RobustBlack: Challenging Black-Box Adversarial Attacks on State-of-the-Art Defenses

Anonymous authors
Paper under double-blind review

Abstract

Although adversarial robustness has been extensively studied in white-box settings, recent advances in black-box attacks (including transfer- and query-based approaches) are primarily benchmarked against weak defenses, leaving a significant gap in the evaluation of their effectiveness against more recent and moderate robust models (e.g., those featured in the Robustbench leaderboard). In this paper, we question this lack of attention from black-box attacks to robust models. We establish a framework to evaluate the effectiveness of recent black-box attacks against both top-performing and standard defense mechanisms, on the ImageNet dataset. Our empirical evaluation reveals the following key findings: (1) the most advanced black-box attacks struggle to succeed even against simple adversarially trained models; (2) robust models that are optimized to withstand strong white-box attacks, such as AutoAttack, also exhibits enhanced resilience against black-box attacks; and (3) robustness alignment between the surrogate models and the target model plays a key factor in the success rate of transfer-based attacks.

1 Introduction

Since the discovery that deep learning models are susceptible to minor input disturbances, yielding adversarial examples (AE) (Goodfellow et al., 2015), the development of robust models has become one of the most active topics in the machine learning community. This topic has been thoroughly explored for white-box settings, leading to standardized attacks (Croce & Hein, 2020b) and benchmarks (Croce et al., 2020).

However, realistic evaluations of adversarial attacks remain less explored. It involves scenarios where the attacker has limited or no knowledge of the target model to be attacked. These scenarios are referred to as black-box (or gray-box with partial knowledge (Guo et al., 2018)). Black-box attacks belong to one of the two following settings: When the output of the model provides both category labels and scores, it is referred to as a score-based black-box attack. Conversely, if the output includes only category labels, it is named a decision-based black-box attack (Brendel et al., 2018). Given that most computer vision Application Programming Interface (APIs) (Google, 2023; Imagga, 2023) generally provide image prediction categories alongside scores, most research in a black-box setting considers score-based black-box attacks. Hence, this work focuses on this category.

State of the art (SOTA) black-box attacks leverage one or both of the following mechanisms: (1) adversarial examples transferability and (2) iterative queries with meta-heuristics. Transferability methods craft adversarial examples on a specifically built surrogate model (with white-box access) such that this example should transfer to (i.e. also fools) a target model (to which they have black-box access). Iterative query methods use query feedback from the target model to guide or refine the optimization process to uncover its vulnerabilities.

Transferability is difficult to achieve when the target differs from the surrogate (in terms of architecture, training, or dataset) and the factors behind transferability (or lack thereof) remain an active field of study (Charles et al., 2020; Dong et al., 2018a; Li et al., 2018; Wu et al., 2020a). This is because adversarial attacks crafted on the surrogate lead to examples that maximize the loss function of the surrogate model (Goodfellow et al., 2015; Kurakin et al., 2017), while the target model has a different loss function. Methods

to improve transferability mostly rely on building diversity during the optimization of adversarial examples (Li et al., 2018; Wu et al., 2020a; Xie et al., 2019a).

Iterative-query attacks also face multiple challenges: With increased search space (large images, for example), common query attacks show a significant decrease in the Attack Success Rate (ASR) or require large amounts of queries to achieve sufficient ASR (Mohaghegh Dolatabadi et al., 2020; Huang & Zhang, 2020; Feng et al., 2022). BASES (Cai et al., 2022b), a recent Black-box Surrogate Ensemble Search attack achieved more than 95% ASR with a few queries (<5) on average across a large scope of non-robust models.

Though the initial motivation for black-box attacks and their evaluation protocol is to improve the realism of robustness evaluations, they generally suffer from a common pitfall: they do not consider robust models which have already demonstrated their improved performance against white-box attacks. We hypothesize that the evaluation results of black-box attacks can be misleading because even simple robustification mechanisms can be sufficient to successfully evade black-box attacks. Going further, we want to assess whether innovations in defenses provide positive benefits against black-box attacks like they did in white-box settings.

To the best of our knowledge, this paper is the first to study the effectiveness of black-box attacks against strandardized defenses to demonstrate the need to confront black-box attacks to existing defenses.

Our contributions can be summarized as follows:

1. We demonstrate that simple adversarial training mechanisms reduce the effectiveness of black-box attacks proven effective against standard models, underscoring the need for more advanced black-box attack strategies to address the robustness of real-world models and systems.

2. We show that white-box robustness could be served as a proxy for black-box robustness. Defenses optimized against AutoAttack generalize well to black-box scenarios.

3. We demonstrate that these effective defense mechanisms can inadvertently contribute to enhancing the success rate of black-box attacks. By using robust models as surrogates, attacks can generate adversarial examples more likely to transfer to robust models, leading to an average increase in success rate of 6.49 percentage points across attacks and target models.

Our contributions send a clear message to the adversarial research community regarding the importance of considering white-box defenses in the evaluation of black-box attacks. They also urge the researchers working on defense mechanisms to study how their defenses can help improve transfer-based attacks.

2 Related Work

2.1 Black-Box Attacks: Transfer-Based and Query-Based Methods

Black-box attacks have seen notable advancements, especially within transfer-based and query-based approaches. Transfer-based attacks rely on adversarial examples generated by a surrogate model, which are then tested on a target model. Early approaches like Projected gradient descent (PGD) (Madry et al., 2017) laid the groundwork for such attacks by iteratively optimizing perturbations to maximize the success rate on the target model . This foundational method was enhanced by the Momentum Iterative Fast Gradient Sign Method (MI-FGSM), which incorporated momentum into PGD to improve transferability and stability across iterations (Dong et al., 2018b). Further innovations, such as the Diverse Input Fast Gradient Sign Method (DI-FGSM) and Translation-Invariant FGSM (TI-FGSM), introduced input transformations and image translations, respectively, to increase the robustness and generalizability of adversarial examples across different models (Xie et al., 2019b; Dong et al., 2019).

Subsequent methods like Variance Tuning (VMI, VNI) refined gradient calculations by considering gradient variance across iterations (Wang & He, 2021), leading to more effective adversarial perturbations. Similarly, ADMIX leveraged diverse inputs by mixing the target image with randomly sampled images to generate more transferable attacks (Wang et al., 2021). Skip Gradient Method (SGM) focused on model architecture,

by using skip connections which facilitate the creation of adversarial examples that have high transferability (Wu et al., 2020b). Universal Adversarial Perturbations (UAP) took a different approach by focusing on universality, generating a single small perturbation that can effectively disrupt a classifier across most natural images (Moosavi-Dezfooli et al., 2017). Notably, Ghost networks (GHOST) and Large Geometric Vicinity (LGV) contributed to transferability by modifying skip connections in surrogate models and by exploring variations in surrogate models vicinity, respectively (Li et al., 2020; Gubri et al., 2022).

In query-based attacks, which generally have higher success rates than transfer attacks, methods like Zeroth Order Optimization (ZOO), Decision Boundary Attack (DBA), Natural Evolutionary Strategies (NES),and HopSkipJump rely on iterative query feedback to optimize perturbations, often not utilizing surrogate models to boost efficiency (Brendel et al., 2018; Chen et al., 2017; Ilyas et al., 2018; Chen et al., 2020). Recent work, such as BASES and Transferable Model-based Embedding (TREMBA) (Cai et al., 2022a; Huang & Zhang, 2019), integrates both surrogate representations and query feedback to create highly effective black-box attacks. BASES leverages an ensemble of surrogates with query feedback to dynamically adjust surrogate weighting, while TREMBA utilizes a generator trained with surrogate models, exploring latent space for more effective query attacks.

2.2 Defenses Against AutoAttack

In parallel, adversarial defenses have evolved to mitigate the risks posed by increasingly sophisticated white-box attacks. As defenses became more robust, the need for a reliable, standardized benchmark to accurately assess their effectiveness grew increasingly apparent. In response, AutoAttack emerged as a reliable, parameter-free method for evaluating adversarial robustness, offering a computationally affordable, and a standardized benchmark applicable to various models (Croce & Hein, 2020b). Building on this, Robustbench established a leaderboard specifically for evaluating adversarial robustness, as outlined by (Croce et al., 2020). This leaderboard includes results on the ImageNet large-scale datasets (Deng et al., 2009), and offers a ranking comparison of various defense strategies across specific perturbation budgets. Among the defenses featured on the leaderboard are those proposed by (Salman et al., 2020; Singh et al., 2024; Liu et al., 2024; Bai et al., 2024; Madry et al., 2017), offering a thorough assesment on the effectiveness of the leading defense mechanisms. Traditional defenses, such as Madry's adversarial training, initially combined clean and adversarial data during training, establishing a baseline for robustness (Madry et al., 2017). Recent advancements have incorporated complex pre-training schemes and data augmentations (e.g., ConvStem architectures, RandAugment, MixUp, and CutMix). Singh et al. (2024) made alterations to the convnets architecture, particularly substituting PatchStem with ConvStem, and modified the training scheme to enhance the robustness against previously unencountered l1 and l2 threat models. Liu et al. (2024) further advances the training scheme with large-scale pre-training, combined with label smoothing, weight decay, and Exponential Moving Averaging (EMA) to enhance generalization against unseen adversarial examples. Bai et al. (2024) builds on a giving defense framework (Liu et al., 2024), focusing specifically on balancing clean and robust accuracy through the nonlinear mixing of robust and standard models, addressing the often-seen tradeoff between clean and robust performance.

2.3 Robustness Under Black-Box Setting

In this part, we focus on how our work introduces a novel contribution compared to previous studies on defenses against black-box adversarial attacks in image classification tasks. We discuss five representative works (Papernot et al., 2017; Dong et al., 2020; Mahmood et al., 2021a;b; Ghaffari Laleh et al., 2022), highlighting their key contributions and how our approach builds upon or complements these efforts by addressing gaps and introducing new perspectives in this domain.

Papernot et al. (2017) introduced a black-box attack strategy targeting deep neural network (DNN) models and evaluated its effectiveness against adversarial training and defensive distillation . They varied the perturbation magnitude used during training and the attack phase, and demonstrate that small adversarial training perturbations in training led to gradient masking, which their attack could bypass by doubling the perturbation budget in the attacking phase, whereas larger perturbations in training improved robustness.

In contrast, our work focus on evaluating SOTA adversarially trained defenses against black-box attacks using a fixed small perturbation 4 over 255 in the attack phase.

Dong et al. (2020) evaluated adversarial robustness in image classification tasks against white-box and black-box attacks. Their work addressed various defense techniques such as robust training, input transformation, randomization, certified defenses, and model ensembling. Our study seeks to complement this by focusing specifically on SOTA robust training-based defenses. Moreover, our work explores a broader landscape of black-box attacks by including recent methods such as ADMIX, BASES, and TREMBA. In doing so, we provide an analysis of leading robust training defenses against recent black-box adversarial strategies.

Mahmood et al. (2021a) noted that the majority of existing defenses primarily address white-box attacks, neglecting the crucial aspect of black-box adversarial robustness. They provided a wide evaluation of adversarial defenses with a Convolutional neural network (CNN) architecture on benchmark datasets such as CIFAR-10 (Krizhevsky, 2009) and Fashion-MNIST (Xiao et al., 2017) against 12 attacks. In a follow-up work, Mahmood et al. (2021b) extended this investigation to Vision Transformers (ViTs), evaluating their adversarial robustness against white-box attacks and two black-box attacks on ImageNet. They suggested that ensemble defenses can enhance robustness when attackers lack access to model gradients. Additionally, They examined the transferability of adversarial examples between CNNs and transformers. Building upon and complementing these efforts, we conduct our evaluations on ImageNet, incorporating both CNN and transformer architectures, and including an ensemble defense strategy that combines robust and standard base models. Our approach further investigates transferability between adversarially trained and standardly trained models. We leverage SOTA robust models to challenge the attacks used to evaluate robustness.

In the field of computational pathology, Ghaffari Laleh et al. (2022) examined adversarial robustness with a focus on CNN and transformer models. This work addressed certain adversarial attacks (including Fast Adaptive Boundary (FAB) (Croce & Hein, 2020a), and Square (Andriushchenko et al., 2020)), as well as adversarial training using (PGD) with Dual-Batch Normalization (DBN). Our work extends this line of inquiry by incorporating SOTA adversarially trained models and a more comprehensive suite of black-box attacks, providing additional context for these attacks by increasing their number and offering a more challenging confrontation against SOTA defenses.

3 Success Rate of Black-Box Attacks Against Adversarial Training

As preliminaries, we investigate the effectiveness of 12 black-box adversarial attacks against Madry adversarial training (Madry et al., 2017) on a single, relatively simple ResNet50 model (He et al., 2016). The purpose of these experiments is to check to what extent black-box attacks, typically successful against vanilla models, remain as successful when confronted to adversarially trained models. This preliminary experiment provides us with the first insight into the impact of defenses studied in white-box settings on the effectiveness of black-box attacks.

3.1 Experimental Setup

We evaluated the attack success rate of 11 black-box adversarial attacks against a standardly trained model (vanilla) and an adversarially trained model. We selected a ResNet50 vanilla target and a ResNet50 robust target that underwent adversarial training. We opted for the same model architecture to eliminate the variations that could arise from differing model architectures. We followed the Robustbench evaluation protocol (Croce et al., 2020) and reused the original implementation of each attack. All attacks were untargeted and bound to an L_∞ maximum perturbation with a distance of 4/255. We evaluated the ASR on 5000 examples from the ImageNet validation set, we used the same image identifiers (IDs) as in RobustBench. All the experiments are repeated with three randoms seeds.

As outlined in 2.1, we selected nine transfer-based and two query-based adversarial attacks to provide a comprehensive evaluation of adversarial techniques, balancing foundational, and SOTA approaches. For transfer attacks, our selection emphasizes the diversity of techniques by including methods targeting input transformations (TI, DI), gradient optimization (MI, VNI, VMI, ADMIX), universality (UAP), architectural nuances (GHOST), and geometry (LGV). For query-based attacks, we focused on methods that improve

query efficiency and performance by leveraging surrogate models, and that rely on different optimization techniques. BASES uses a small dimensional search space by modifying the weights given to each surrogate based on the feedback received from querying the target, while TREMBA learns a low-dimensional embedding then performs an efficient search within this space. As discussed in 2.1, all query attacks in this study require surrogate models. The hyperparameters for all transfer and query attacks are detailed in the Appendix A and B, respectively.

We employed surrogate models that are either individually larger than the target model or collectively larger as an ensemble, allowing us to leverage the full capacity of the attack. For all single-surrogate attacks, including the five FGSM-based baseline attacks, ADMIX, UAP, GHOST, and LGV, the surrogate is a WideResNet50-2 (Zagoruyko, 2016) with standard training. For TREMBA, the surrogates are an ensemble of four models with standard training: VGG16 (Simonyan & Zisserman, 2014), ResNet18, SqueezeNet (Iandola et al., 2016), and GoogLeNet (Szegedy et al., 2015) as provided in their repository. For BASES, the surrogates are an ensemble of 10 models with standard training: VGG16_{BN}, ResNet18, SqueezeNet_{11}, GoogLeNet, MnasNet_{10} (Tan et al., 2019), DenseNet161 (Huang et al., 2017), EfficientNet_{B0} (Tan & Le, 2019), RegNet_{Y400MF} (Radosavovic et al., 2020), $\text{ResNeXt101}_{32x8d}$ (Xie et al., 2017), ConvNeXt_{Small} (Liu et al., 2022) from . The models used in this study were obtained from the torchvision library (maintainers & contributors, 2016) and the RobustBench benchmark Croce et al. (2020). We excluded the SGM attack from this section because, at the time of this study, its implementation supports only ResNet and DenseNet surrogates, while our setup uses WideResNet50-2.

3.2 Results

We evaluate the ASR of the black-box attacks against standard and adversarially trained Resnet-50 models. We illustrate the mean ASR using bars, and the standard deviation (STD) as error bars in Figure 1. Against the undefended model, all attacks except UAP achieve more than 40% ASR, with the most recent attacks LGV, BASES and TREMBA reaching 90.11% $\pm$ 0.12, 81.08% $\pm$ 0.49, and 89.56% $\pm$ 0.09, respectively. However, all black-box attacks fail against the robustified model with an ASR of less than 4%. The most recent attacks only bring about marginal improvement over the simplest FGSM attacks. The most effective attack against the defended model in our evaluation, TREMBA achieves 3.56% ASR while TI-FGSM reaches 2.81% ASR. To ensure that the reduced effectiveness of black-box attacks against the adversarially trained ResNet50 model is not due to budget limitation, we quadruple the budget of two of the best black-box attacks in our study (BASES and TREMBA). Even with this increased budget, both attacks still significantly underperform against the adversarially trained ResNet50 compared to their performance on the vanilla ResNet50 model, as detailed in Appendix C.

Insight 1

The effectiveness of black-box attacks drastically decreases against a simple adversarially trained model, with the strongest attack seeing its ASR drops from 89.56% to 3.56%.

4 Effectiveness of White-Box Defenses Against Black-Box Attacks

Given that even the most recent advances in black-box attacks remain ineffective against Madry adversarial training, we investigate whether recent innovations in defense mechanisms (studied in white-box settings) further reduce the effectiveness of black-box attacks. In particular, we want to observe whether these defenses designed against white-box attacks (viz. AutoAttack) and established benchmarks (viz. Robustbench), generalize to black-box attacks. A positive answer to these questions would maintain (and even raise) the motivation to develop defenses in white-box settings, whereas a negative answer would raise an urgent necessity of investigating specific defenses against black-box attacks.

4.1 Experimental Setup

We followed the same protocol as in the previous section, in terms of datasets, and adversarial black-box attacks. We selected nine robust target models from the Robustbench leaderboard. We ensured to have

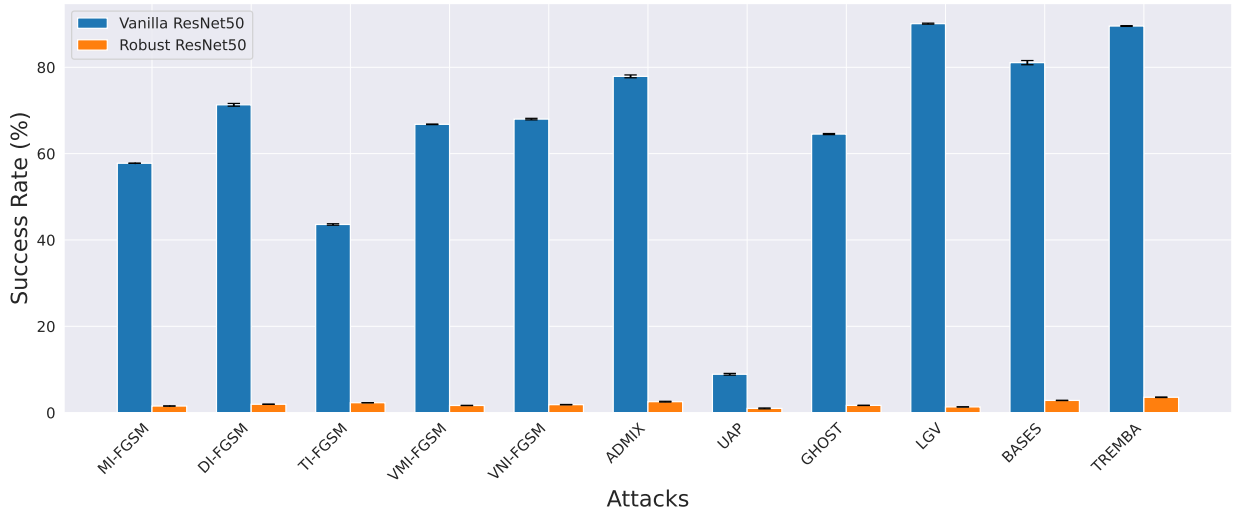

Figure 1: The blue bars (with error bars) show the success rates and standard deviations for the vanilla ResNet50 model, while the orange bars (with error bars) show the results for the robust ResNet50 model.

an even distribution across the robustness spectrum for the nine targets. The robust accuracy of the most robust model, a Swin-L (Liu et al., 2024) is 59.64% (that is, a success rate of AutoAttack of 25.97%), and the robust accuracy of the least robust model, a ResNet18 (Salman et al., 2020) is 25.32% (for a success rate of AutoAttack of 50.53%). Table 1 presents the architecture and the number of parameters of the models. These nine models have been defended using different categories of mechanisms, listed in Table 2.

We kept the same surrogate ensemble for BASES and TREMBA, and we used the surrogate model ResNet50 with standard training for single-surrogate attacks. In addition to the selected black-box attacks, we incorporated the SGM attack due to its compatibility with the ResNet50 surrogate model.

Table 1: Summary of different models from the RobustBench leaderboard (Croce et al., 2020) and their configurations. Legend of the defenses in Table 2.

Rank	Reference	Architecture	Robust Accuracy	Defense	Architecture Type	Parameters
1	Liu et al. (2024)	Swin-L (Liu et al., 2021)	59.56 %	A	Transformer	196.53M
2	Bai et al. (2024)	ConvNeXtV2+Swin-L (Liu et al., 2022; 2021)	58.50 %	A, B	Transformer & Convolution	394.49M
3	Liu et al. (2024)	ConvNeXt-L (Liu et al., 2022)	58.48 %	A	Convolution	197.77M
5	Liu et al. (2024)	Swin-B (Liu et al., 2021)	56.16 %	A	Transformer	87.77M
7	Liu et al. (2024)	ConvNeXt-B (Liu et al., 2022)	55.82 %	A	Convolution	88.59M
12	Singh et al. (2024)	ViT-S+ConvStem (Dosovitskiy et al., 2020; Xiao et al., 2021)	48.08 %	C	Transformer	22.78M
17	Salman et al. (2020)	WideResNet50.2 (Zagoruyko, 2016)	38.14 %	D	Convolution	68.88M
18	Salman et al. (2020)	Resnet50 (He et al., 2016)	34.96 %	D	Convolution	25.56M
21	Salman et al. (2020)	Resnet18 (He et al., 2016)	25.32 %	D	Convolution	11.69M

Table 2: List of the defense mechanisms.

Label	Description
A	Adv training with large data augmentation, regularisation, weights averaging, pretraining
B	Non linear ensemble of two base models robust and vanilla
C	Downsample convolutional layers before subsequent network layers
D	Standard adversarial training

Figure 2: Success rate of blackbox attacks against SoTA defenses.

4.2 Results

We analyze the performance of defenses against black-box attacks by examining the impact of different defense strategies, size of models, and architecture of models, as well as the relationship between AutoAttack success rates and black-box attacks success rates.

Impact of defense mechanism. We categorize the defended models from Table 1 into four distinct families (A, B, C, and D), reflecting the various training strategies employed to enhance model robustness. In Figure 2 we show the ASR of the 12 black-box attacks against the four defenses (robust models). The models are sorted according to the Robustbench leaderboard, with stronger defenses against AutoAttack positioned to the right. Models using the same category of defense mechanisms but differing architectures are grouped in the same color. We observe that stronger defenses against AutoAttack lead to stronger defenses against black-box attacks. For instance, Defense A (red) outperforms Defense C (yellow), which in turn surpasses Defense D (blue) across all black-box attacks. An exception to this trend is observed with the ensemble Conv+Swin Defense, ranked #2 on Robustbench and mixing mechanisms of types A and B. Although ranked second against AutoAttack, it performs similarly to Defense D when subjected to black-box attacks. This is explained by the composition of defense mechanisms A and B, which combines the adversarially trained model from Defense A with a standard (vanilla) model, as explained in Section 2.2. This hybrid nature introduces a vulnerability: while this defense ranks among the best defenses against AutoAttack, the inclusion of a standard model component causes its performance against black-box attacks to align with that of weaker AutoAttack defenses. This vulnerability arises because black-box attacks often exploit vanilla surrogates, making the defense susceptible due to the standard model's presence.

Impact of model size. We compare in Table 3 the success rates of the black-box attacks against similar adversarially trained models. These models do not benefit from advanced defense mechanisms (e.g., synthetic data augmentation, ...) and only differ in size. While larger models with adversarial training tend to

Table 3: Success rate on Resnet models trained by Salman et al. (2020).

Attacks	Salman ResNet18	Salman ResNet50	Salman WideResNet50.2
AUTOATTACK	50.53	43.84	42.87
MI-FGSM	1.28 ± 0.00	1.68 ± 0.00	1.34 ± 0.00
DI-FGSM	1.41 ± 0.02	2.20 ± 0.10	1.99 ± 0.04
TI-FGSM	2.26 ± 0.04	2.81 ± 0.07	2.78 ± 0.10
VMI-FGSM	1.24 ± 0.05	1.89 ± 0.04	1.48 ± 0.05
VNI-FGSM	1.35 ± 0.02	1.72 ± 0.03	1.59 ± 0.05
ADMIX	1.68 ± 0.06	2.68 ± 0.13	2.12 ± 0.08
UAP	0.86 ± 0.10	1.02 ± 0.01	0.93 ± 0.07
SGM	1.25 ± 0.00	1.62 ± 0.00	1.43 ± 0.00
GHOST	1.21 ± 0.18	1.69 ± 0.05	1.49 ± 0.06
LGV	2.27 ± 0.11	2.92 ± 0.01	2.58 ± 0.09
BASES	2.36 ± 0.03	2.83 ± 0.04	2.35 ± 0.06
TREMBA	4.06 ± 0.06	3.56 ± 0.06	3.22 ± 0.06

achieve higher robustness to AutoAttack, the increase in size does not lead to lower success rates. Similar insights are revealed in Table 4, where we compare the robustness of two sizes of ConvNext models and SWIN Transformers. Larger models do not consistently demonstrate stronger robustness to adversarial perturbations in black-box settings, although increased size improves robustness against AutoAttack.

Impact of model architecture. Liu et al. (2024) demonstrate that the best robustness against AutoAttack can be achieved using indiscriminately Convnets or Transformers models. Our results in Table 4 confirm that for a fixed size, both architectures demonstrated similar robustness against black-box attacks.

Table 4: Success rate on very large models trained by Liu et al. (2024).

Attacks	Liu ConvNeXt-B	Liu Swin-B	Liu ConvNeXt-L	Liu Swin-L
AUTOATTACK	28.01	27.14	25.95	25.98
MI-FGSM	0.63 ± 0.00	0.37 ± 0.00	0.49 ± 0.00	0.56 ± 0.00
DI-FGSM	0.90 ± 0.02	0.63 ± 0.03	0.74 ± 0.09	0.85 ± 0.04
TI-FGSM	0.97 ± 0.02	0.79 ± 0.07	1.07 ± 0.03	1.14 ± 0.08
VMI-FGSM	0.74 ± 0.02	0.41 ± 0.01	0.58 ± 0.03	0.55 ± 0.01
VNI-FGSM	0.74 ± 0.04	0.45 ± 0.02	0.49 ± 0.02	0.64 ± 0.04
ADMIX	1.04 ± 0.04	0.72 ± 0.02	0.74 ± 0.01	0.87 ± 0.00
UAP	0.19 ± 0.04	0.16 ± 0.02	0.18 ± 0.06	0.21 ± 0.04
SGM	0.69 ± 0.00	0.37 ± 0.00	0.49 ± 0.00	0.56 ± 0.00
GHOST	0.63 ± 0.03	0.46 ± 0.07	0.49 ± 0.09	0.49 ± 0.04
LGV	0.92 ± 0.07	0.91 ± 0.03	1.07 ± 0.03	1.03 ± 0.03
BASES	1.38 ± 0.06	0.96 ± 0.02	1.16 ± 0.00	1.26 ± 0.02
TREMBA	1.34 ± 0.04	1.14 ± 0.03	1.47 ± 0.01	1.25 ± 0.01

Relationship between AutoAttack success rate and black-box success rate. As shown in Figure 3, the scatter plot illustrates the relationship between the ASR of AutoAttack and that of 12 black-box attacks. Each data point represents the ASR on the y-axis of a given black-box attack (indicated by a color) against a specific robust model (determined by its AutoAttack success rate on the x-axis).

Although the initial ASR of black-box attacks is low (less than 5%), we observe a clear downward trend as robust models improve their effectiveness against AutoAttack, until the best four robust models (with AutoAttack success rate ranging from 25.48% to 27.14%). This vulnerability stems from the lack of correlation between model architecture and robustness, as these models all employ defense (A) with diverse sizes and architectures, as shown in Table 1.

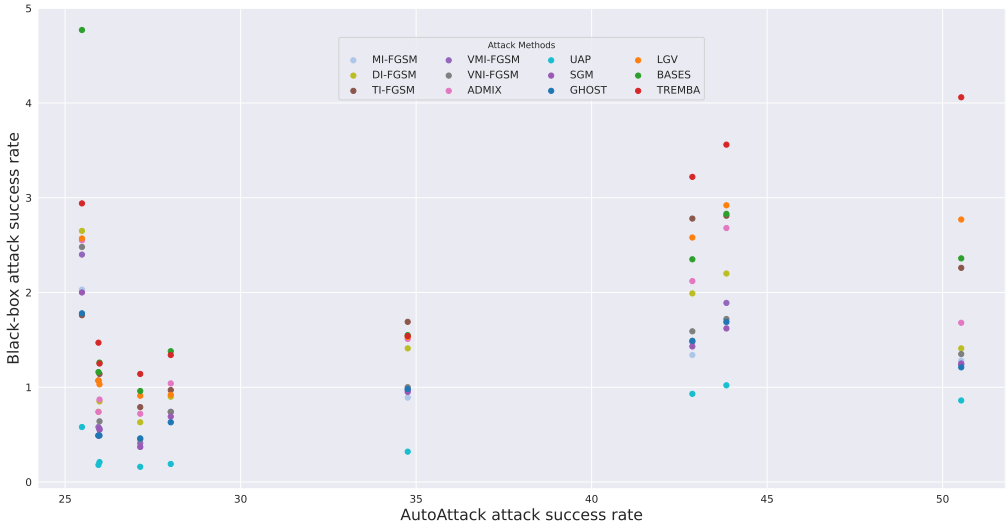

Figure 3: Relation between success rate of AutoAttack and the success rate of black-box attacks.

Furthermore, the best robust model (25.48% AutoAttack success rate) underperforms compared to the other top robust models against black-box attacks. This occurs because it uses the ensemble defense (B), which, as discussed earlier, combines a robust model with a standard model, making it vulnerable to black-box attacks that often exploit surrogates resembling standard models.

Insight 2

SoTA defenses against AutoAttack generally correlate with improved black-box robustness. However, having a large model size does not always guarantee better robustness. Additionally, an ensemble defense that achieves similar robustness to a single defense against AutoAttack may be less effective against black-box attacks, especially those using surrogate models similar to the ensemble's components.

5 Robust surrogates to Improve Black-Box Attacks

The results from the previous section demonstrated that models robust against AutoAttack remain relevant (to some extent) against black-box attacks. A critical concern then is whether these robust "defenses" could not be used by the attacker to improve its black-box attack. Indeed, most of the black-box require designing a surrogate and can therefore leverage the recent advances in model defense to strengthen their attacks. We hypothesize that these new defenses raise security risks by providing stronger surrogates to attackers.

5.1 Experimental Setup

We adhered to the same protocol as in previous sections, maintaining consistency in datasets and adversarial black-box attacks. We considered various configurations regarding the robustness of both the surrogate and target models. For the target models, we selected highly robust models (Swin-L, ConvNeXtV2+Swin-L, ConvNext-L from (Liu et al., 2024) and (Bai et al., 2024)), a moderately robust model (ResNet18 from (Salman et al., 2020)), and a non-robust model (ResNet50). This way, we can measure the effects of using robust surrogates on targets with top defense (A), a simple defense (D) and no defense.

As for surrogate, we used robust and non-robust models. The selection of robust surrogates was based on choosing the highly ranked robust models that are not used as target and are compatible with the attack methods. Hence, we used the robust surrogate RaWideResNet-101-2 from (Peng et al., 2023) for single surrogate attacks, because it is the best defense in Robustbench that supports skip connections (which are essential for the GHOST attack). For ensemble surrogate attacks, we also added the robust surrogate Swin-B from (Liu et al., 2024), as it is the best robust model available, excluding the target models. For single, non-robust surrogate attacks, we used a vanilla WideResNet-101-2 in order to have the same architecture as in the robust surrogate, ensuring fair comparison. For BASES and TREMBA (who use an ensemble of surrogates), we kept the same ensemble of vanilla models as in our previous experiments. We excluded the SGM attack that does not support the WideResNet-101-2 surrogates used in our setup.

5.2 Results

We present in Figure 4 the success rate of black-box attacks using non-robust surrogates and robust surrogates against a non robust target and robust targets.

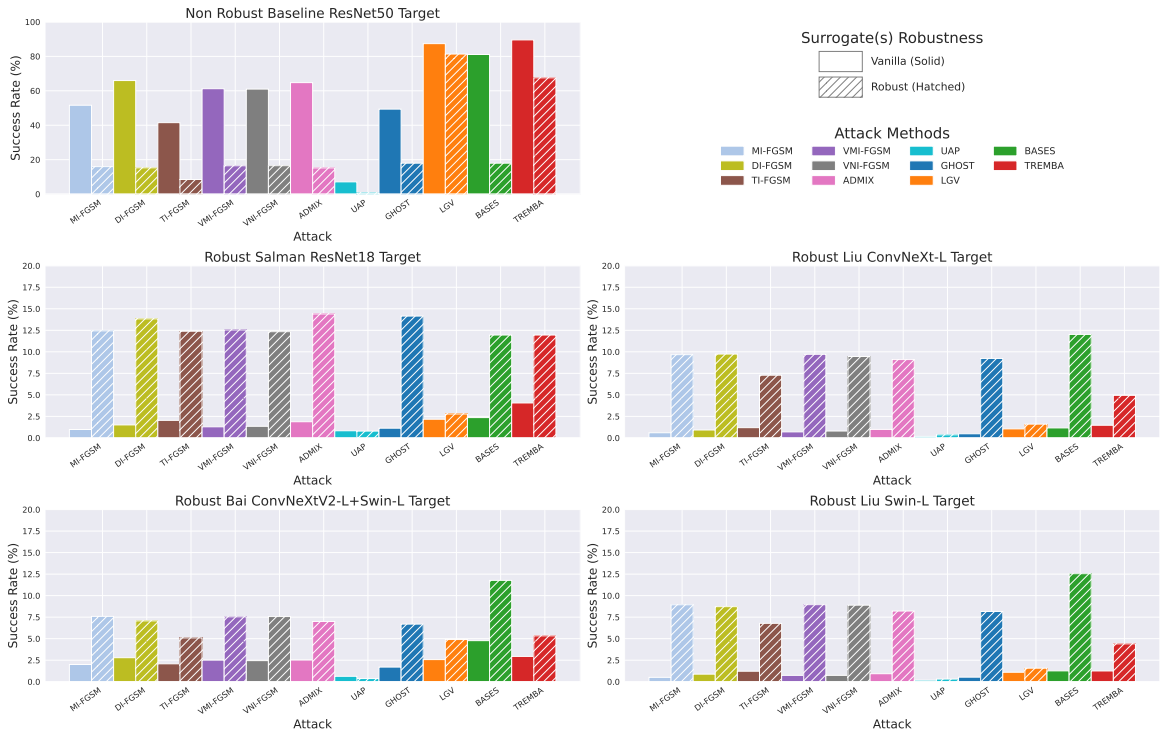

Figure 4: Success rates of black-box attacks using vanilla and robust surrogates against a vanilla target and robust targets.

Non-robust target: The results clearly show that using robust surrogate models is detrimental to attack success rates. Across all black-box attacks, non-robust surrogates consistently outperform robust surrogates, with an average improvement of 35.19 percentage points in attack success rates on the vanilla ResNet50 model. For instance, GHOST improves from 17.95% to 49.36%, BASES from 17.94% to 81.08%, and TREMBA from 67.54% to 89.56%. Even UAP that had the lowest success rate improves from 0.73% to 7.11%. Notably, attacks that leveraged a training phase before the attacking phase (LGV and TREMBA) exhibited a relatively smaller improvement gap compared to other attacks as robust surrogates were already performing well. Specifically, LGV improved from 81.34% to 87.39%, while TREMBA increased from 67.54% to 89.56%.

Robust target: By contrast, robust surrogates yield significant improvement in the success rate of black-box attacks when targeting robust models, resulting in an average improvement of 6.42 percentage points in ASR across attacks and target models. Notably, attacks using robust surrogates on top-tier robust models (e.g., Liu ConvNeXt-L and Liu Swin-L) yield up to 15 times higher success rates than their non-robust counterparts. For instance, when using the GHOST attack to target the highly robust Swin-L model, the robust surrogate achieves a success rate of 8.16%, compared to just 0.53% with the non-robust surrogate. Similarly, BASES improves its success rate from 1.26% to 12.55% against the Swin-L model. An improvement of greater than or equal to 2 percentage points in ASR is consistent across all defense mechanisms and all attacks, except LGV and UAP. We hypothesize that the initial training phase of the LGV models, where a high learning rate is used over additional epochs, potentially weakens the model's robustness during the early sampling phase, which contrasts with GHOST networks that maintain stronger robustness after their creation. This hypothesis is explored in Appendix D.

The results highlight that the choice of surrogate model plays a critical role in the effectiveness of transfer-based black-box attacks, particularly in restricted environments where the robustness of the target is unknown. Robust models, by design, tend to secure blind spot regions in the input space in which vanilla models often fail to predict adversarial examples correctly. As a result, transfer attacks on robust surrogates aim to fool other regions of the input space that are not yet secured by adversarial training, making them better suited for attacking robust models. On the other hand, transfer attacks on vanilla surrogates focuses more on blind spot regions of the input space that are also likely to be mispredicted by other vanilla target models.

Insight 3

The choice of the surrogate model, robust or vanilla, impacts the success of black-box attacks. Robust surrogates are more effective against robust targets, while vanilla surrogates excel against vanilla targets. Attackers can therefore improve attack effectiveness by adaptively selecting surrogates based on the perceived robustness of the target model.

6 Conclusion and Perspectives

This research investigated the effectiveness of black-box adversarial attacks against SoTA and standard robust models using ImageNet, structured around three series of experiments, uncovering key insights.. Our first findings demonstrate that even advanced black-box attacks often struggle against a relatively simple adversarial defense. This suggests that adversarial training can be a highly effective baseline defense for real-world systems accessible via APIs. Consequently, understanding the limitations of current black-box attacks against robust models is crucial for developing more effective attack strategies. Based on our second experiments, we found that defense mechanisms optimized against AutoAttack often generalize effectively to resist black-box attacks, even though performance differences are less pronounced in black-box settings. This finding supports the relevance of AutoAttack as a benchmark for black-box robustness and suggests that current efforts focused on white-box defenses provide a meaningful degree of robustness in black-box scenarios as well – though ensemble defenses may not always translate their AutoAttack robustness to black-box robustness, especially against attacks using similar surrogate models as the ensemble. Finally, we observed that selecting a surrogate with a level of robustness aligned with that of the target can significantly impact the success rate of transfer-based attacks. This finding reveal the possibility for strategic selection of surrogate models when planning black-box attacks. Future work should consider designing black-box attacks that specifically target SOTA robust models to better reflect real-world defensive capabilities. Additionally, further exploration of why SoTA defenses optimized for AutoAttack are also relevant for black-box robustness could yield valuable insights into cross-attack resilience mechanisms. Expanding the study of transfer-based attacks to consider varying levels of robustness between surrogate and target models would also help to deepen our understanding of transferability dynamics. Finally, while this study focused on image classification tasks, future research could extend the analysis to other tasks, such as object detection or segmentation, to understand how black-box robustness may vary across applications. In conclusion, this study demonstrates that SoTA defenses are notably resilient against black-box attacks, underscoring the importance of developing

more targeted attack strategies to effectively challenge these modern robust models. As the field advances, these insights can guide both the creation of more sophisticated black-box attacks and the design of improved defense mechanisms that support the deployment of robust AI systems across a wide range of applications.

Acknowledgment

This research is supported by the Luxembourg National Research Fund under the AFR grant. We extend our gratitude to Martin Gubri (ParameterLabs) for his invaluable support with the LGV implementation.

References

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pp. 484–501. Springer, 2020.

Yatong Bai, Mo Zhou, Vishal M Patel, and Somayeh Sojoudi. Mixednuts: Training-free accuracy-robustness balance via nonlinearly mixed classifiers. *arXiv preprint arXiv:2402.02263*, 2024.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations (ICLR)*, 2018.

Zikui Cai, Chengyu Song, Srikanth Krishnamurthy, Amit Roy-Chowdhury, and M. Salman Asif. Blackbox attacks via surrogate ensemble search, 2022a.

Zikui Cai, Chengyu Song, Srikanth Krishnamurthy, Amit Roy-Chowdhury, and M. Salman Asif. Blackbox attacks via surrogate ensemble search, 2022b. URL `https://arxiv.org/abs/2208.03610`.

Zachary Charles, Harrison Rosenberg, and Dimitris Papailiopoulos. A geometric perspective on the transferability of adversarial directions. In *AISTATS 2019*, 11 2020. URL `http://arxiv.org/abs/1811.03531`.

Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 IEEE Symposium on Security and Privacy (SP)*, pp. 1277–1294. IEEE, 2020.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 15–26, 2017.

Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning*, pp. 2196–2205. PMLR, 2020a.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020b.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting Adversarial Attacks with Momentum. In *CVPR*, pp. 9185–9193, 10 2018a. ISBN 9781538664209. doi: 10.1109/CVPR.2018.00957.

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

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4312–4321, 2019.

Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 321–331, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Yan Feng, Baoyuan Wu, Yanbo Fan, Li Liu, Zhifeng Li, and Shu-Tao Xia. Boosting black-box attack with partially transferred conditional adversarial distribution. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pp. 15095–15104, 2022.

Narmin Ghaffari Laleh, Daniel Truhn, Gregory Patrick Veldhuizen, Tianyu Han, Marko van Treeck, Roman D Buelow, Rupert Langer, Bastian Dislich, Peter Boor, Volkmar Schulz, et al. Adversarial attacks and adversarial robustness in computational pathology. *Nature communications*, 13(1):5711, 2022.

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

Google. Google cloud vision api. `https://cloud.google.com/vision`, 2023. Accessed:2023-08-15.

Martin Gubri, Maxime Cordy, Mike Papadakis, Yves Le Traon, and Koushik Sen. Lgv: Boosting adversarial example transferability from large geometric vicinity, 2022. URL `https://arxiv.org/abs/2207.13129`.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *Proc. Int. Conf. Learn. Representat.*, 2018.

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

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. In *International Conference on Learning Representations*, 2019.

Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. In *Proc. Int. Conf. Learn. Representat.*, 2020.

Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and$<$ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pp. 2137–2146, 2018.

Imagga. Ai-powered image tagging api. "`https://imagga.com/solutions/auto-tagging`", 2023. Accessed:2023-08-15.

Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, 7 2017. URL http://arxiv.org/abs/1607.02533.

Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. Learning Transferable Adversarial Examples via Ghost Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11458–11465, 12 2018. ISSN 2374-3468. doi: 10.1609/aaai.v34i07.6810. URL http://arxiv.org/abs/1812.03413.

Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. Learning transferable adversarial examples via ghost networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11458–11465, 2020.

Chang Liu, Yinpeng Dong, Wenzhao Xiang, Xiao Yang, Hang Su, Jun Zhu, Yuefeng Chen, Yuan He, Hui Xue, and Shibao Zheng. A comprehensive study on robustness of image classification models: Benchmarking and rethinking. *International Journal of Computer Vision*, pp. 1–23, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Kaleel Mahmood, Deniz Gurevin, Marten van Dijk, and Phuoung Ha Nguyen. Beware the black-box: On the robustness of recent defenses to adversarial examples. *Entropy*, 23(10):1359, 2021a.

Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7838–7847, 2021b.

TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016.

Hadi Mohaghegh Dolatabadi, Sarah Erfani, and Christopher Leckie. Advflow: Inconspicuous black-box adversarial attacks using normalizing flows. In *Proc. Int. Conf. Neural Inf. Process. Sys.*, pp. 15871–15884, 2020.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 86–94. IEEE, 2017.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

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

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10428–10436, 2020.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better?, 2020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Naman Deep Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2820–2828, 2019.

Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1924–1933, 2021.

Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16158–16167, 2021.

Spencer Woo. torchattack: Adversarial attacks in pytorch. `https://github.com/spencerwooo/torchattack/tree/main`, 2024. Accessed: 2024.

Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets. In *ICLR*, 2 2020a. URL `https://arxiv.org/abs/2002.05990v1http://arxiv.org/abs/2002.05990`.

Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. *International Conference on Learning Representations*, 2020b.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34:30392–30400, 2021.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pp. 2725–2734, 3 2019a. ISBN 9781728132938. doi: 10.1109/CVPR.2019.00284. URL `http://arxiv.org/abs/1803.06978`.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019b.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Sergey Zagoruyko. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Appendix

A Hyperparameters of Transfer Attacks

Tables 5 and 6 provide a comprehensive overview of the hyperparameters used for the ten transfer-based adversarial attacks in our experiments. They categorize the key hyperparameters into two groups: common hyperparameters shared across most attacks and attack-specific hyperparameters that define the unique characteristics of each method. Most of the attack hyperparameters were kept at their default values, as specified in the repositories or libraries that provide the implementations of adversarial attacks (Kim, 2020; Woo, 2024).

A.1 Common Hyperparameters

Table 5 provides the default values and explanations for these common hyperparameters, such as the step size (Alpha), momentum factor (Decay), and number of iterations (Steps), which are essential for iterative attacks as they control the perturbation magnitude and the optimization process.

A.2 Attack-Specific Hyperparameters

In addition to the shared hyperparameters, each attack incorporates its own unique hyperparameters tailored to its specific mechanisms. Table 6 summarizes these attack-specific hyperparameters, listing the default values, explanations, and the corresponding attacks to which they apply.

B Hyperparameters of Query Attacks

Tables 7 and 8 provides an overview of the hyperparameters used in BASES and TREMBA, respectively. Most of these hyperparameters were kept at their default values, as specified in their repositories, since our experiments aimed to evaluate their performance under standard configurations. For BASES, key hyperparameters include the ensemble size (N_wb), the type of fusion for surrogate models (Fuse), and the number of weight updates (Iterw). Similarly, TREMBA, which operates in two phases (training and attacking), relies on hyperparameters such as the number of training epochs (Epochs), the number of attack iterations (Num_Iters), the number of samples for Natural Evolution Strategy (NES), and latent space parameters like the Gaussian noise standard deviation (Sigma).

C Increasing the Budget of Black-Box Attacks

We quadrupled (multiplied by 4) the budget of the best black-box attacks in section 3 (BASES and TREMBA) against the robust ResNet50 model, to reassure that black-box attacks reduced effectiveness (compared to their effectiveness against the vanilla ResNet50 model) is not due to a limited budget. For BASES, we increased the number of inner iterations of PGD within the perturbation machine (Iters) to 40. For TREMBA, we increased the Number of iterations for NES (Num_Iters) to 800.

Figure 5 presents the success rates of BASES and TREMBA attacks against the vanilla ResNet-50 model with standard attack budget, and against the adversarially trained ResNet-50 model under both standard

Table 5: Common hyperparameters for transfer-based attacks.

PARAMETER	DEFAULT VALUE	EXPLANATION	APPLIES TO
Alpha	2/255	Step size or update rate for perturbation.	All attacks except UAP
Decay	1.0	Momentum factor.	All attacks except UAP
Steps	10	Number of iterations.	All attacks except UAP
Epsilon (eps)	4/255	Maximum perturbation for adversarial example.	All attacks

Table 6: Attack-specific hyperparameters for transfer-based attacks.

PARAMETER	DEFAULT VALUE	EXPLANATION	APPLIES TO
Resize Rate	0.9	Rate at which the input images are resized.	DIFGSM, TIFGSM
Diversity Prob	0.5	Probability to apply input diversity.	DIFGSM, TIFGSM
Kernel Name	Gaussian	Type of kernel used.	TIFGSM
Kernel Length	15	Length of the kernel.	TIFGSM
Sigma (nsig)	3	Radius of the Gaussian kernel.	TIFGSM
N	5	Number of samples in the neighborhood.	VMI, VNI
Beta	1.5	Upper bound for neighborhood.	VMI, VNI
Prior Type	no_data	Type of prior used for attack optimization.	UAP
Patience Interval	5	Number of iterations to wait before checking convergence.	UAP
Gamma	0.5	Decay factor for gradients from skip connections.	SGM
LGV Epochs	5	Number of epochs for training the LGV models.	LGV
LGV Models Epoch	8	Number of models collected per epoch.	LGV
LGV Learning Rate	0.1	Learning rate for the LGV training phase.	LGV
LGV Batch Size	256	Batch size for the LGV training phase.	LGV
Randomized Modulating Scalar	0.22	Drawn from the uniform distribution 1-scalar, 1+scalar.	GHOST
Portion	0.2	Portion for the mixed image.	Admix
Size	3	Number of randomly sampled images.	Admix

Table 7: Hyperparameters for the BASES attack.

PARAMETER	DEFAULT VALUE	EXPLANATION
Epsilon (eps)	4	Perturbation bound magnitude (0 to 255 pixel range)).
N_wb	10	Number of models in the ensemble.
Bound	linf	Perturbation bound norm type.
Iters	10	Number of inner iterations.
Fuse	loss	Fuse method, e.g., loss or logit.
Loss_name	CW	Loss function used for optimization.
Algo	PGD	White-box algorithm used for the perturbation machine
X	3	Factor to scale the step size alpha.
Learning Rate (lr)	0.005	Learning rate for weight updates.
Iterw	20	Number of iterations to update weights.

and quadrupled attack budgets. Both attacks underperform against the adversarially trained ResNet50 under quadrupled budgets compared to the vanilla model with standard budgets. For instance, BASES achieves 2.56% ASR on the robust ResNet50 Budget (x4) versus 81.08% on the vanilla ResNet50. Similarly, TREMBA's ASR is at 5.24% on the robust ResNet50 Budget (x4) compared to 89.56% on the vanilla ResNet50. These findings confirm that our earlier results in section 3 were not a consequence of budget limitations.

D AutoAttack Robustness of LGV and GHOST Models

To evaluate the impact of the LGV weight collection phase, and the GHOST perturbing phase on model robustness, we collected 10 models after finishing the first phase for both LGV and GHOST, prior to proceeding to the attacking phase. Specifically, the LGV variations were obtained by training the surrogate model for additional epochs using a high learning rate, during which the weights of the models were collected periodically. For GHOST networks, the variations were obtained by introducing a random perturbation to the scaling factor in the skip connections. Using the robust surrogate model *WideResNet-101-2* introduced in (Peng et al., 2023) and described in Section 5, we evaluated the robustness of the resulting LGV models and compared it with the robustness of GHOST models. The robust accuracies for the collected models are reported in Table 9.

Table 8: Hyperparameters for the TREMBA attack.

PARAMETER	DEFAULT VALUE	EXPLANATION
Epsilon (eps)	8/255 (Training Phase), 4/255 (Attacking Phase)	Perturbation bound magnitude.
Learning Rate (G)	0.01 (Training Phase), 5.0 (Attacking Phase)	for gradient updates, for latent updates
Momentum	0.9 (Training Phase), 0.0 (Attacking Phase)	for SGD, for attack on the embedding space.
Num_images	49k (Training Phase)	Size of the training set
Epochs	500	Number of epochs for Training.
Schedule	10	Epochs between Training learning rate decay.
Gamma	0.5	Training Learning rate decay factor.
Margin	200.0 (Training Phase), 5.0 (Attacking Phase)	Margin for the loss function.
Sample_Size	20	Number of samples for NES
Num_Iters	200	Number of iterations for NES.
Sigma	1.0	Standard deviation of the Gaussian noise for NES.
Learning Rate Min	0.1	Minimum learning rate in the Attacking Phase.
Learning Rate Decay	2.0	Learning rate decay factor in the Attacking Phase.
Plateau_Length	20	Number of iterations to monitor the objective function
Plateau_Overhead	0.3	Allowed loss threshold to change before the decay

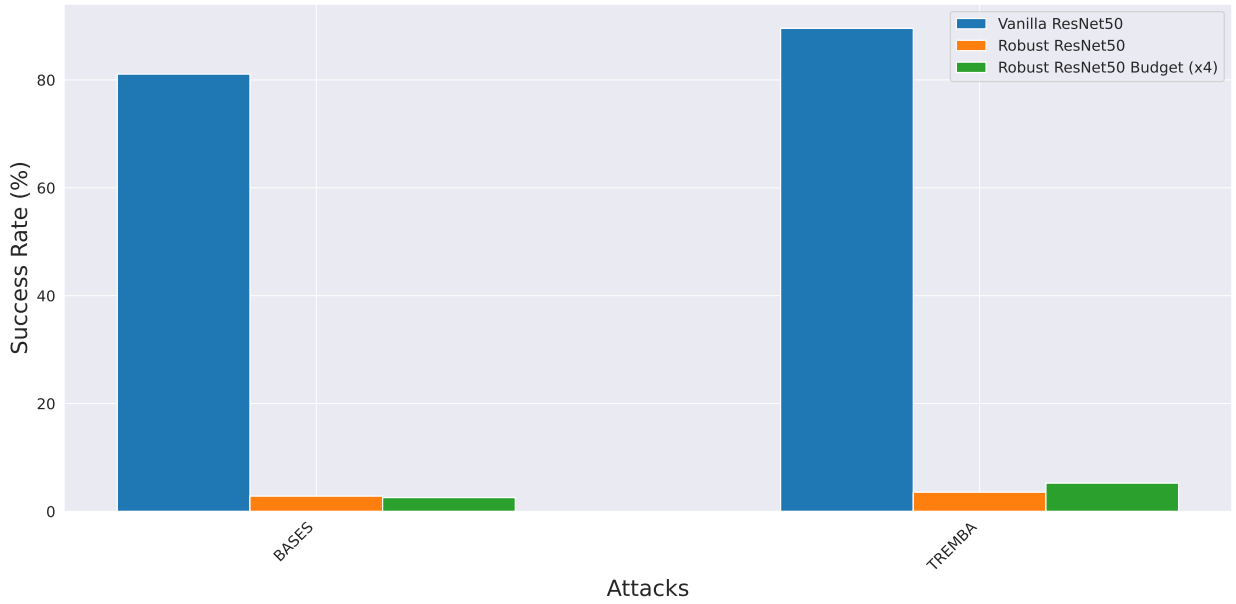

Figure 5: The blue bars show the success rate for the vanilla ResNet50 model, while the orange bars show the results for the robust ResNet50 model, and the green bars show the success rate for the ResNet50 model by quadrupling the budget of the attacks.

Given that both LGV and GHOST models are derived from the same robust surrogate model, we observed a difference in their robustness. The robust accuracies for all LGV models is 0% (indicating 100% attack success rates), while the robust accuracies for GHOST networks is on average 51.54% across the 10 collected models. This confirms that the additional training phase of the LGV models with a high learning rate decreases the robustness of the surrogate models. Consequently, this explains why the performance improvement for LGV models when using a robust surrogate against robust targets is lower compared to the GHOST networks. As a good practice, it is essential to evaluate the robustness of surrogate models prior to the attacking phase. This ensures that any preliminary training phase does not inadvertently weaken the surrogate robustness.

Table 9: Robust accuracies of LGV and GHOST models before the attacking phase.

	Num 1	**Num 2**	**Num 3**	**Num 4**	**Num 5**	**Num 6**	**Num 7**	**Num 8**	**Num 9**	**Num 10**	**Average**
GHOST	51.39	51.86	51.74	51.02	51.58	50.76	52.39	52.81	51.11	50.78	51.54
LGV	0	0	0	0	0	0	0	0	0	0	0

Table 10: Success rates of black-box attacks against one vanilla, and one robust ResNet50 model.

MODELS	**MI-FGSM**	**DI-FGSM**	**TI-FGSM**	**VMI-FGSM**	**VNI-FGSM**	**ADMIX**	**UAP**	**GHOST**	**LGV**	**BASES**	**TREMBA**
Vanilla ResNet50	57.77 ±0.00	71.31 ±0.32	43.59 ± 0.17	66.79 ± 0.07	68.00 ± 0.18	99.95 ± 0.00	8.89 ± 0.20	64.53 ± 0.13	90.11 ± 0.12	81.08 ±0.49	89.56 ±0.09
Robust ResNet50	1.53 ± 0.00	1.94 ± 0.03	2.3 ± 0.01	1.66 ± 0.03	1.85 ± 0.04	2.54 ± 0.09	0.98 ± 0.10	1.68 ±0.03	1.33 ±0.01	2.83 ±0.04	3.56 ±0.06

E Detailed Tabular Results

We provide the tabular representation corresponding to the results presented in the main paper. Tables 10, 11, 12, 13 represent the success rates for section 3, 4, and 5 with vanilla surrogates and robust surrogates, respectively. These tables display the mean and standard deviation of success rates computed across three random seeds.

Table 11: Success rates of white-box AutoAttack and black-box attacks against robust targets from Robustbench.

RANK	MODELS	AUTOATTACK	MI-FGSM	DI-FGSM	TI-FGSM	VMI-FGSM	VNI-FGSM	ADMIX	UAP	SGM	GHOST	LGV	BASES	TREMBA
21	Salman ResNet18	50.53	1.28 ±0.00	1.41 ±0.02	2.26 ± 0.04	1.24 ± 0.05	1.35 ± 0.02	1.68 ± 0.06	0.86 ± 0.10	1.25 ± 0.00	1.21 ±0.18	2.27 ±0.11	2.36 ±0.04	4.06 ±0.06
18	Salman ResNet50	43.84	1.68 ± 0.00	2.2 ± 0.10	2.81 ± 0.07	1.89 ± 0.04	1.72 ± 0.03	2.68 ± 0.13	1.02 ± 0.01	1.62 ± 0.00	1.69 ±0.05	2.92 ±0.01	2.83 ±0.04	3.56 ±0.06
17	Salman WideResNet50.2	42.87	1.34 ± 0.00	1.99 ± 0.04	2.78 ± 0.10	1.48 ± 0.05	1.59 ± 0.05	2.12 ± 0.08	0.93 ± 0.07	1.43 ± 0.00	1.49 ±0.06	2.58 ±0.09	2.35 ±0.06	3.22 ±0.06
12	Sing ViT-S, ConvStem	34.76	0.89 ± 0.00	1.41 ± 0.05	1.69 ± 0.05	0.97 ± 0.04	1.00 ± 0.00	1.51 ± 0.05	0.32 ± 0.06	0.95 ± 0.00	0.98 ±0.03	1.53 ±0.08	1.55 ±0.00	1.54 ±0.03
7	Liu ConvNeXt-B	28.01	0.63 ± 0.00	0.9 ± 0.02	0.97 ± 0.02	0.74 ± 0.02	0.74 ± 0.04	1.04 ± 0.04	0.19 ± 0.04	0.69 ± 0.00	0.63 ±0.03	0.92 ±0.07	1.38 ±0.06	1.34 ±0.04
5	Liu Swin-B	27.14	0.37 ± 0.00	0.63 ± 0.03	0.79 ± 0.07	0.41 ± 0.01	0.45 ± 0.02	0.72 ± 0.02	0.16 ± 0.02	0.37 ± 0.00	0.46 ±0.07	0.91 ±0.03	0.96 ±0.02	1.14 ±0.03
3	Liu ConvNeXt-L	25.95	0.49 ± 0.00	0.74 ± 0.09	1.07 ± 0.03	0.58 ± 0.03	0.49 ± 0.02	0.74 ± 0.01	0.18 ± 0.06	0.49 ± 0.00	0.49 ±0.09	1.07 ±0.03	1.16 ±0.00	1.47 ±0.01
2	Bai ConvNeXtV2-L, Swin-L	25.48	2.03 ± 0.00	2.65 ± 0.03	1.76 ± 0.05	2.40 ± 0.05	2.48 ± 0.03	2.55 ± 0.05	0.58 ± 0.05	2.00 ± 0.00	1.78 ±0.16	2.57 ±0.10	4.77 ±0.13	2.94 ±0.12
1	Liu Swin-L	25.98	0.56 ± 0.00	0.85 ± 0.04	1.14 ± 0.08	0.55 ± 0.01	0.64 ± 0.04	0.87 ± 0.00	0.21 ± 0.04	0.56 ± 0.00	0.49 ±0.04	1.03 ±0.03	1.26 ±0.02	1.25 ±0.01

Table 12: Success rates of black-box attacks using vanilla surrogates against a vanilla target and robust targets.

Models	MI-FGSM	DI-FGSM	TI-FGSM	VMI-FGSM	VNI-FGSM	ADMIX	UAP	GHOST	LGV	BASES	TREMBA
Standard ResNet50	51.62 ± 0.0	65.95 ± 0.75	41.5 ± 0.35	61.22 ± 0.18	61.10 ± 0.13	64.75 ± 0.17	7.11 ± 0.25	49.36 ± 0.14	87.39 ± 0.26	81.08 ± 0.49	89.56 ± 0.09
Salman ResNet18	0.98 ± 0.0	1.51 ± 0.03	2.04 ± 0.06	1.29 ± 0.05	1.35 ± 0.06	1.88 ± 0.08	0.83 ± 0.03	1.13 ± 0.08	2.17 ± 0.05	2.36 ± 0.04	4.06 ± 0.06
Liu ConvNeXt-L	0.6 ± 0.0	0.92 ± 0.05	1.19 ± 0.1	0.7 ± 0.02	0.8 ± 0.04	0.99 ± 0.03	0.18 ± 0.04	0.48 ± 0.09	1.05 ± 0.02	1.16 ± 0.0	1.47 ± 0.01
Liu ConvNeXtV2-L+Swin-L	2.0 ± 0.0	2.79 ± 0.07	2.08 ± 0.04	2.51 ± 0.02	2.45 ± 0.05	2.52 ± 0.06	0.63 ± 0.03	1.7 ± 0.19	2.57 ± 0.07	4.77 ± 0.13	2.94 ± 0.12
Liu Swin-L	0.51 ± 0.0	0.88 ± 0.03	1.22 ± 0.05	0.74 ± 0.02	0.74 ± 0.04	0.93 ± 0.05	0.18 ± 0.0	0.53 ± 0.03	1.1 ± 0.11	1.26 ± 0.02	1.25 ± 0.01

Table 13: Success rates of black-box attacks using robust surrogates against a vanilla target and robust targets.

Models	MI-FGSM	DI-FGSM	TI-FGSM	VMI-FGSM	VNI-FGSM	ADMIX	UAP	GHOST	LGV	BASES	TREMBA
Standard ResNet50	15.96 ± 0.0	15.28 ± 0.21	8.48 ± 0.18	16.47 ± 0.02	16.51 ± 0.04	15.22 ± 0.11	0.73 ± 0.0	17.95 ± 0.36	81.34 ± 0.43	17.94 ± 0.00	67.54 ± 0.13
Salman ResNet18	12.44 ± 0.0	13.83 ± 0.14	12.39 ± 0.04	12.54 ± 0.02	12.34 ± 0.05	14.4 ± 0.17	0.79 ± 0.0	14.15 ± 0.22	2.78 ± 0.21	11.95 ± 0.00	10.58 ± 0.14
Liu ConvNeXt-L	9.67 ± 0.0	9.74 ± 0.0	7.28 ± 0.04	9.71 ± 0.01	9.46 ± 0.00	9.1 ± 0.06	0.36 ± 0.0	9.23 ± 0.13	1.6 ± 0.11	12.02 ± 0.00	4.94 ± 0.06
Liu ConvNeXtV2-L+Swin-L	7.59 ± 0.0	7.07 ± 0.09	5.09 ± 0.04	7.54 ± 0.03	7.57 ± 0.00	7.0 ± 0.02	0.37 ± 0.0	6.69 ± 0.08	4.9 ± 0.06	11.77 ± 0.01	5.33 ± 0.11
Liu Swin-L	8.94 ± 0.0	8.72 ± 0.04	6.78 ± 0.07	8.95 ± 0.01	8.89 ± 0.02	8.21 ± 0.01	0.31 ± 0.0	8.16 ± 0.09	1.57 ± 0.04	12.55 ± 0.00	4.4 ± 0.08