confirming RAP's central role in achieving highly transferable attacks. Second, style loss(L_s) is equally indispensable; its removal caused the ASR to drop to 0.72, indicating that style transformation is not merely for aesthetics but is a crucial vehicle for effectively embedding adversarial perturbations. Finally, the removal of edge loss (L_{edge}) had a slightly smaller direct impact on ASR (decreasing to 0.73), but it significantly lowered the image's NIMA score to 5.10, compromising the image's structural integrity and visual fidelity.

Table 2. Quantitative Results of Ablation Study

	ASR ↑	**SSIM** ↑	**LPIPS** ↓	**NIMA** ↑
Full	0.99	0.83	0.17	5.98
Without L_{edge}	0.73	0.79	0.24	5.10
Without L_{style}	0.72	0.73	0.20	5.16
Without L_{adv}	0.85	0.51	0.45	5.00

In summary, these results collectively demonstrate the necessity of each component within the STAR framework, which work synergistically to achieve an effective balance between attack performance and visual quality.

5 CONCLUSION

This work presents STAR, a framework synthesizing artistic style transfer with Reverse Adversarial Perturbation optimization. This integration explicitly guides adversarial search toward geometrically flat regions of the loss landscape, uncovering inherently model-agnostic vulnerabilities. Concurrently, the stylization process coherently integrates transfer-optimized perturbations into natural image textures.

Extensive experiments on the ImageNet and CIFAR-100 datasets demonstrate that STAR achieves high success rates when attacking various black-box models, outperforming many existing stylized attack methods. The generated adversarial examples also maintain robustness when confronted with common image processing defenses. Furthermore, ablation studies confirm the necessity of each component within the framework: the **RAP module is crucial for enhancing transferability**, and the artistic loss terms are key to ensuring visual quality.

Collectively, STAR establishes that directly optimizing for flat-region convergence within a stylized adversarial framework constitutes an effective mechanism for generating highly transferable yet visually plausible attacks.

References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. & Others An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv Preprint ArXiv:2010.11929*. (2020)
2. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Innamorati, S. & He, K. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 16133-16142 (2023)
3. Liu, S., Zeng, F., Li, L., Zhang, H., Geng, A., Li, Y., Su, H., Yang, J., Li, H., Su, H. & Others Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *European Conference On Computer Vision*. pp. 232-252 (2024)
4. Ravi, N., Gelli, S., Vu, T., Kusupati, A., Duarte, K., Mintun, E., Fan, H. & Others Sam 2: Segment anything in images and videos. *ArXiv Preprint ArXiv:2408.00714*. (2024)
5. Wang, Z., Bovik, A., Sheikh, H. & Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions On Image Processing*. **13**, 600-612 (2004)
6. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y. & Dai, J. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. (2024)
7. Kim, W., Hong, S. & Yoon, S. Diverse generative perturbations on attention space for transferable adversarial attacks. *2022 IEEE International Conference On Image Processing (ICIP)*. pp. 281-285 (2022)
8. Kshatri, S. & Singh, D. Convolutional neural network in medical image analysis: a review. *Archives Of Computational Methods In Engineering*. **30**, 2793-2810 (2023)
9. Chen, X., Gao, X., Zhao, J., Ye, K. & Xu, C. Advdiffuser: Natural adversarial example synthesis with diffusion models. *Proceedings Of The IEEE/CVF International Conference On Computer Vision*. pp. 4562-4572 (2023)
10. Zhang, C., Zhou, L., Xu, X., Wu, J. & Liu, Z. Adversarial attacks of vision tasks in the past 10 years: A survey. *ACM Computing Surveys*. (2024)
11. Zhang, J. & Li, C. Adversarial examples: Opportunities and challenges. *IEEE Transactions On Neural Networks And Learning Systems*. **31**, 2578-2593 (2019)
12. Goodfellow, I., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. *ArXiv Preprint ArXiv:1412.6572*. (2014)
13. Jia, W., Lu, Z., Zhang, H., Liu, Z., Wang, J. & Qu, G. Fooling the Eyes of Autonomous Vehicles: Robust Physical Adversarial Examples Against Traffic Sign Recognition Systems. (2022), https://arxiv.org/abs/2201.06192
14. Liang, K. & Xiao, B. Styless: Boosting the transferability of adversarial examples. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 8163-8172 (2023)
15. Wang, X., He, X., Wang, J. & He, K. Admix: Enhancing the transferability of adversarial attacks. *Proceedings Of The IEEE/CVF International Conference On Computer Vision*. pp. 16158-16167 (2021)
16. Bhattad, A., Liu, M., Singh, V., Wu, D. & Hong, Y. Unrestricted adversarial examples via semantic manipulation. *ArXiv Preprint ArXiv:1904.06347*. (2019)
17. Zhao, Z., Liu, Z. & Larson, M. Adversarial color enhancement: Generating unrestricted adversarial images by optimizing a color filter. *ArXiv Preprint ArXiv:2002.01008*. (2020)

18. Chen, Z., Liu, K., Zhang, J., Liu, Z., Zhang, Z., Dong, Y., Liu, W. & Sun, H. Content-based unrestricted adversarial attack. *Advances In Neural Information Processing Systems*. **36** pp. 51719-51733 (2023)
19. Shamsabadi, A., Sanchez-Matilla, R. & Cavallaro, A. Colorfool: Semantic adversarial colorization. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 10116-10125 (2020)
20. Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 618-626 (2017)
21. Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O. & Frossard, P. Universal adversarial perturbations. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 1765-1773 (2017)
22. Zou, Y., Liu, Y., Hu, Y., Li, Y. & Li, R. Flatten long-range loss landscapes for cross-domain few-shot learning. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 23575-23584 (2024)
23. Zhang, Y., Hu, S., Zhang, L., Shi, J., Li, M., Liu, X., Wan, W. & Jin, H. Why does little robustness help? a further step towards understanding adversarial transferability. *2024 IEEE Symposium On Security And Privacy (SP)*. pp. 3365-3384 (2024)
24. Ge, Z., Zhang, F., Zhu, H., Zhang, Z., Zhang, W. & Ge, Y. Improving the transferability of adversarial examples with arbitrary style transfer. *Proceedings Of The 31st ACM International Conference On Multimedia*. pp. 5922-5931 (2023)
25. Ke, W., He, Y., Liu, M. & Wang, J. Improving the transferability of adversarial examples through neighborhood attribution. *Knowledge-Based Systems*. **296** pp. 111909 (2024)
26. Qin, Z., Liu, Z., Yi, J., Zhang, Z., Zhou, F. & Chen, Y. Boosting the transferability of adversarial attacks with reverse adversarial perturbation. *Advances In Neural Information Processing Systems*. **35** pp. 29845-29858 (2022)
27. Wang, X., Guo, J., Zhang, H., Qin, Y., Li, X. & Huang, Q. Advst: Generating unrestricted adversarial images via style transfer. *IEEE Transactions On Multimedia*. **26** pp. 4846-4858 (2023)
28. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z. & Yuille, A. Improving transferability of adversarial examples with input diversity. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 2730-2739 (2019)
29. Dong, Y., Pang, T., Su, H. & Zhu, J. Evading defenses to transferable adversarial examples by translation-invariant attacks. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 4312-4321 (2019)
30. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X. & Li, J. Boosting adversarial attacks with momentum. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 9185-9193 (2018)
31. Das, N., Shanbhogue, M., Chen, S., Hohman, F., Chen, L., Kounavis, M. & Chau, D. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *ArXiv Preprint ArXiv:1705.02900*. (2017)
32. Xu, W., Evans, D. & Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. *ArXiv Preprint ArXiv:1704.01155*. (2017)
33. Duda, R. & Hart, P. Pattern classification and scene analysis. *A Wiley-interscience Publication*. (1973)
34. Talebi, H. & Milanfar, P. NIMA: Neural image assessment. *IEEE Transactions On Image Processing*. **27**, 3998-4011 (2018)

35. Yuan, S., He, Q., Wang, J., Zhang, J., Liu, Z., Chen, S. & Song, L. Natural color fool: Towards boosting black-box unrestricted attacks. *Advances In Neural Information Processing Systems*. **35** pp. 7546-7560 (2022)
36. Chen, J., Zhang, K., Zheng, H., Mao, Z., Wang, W., Liu, C. & Chen, C. Diffusion models for imperceptible and transferable adversarial attack. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. (2024)
37. Krizhevsky, A., Sutskever, I. & Hinton, G. Imagenet classification with deep convolutional neural networks. *Advances In Neural Information Processing Systems*. **25** (2012)
38. Zhang, R., Isola, P., Efros, A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 586-595 (2018)

Appendix

A Pseudo-code for STAR

Algorithm 1 The STAR (Stylized and Transferable Adversarial Robustness) Framework

Input: Original image I_{ori}, Style reference image I_{sty}, True label y, Surrogate model M_s, Laplace operator S **Parameter:** Loss weights: α, β , γ, τ, λ, Total iterations K, step size η **Output:** A stylized and highly transferable adversarial image I_{adv}

Initialize: $I_{adv}, I_{adv} \leftarrow I_{ori}, n_{adv} \leftarrow 0$
while $k = 1 < K$ **do**
  $I_{adv} + n_{adv} \leftarrow$ Compute_RAP_Gradient(I_{adv}, y)
  $\mathbf{L}_c \leftarrow$ ContentLoss(I_{adv}, I_{ori})
  $\mathbf{L}_s \leftarrow$ StyleLoss(I_{adv}, I_{sty})
  $\mathbf{L}_{edge} \leftarrow$ EdgeLoss(I_{adv}, I_{ori})
  $\mathbf{L}_{tv} \leftarrow$ TotalVariationLoss(I_{adv})
  $\mathbf{L}_{adv} \leftarrow$ TotalVariationLoss(I_{adv})
  $\mathcal{L}_{STAR} \leftarrow \mathcal{L}_{art} + \lambda \mathcal{L}_{adv}$
  Update I_{adv}
end while
return I_{adv}

B ASR for Cifar100

Table 3. Performance comparison of different attack methods in terms of Attack Success Rate (ASR). All models were evaluated on the CIFAR-100 dataset.

Surrogate	Target	Attack							
		cadv	CF	ACE	NCF	ACA	Adv	Diff	our
Resnet56	Res	0.37	0.30	0.56	0.50	0.49	0.89	0.86	**0.92**
	VGG	0.39	0.18	0.18	0.71	0.60	0.86	0.74	**0.97**
	Den	0.21	0.36	0.38	0.48	0.33	**0.99**	0.71	**0.99**
	Vit	0.43	0.23	0.11	0.89	0.27	0.69	0.63	**0.79**
	Mob	0.41	0.43	0.57	0.66	0.34	**1.00**	0.77	**1.00**
VGG19	Res	0.24	0.50	0.35	0.57	0.44	0.90	0.70	**0.93**
	VGG	0.34	0.46	0.40	0.50	0.63	0.87	0.90	**0.91**
	Den	0.26	0.53	0.56	0.64	0.57	**0.99**	0.74	**0.99**
	Vit	0.48	0.50	0.38	0.64	0.78	0.67	0.57	**0.83**
	Mob	0.36	0.51	0.35	0.49	0.40	0.90	0.88	**0.92**
Densenet121	Res	0.25	0.10	0.08	0.42	0.33	0.91	0.75	**1.00**
	VGG	0.23	0.09	0.06	0.64	0.59	0.89	0.69	**0.92**
	Den	0.16	0.28	0.55	0.71	0.81	0.88	0.80	**0.94**
	Vit	0.53	0.05	0.10	0.49	0.39	0.66	0.61	**0.76**
	Mob	1.00	0.14	0.12	0.67	0.71	**0.91**	0.70	0.87
MobilenetV2	Res	0.30	0.10	0.56	0.51	0.78	0.87	0.91	**1.00**
	VGG	0.28	0.12	0.18	0.47	0.40	0.80	0.80	**0.86**
	Den	0.26	0.16	0.38	0.67	0.65	0.71	0.79	**1.00**
	Vit	0.47	0.16	0.21	0.62	**0.66**	0.65	0.53	0.58
	Mob	**1.00**	0.23	0.56	0.50	0.73	0.93	0.64	**1.00**

Bold: best results; underline: second best