S^4ST: A Strong, Self-transferable, faSt, and Simple Scale Transformation for Transferable Targeted Attack

Yongxiang Liu, Bowen Peng, Li Liu, Xiang Li

Abstract—Transferable targeted adversarial attacks against deep neural networks have been proven significantly more challenging than untargeted ones, yet they remain relatively underexplored in current research. Existing solutions either suffer from the requirements for extensive additional data and prolonged training for each target label (generative and universal attacks) or risk severe overfitting to the surrogate model (gradient-based attacks). These limitations significantly impede their practicality and may lead to underestimating the potential risks of targeted attacks. This paper sheds new light on performing highly efficient yet transferable targeted attacks leveraging the simple gradient-based baseline. Our research underscores the critical importance of image transformations within gradient calculations, marking a shift from the prevalent emphasis on loss functions to address the gradient vanishing problem. Moreover, we have developed two effective blind estimators that facilitate the investigation and design of transformation strategies to enhance targeted transferability under black-box conditions. The adversarial examples' self-transferability to geometric transformations such as rotation, scaling, and shear has been identified as strongly correlated with their black-box transferability, featuring these basic operations as potent yet overlapped proxies for facilitating targeted transferability. The surrogate self-alignment assessments further highlight the simple scaling transformation's exceptional efficacy, which rivals that of most existing advanced methods. Building on these insights, we introduce a scaling-centered transformation strategy termed Strong, Self-transferable, faSt, and Simple Scale Transformation (S^4ST) to enhance transferable targeted attacks. In experiments conducted on the ImageNet-Compatible benchmark dataset, our proposed S^4ST attains a state-of-the-art (SOTA) average targeted transfer success rate across various challenging black-box models, outperforming the previous leading method by over 14% while requiring only 25.7% of the execution time. Additionally, our approach eclipses SOTA generative attacks by a considerable margin and exhibits remarkable effectiveness against real-world APIs. This work marks a significant leap forward in the field of transferable targeted adversarial attacks, revealing the realistic threats they pose and providing an effective and efficient generation method for future research.

Index Terms—Adversarial examples, gradient-based attack, transferability, targeted adversarial attack, deep neural networks

arXiv:2410.13891v1 [cs.CR] 13 Oct 2024

1 Introduction

Deep Neural Networks (DNNs) have been shown to be vulnerable to adversarial examples (AEs) [1], which refer to images with small, intentionally crafted perturbations that are usually imperceptible to the human visual system but can fool DNNs. The presence of AEs raises significant concerns when deploying DNNs in security-critical domains, including autonomous driving, face recognition, and medical image analysis. Therefore, finding, understanding, and defending AEs have received significant attention [2]–[5]. AEs exhibit an intriguing property known as transferability [6], which means that AEs designed for one DNN can successfully deceive other DNNs with distinct architectures. The transferability of AEs facilitates the implementation of strict black-box attacks where queries to the victims are forbidden. As a result, significant attention has been paid to enhancing these examples' transferability in the last decade [7].

While transferable untargeted attacks have been extensively investigated, yielding a substantial body of research [8]–[13], Targeted Transferable Attacks (TTAs) are significantly more challenging and consequently less explored. This is primarily due to the increased complexity of directing a model to a specific target class rather than merely causing a misclassification to any class other than the original. More specifically, targeted transferability hinges on the generation of AEs that subtly align with the visual semantics of the targeted label [14], [15], as opposed to creating untargeted perturbations that just disrupt the existing visual cues. Achieving effective TTAs often entails optimizing a universal perturbation [14], [16] across tens of thousands of additional data points or indirectly through the generative learning [15], [17], [18]. Such approaches, known as *resource-intensive TTAs* [19], are significantly hampered in their practicality due to their reliance on extensive datasets (often a subset of the victims' training dataset), time-consuming training processes and their inherent limitation to single-target labels. In contrast, we focus on *simple TTAs*, which involve executing gradient descent on any (sample, label) pair, offering both efficiency and flexibility. However, these are plagued by severe overfitting and poor transferability [8], [20].

The limitations of simple TTAs have been attributed to the gradient vanishing issue associated with the cross-entropy loss function in targeted optimizations, prompting the proposal of a series of alternative loss functions aimed at circumventing or mitigating this issue [19], [21]–[23]. In this paper, however, we shift our focus to the role of input transformations, which, despite being inherently critical within simple TTAs, have yet to be underexplored. An illustrative example of a transformation-involved attack is

• *The authors are with the College of Electronic Science and Technology, National University of Defense Technology (NUDT), Changsha 410073, China. Email: lyx_bible@sina.com, pbow16@nudt.edu.cn, liuli_nudt@nudt.edu.cn, lixiang01@vip.sina.com.*

Corresponding Author: Xiang Li.
This work was supported partially by the National Key Research and Development Program of China under Grant 2021YFB3100800, the National Natural Science Foundation of China under Grant 62376283, 61921001, 62022091, and 62201588.

(a) Results by SOTA resource-intensive method, M3D [18]. When the generator is more universal to to more additional samples, it exhibits stronger transferability while inevitably consuming formidable time for training.

(b) Results by TI-MI-FGSM (TMI) [8], [9] with simple scaling transformation and our S^{4}ST (simple TTA). When the perturbation is more transferable to more transformed self-copies, it exhibits stronger transferability while maintaining efficiency.

Fig. 1. Comparison between (a) the SOTA resource-empowered attack and (b) our findings-supported simple attack. **Observations:** Scaling-based simple attack achieves performance on par with M3D and our S^{4}ST further excels by significant margins in both terms of effectiveness (81.9% *vs* 61.9% average tSuc) and efficiency (24.4 *vs* 606.1 minutes). And our method further excels at its flexibility in handling arbitrary target labels. **Settings:** The targeted transferability is measured by the targeted transfer success rate (tSuc) averaged over 14 black-box models and 1,000 images, where the surrogate model is ResNet-50 and the target label is `Street sign`. The time consumption includes both training and attacking for all 1,000 images. The perturbation budget is $\ell_\infty = 16/255$. Extended results comparing the proposed S^{4}ST and resource-intensive methods are in Section 5.4.

presented in Fig. 1b. A transformation facilitates simple TTAs through a dual mechanism that diversifies features and softens the prediction distribution, naturally alleviating the gradient vanishing problem. Through experimental combinations of various transformations and advanced loss functions, we observe that all loss functions are practical only when coupled with transformations. This underscores the pivotal role of transformations in simple TTAs and urges further investigation into understanding and unlocking their untapped benefits for further research.

To this end, we aim to find certain black-box measures that enable the design of highly potent transformation methods, *e.g.*, identifying a strategy to combine various basic transformations without accessing the victim models. We began by re-examining several existing consensuses, which are frequently discussed in the context of transformation-based transfer attacks for both untargeted and targeted scenarios. These include the transformation-induced data diversity [9], [10], [24], [25], the resultant attention deviation [9], [26], and the magnitude of gradients during optimization [19], [23], [27]. We incorporated 12 renowned transformation techniques and broadened their hyperparameters to facilitate cross-method and varied intra-method evaluation scenarios. Our findings indicate that the correlation between metrics derived from the established consensuses and the targeted transferability induced by transformations is weak; thus, these metrics do not meet our criteria.

Our analysis was extended to examining AEs' self-transferability across basic transformations, inspired by observations that collaborate distinct data from diverse classes and transformed replicas of a single image. As illustrated in Fig. 1, harnessing both types of data enhances targeted transferability in a similar way. Intriguingly, despite the AEs being generated through various complex transformations, their transferability to black-box models is highly correlated with their self-transferability to certain basic geometric transformations like rotation and scaling. This suggests an overlap in the role of diverse transformations in aiding attacks on invariant features, hinting at the potential simplicity of transformation design, and underscores the need for further research into the utility of basic transformations. To enable blind estimation of the effectiveness of directly attacking various basic transformations, we further devised a measure called surrogate self-alignment. In essence, it simplifies the role of transformations in mitigating overfitted perturbations' utility, thereby exploiting the intrinsic alignment between the surrogate model and various black-box models on the distribution of clean data. The self-alignment evaluations identified the scaling transformation as exceptionally effective

Fig. 2. Comparison against existing transformation methods-based simple TTAs at incremental attack iterations. **Observations:** Compared with the previous SOTA transformation, H-Aug [26], S^{4}ST yields an absolute improvement of 14.2% on average tSuc to 83.0%, and a relative reduction of 74.3% on time consumption to 1.45s. Additionally, compared with the previous setting ($T = 300$), more attack iterations benefit a lot for potent transformations like our S^{4}ST. **Settings:** The average tSuc is averaged over 14 black-box models and 1,000 images with distinct target labels with the ResNet-50 as the surrogate model. The perturbation budget is $\ell_\infty = 16/255$, and the baseline attack is TMI. Time consumption here refers to the execution duration to attack a single image. Extended results comparing the proposed S^{4}ST and other transformations are in Section 5.3.

in facilitating targeted transferability. Subsequent experiments confirmed that the simple scaling transformation alone can achieve performance comparable to that of many advanced methods, confirming the validity of our self-transferability and self-alignment analyses.

All the findings above enabled the development of a novel, straightforward, yet powerful scaling-centered transformation method for simple TTAs, named Strong, Self-transferable, faSt, and Simple Scale Transformation (S^{4}ST). S^{4}ST amalgamates three efficient strategies to enhance the simple scaling transformation: *1)* Modifications that maintain dimensional consistency while injecting greater diversity and reducing computational burden; *2)* Integration with complementary transformations that are less correlated with scaling (determined according to the self-transferability analysis in a black-box manner); and *3)* Implementation of block-wise scaling. The proposed S^{4}ST transformation underwent extensive evaluation and was demonstrated to be both efficient and effective in promoting targeted transferability, surpassing state-of-the-art (SOTA) methods in both resource-intensive and simple solutions. The results are outlined in Figs. 1 and 2. Moreover, it exhibited exceptional transferability to the real-world commercial API. Additionally, our method was found to significantly improve the exploitation of visual elements within individual images to produce highly transferable target semantics, circumventing the

necessity to learn this capability from extensive multi-class data.

In summary, our main contributions are as follows:

- We revisit various relevant transformation techniques and the prevailing consensus on designing them for transfer attacks. Our experimental results elucidate the critical role of image transformations in simple TTAs and underscore the current consensus's limitations in understanding these transformations' effectiveness.
- We introduce effective black-box measures and provide valuable insights. Self-transferability against basic geometric transformations is verified as a reliable indicator of black-box transferability. We further propose surrogate self-alignment to blindly estimate the benefits of attacking basic transformations in enhancing targeted transferability. These two allow for feasible analysis of basic transformations under the black-box setting, significantly diminishing the dependence on empirical or intuitive choices common in current research.
- To the best of our knowledge, we are the pioneers in demonstrating the uniquely superior efficacy of simple scaling transformations in promoting targeted transferability. An advanced scaling-centric transformation, S^4ST, is further devised, ingeniously integrating modified scaling with complementary transformations and leveraging the benefits of block-wise operations.
- Through extensive and comprehensive experiments on the ImageNet-Compatible dataset, we substantiate our proposed method's effectiveness and efficiency, outperforming existing resource-intensive and simple solutions by considerable margins, thereby establishing a new SOTA for TTAs.

In concert with numerous studies in adversarial machine learning, this research strives to enhance our comprehension of adversarial vulnerabilities to foster the development of trustworthy artificial intelligence systems [28]. The concept of targeted transferability is instrumental not only for bolstering privacy and intellectual property protection [29], [30] and robust inference [2] but also for pinpointing potential risks in critical applications like autonomous driving [31], [32] and identity verification [33], [34]. Although the possibility of malicious exploitation of our findings cannot be eliminated, we are optimistic that this work will serve as a clarion call to stakeholders and the research community, spurring the creation of more robust defenses and ultimately leading to beneficial outcomes.

The rest of this paper is organized as follows. Section 2 introduces the problem definition and related works. Section 3 provides experimental validations of existing methods and consensus. Section 4 elaborates on the motivation, analysis, findings, and details of our method. Section 5 offers the experimental settings, results, and analysis. Finally, Section 6 concludes this work.

2 Background and Preliminary

2.1 Problem Setting

Consider an image classifier F that predicts a label among K categories, $\boldsymbol{y} = \{y_i | y_i \in \mathbb{Z}, 0 \leq i < K\}$, for an input with RGB channels, $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$, where H and W denote the image height and width, respectively. Without loss of generality, we denote $F = C \circ f$, where $f(\boldsymbol{x}) = \boldsymbol{p} \in [0,1]^K$ represents the prediction distribution and $C(\boldsymbol{p}) = y_{\text{argmax}(\boldsymbol{p})}$ indexes the label with the highest confidence. TTA aims to manipulate black-box classifiers, $G = C \circ g$, to produce an arbitrary targeted prediction, y_t, for any input:

$$G(\boldsymbol{x}^{adv}) = y_t, \quad \text{s.t.} \quad \|\boldsymbol{x}^{adv} - \boldsymbol{x}\|_p \leq \epsilon, \tag{1}$$

where $\boldsymbol{x}^{adv}$ is the AE crafted to be within a p-norm ball centered at $\boldsymbol{x}$ with radius ϵ (the perturbation budget). In the context of transfer attacks, $\boldsymbol{x}^{adv}$ is generated using the surrogate model f:

$$\boldsymbol{x}^{adv} = \underset{\boldsymbol{x}' : \|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \epsilon}{\arg\max} \mathcal{L}\left(f(\boldsymbol{x}'), y_t\right), \tag{2}$$

where $\mathcal{L}$ is the loss function designed to induce the targeted misclassification.

2.2 Transferable Targeted Attacks

We adopt the categorization from Zhao *et al.* [19], which divides TTAs into simple and resource-intensive solutions. A schematic overview of these methods is presented in Fig. 1, with detailed descriptions provided below.

2.2.1 Simple TTAs

Simple TTAs employ gradient-based iterative optimization on a target (sample, label) pair using a surrogate model. The methodology for untargeted transferable attacks can be adapted for targeted scenarios. For instance, I-FGSM [35], which is indicative of this approach, addresses (2) through multi-step (approximately 10) gradient descent under the ℓ_∞ constraint,, employing the cross-entropy loss function $\mathcal{L}_{\text{CE}}$:

$$\boldsymbol{x}_0^{adv} = \boldsymbol{x}, \boldsymbol{x}_{i+1}^{adv} = \Pi_\epsilon(\boldsymbol{x}_i^{adv} - \alpha \cdot \text{sign}(\nabla_{\boldsymbol{x}} \mathcal{L}_{\text{CE}}(f(\boldsymbol{x}_i^{adv}), y_t))), \tag{3}$$

where α represents the step size and Π_ϵ denotes the projection operation. While effective in untargeted attacks, the cross-entropy loss faces challenges in TTAs due to gradient vanishing issues [19], [21]. Let $\boldsymbol{\phi}$ denote the representation produced by f, such as features preceding the final linear layer. The logit values are computed as $l_i = w_i^\top \boldsymbol{\phi} + b_i$, with w_i and b_i being the class-specific weight and bias, respectively. The normalized probability with respect to each category is given by the SoftMax function: $p_i = \frac{e^{l_i}}{\sum e^{l_j}}$. The gradient

$$\frac{\partial \mathcal{L}_{\text{CE}}}{\partial \boldsymbol{x}} = \frac{\partial \mathcal{L}_{\text{CE}}}{\partial \boldsymbol{\phi}} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{x}} = [(p_t - 1)\, w_t^\top + \sum_{i \neq t} p_i w_i^\top] \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{x}} \tag{4}$$

tends to vanish as the model becomes confident ($p_{i\neq t} \to 0$). To address this, Li *et al.* introduced a Poincaré distance-based triplet loss, $\mathcal{L}_{\text{Po+Trip}}$, to facilitate effective optimization [21]. The Logit loss, proposed by Zhao *et al.* [19], $\mathcal{L}_{\text{Logit}}$, directly maximizes the target logit and provides a direct solution to the gradient vanishing dilemma. Subsequent studies have explored ways to soften $\mathcal{L}_{\text{CE}}$ to stabilize the optimization process, such as through margin-based logit calibration [22], $\mathcal{L}_{\text{CE}}^{\text{MLC}}$, or an adversarial optimization scheme [23], among others.

Beyond the design of loss functions, it is the image transformations, as elaborated upon below and in Sec. 2.3, that have proven crucial for achieving effective simple TTAs and have so far been underexplored. Zhao *et al.* [19] demonstrated that the Diverse Inputs (DI) transformation [10] significantly enhances the transferability of targeted attacks, noting that AEs rarely transfer successfully without it. The optimization process was observed to require substantially more iterations (approximately 300) to converge. Since then, the DI method has become a standard benchmark in subsequent research. However, efforts to understand which transformations are essential for targeted

transferability remain scarce. Wei *et al*. [26] provided empirical evidence suggesting that DI-transformed images diversify the discrimination regions of the surrogate model across iterations, thereby expanding the final discrimination region with respect to the targeted label in black-box models. They also introduced the T/H-Aug method, which applies random combinations of attention-deviation transformations. Byun *et al*. proposed the Object-based DI (ODI) [34] transformation, inspired by the intuition that AEs transferable across various 3D shapes and viewpoints could more effectively compromise black-box models. In our work, we explore in greater depth the role of image transformations within simple TTAs. We also challenge the limitations of existing empirical insights and intuitive approaches, proposing new methods to analyze various transformations. Building on these, we have developed the S^4ST transformation, which has led to notable performance improvements.

2.2.2 Resource-Intensive TTA

Resource-intensive TTAs are diverse in nature. They seek to improve targeted transferability by leveraging the data distribution through a substantial volume of additional data. Targeted universal attacks [14] develop a single perturbation across a large dataset, enabling the universal attack of multiple unseen samples and demonstrating cross-model transferability. Inkawhich *et al*. chose to train auxiliary models at intermediate layers to capture and attack the feature distribution relative to the target class [36]. At present, generative approaches are among the most potent TTAs. The pioneering generative framework for TTA, known as Transferable Targeted Perturbation (TTP) [15], involves training a generator on extensive data to align with the target class in terms of both global distribution and local neighborhood structure. Wang *et al*. [17] integrated TTP with the techniques from [36] and introduced a perturbation dropping strategy to enhance performance further. Zhao *et al*. introduced the Minimizing Maximum Model Discrepancy (M3D) attack [18], which simultaneously discriminates and targets two surrogate models (the original and its duplicate), achieving the SOTA performance. Although these methods are effective and efficient in attacking, their practicality is constrained by the requirement for extensive training data from the same domain or even the same dataset as the target models and the laborious retraining process for each label. By building on our proposed S^4ST transformation, simple TTAs have surprisingly surpassed SOTA resource-intensive techniques, all while preserving remarkable flexibility and efficiency. This establishes our method as a valuable benchmark for future research.

2.3 Image Transformation for Transferability

Image transformation is an effective and necessary strategy in training deep vision models, effectively mitigating the need for extremely large quantities of data and annotations [37]–[39]. For example, horizontal flips and random resized cropping are commonly used in natural image understanding tasks, significantly boosting generalization performance. Furthermore, transformation strategies are typically data-dependent and can be conveniently implemented in a plug-and-play manner across different model architectures. The same benefits extend to the domain of adversarial attacks. There are two primary streams of transformation strategies for enhancing adversarial transferability: *1)* Dynamically searching for optimal combinations from a vast pool of transformations during the attack iteration [40]–[42]; and *2)* Consistently applying a predefined transformation in each iteration. In this paper, we focus exclusively on the latter to avoid the need to access black-box models for optimization. In addition to these TTA-oriented transformation methods [26], [34], we also examine several prominent works in the field of untargeted attacks. This is to validate whether the prevalent consensuses in transformation design is applicable in the context of TTA. An overview of the methods discussed is presented in Fig. 3, introduced below.

Fig. 3. Illustration of the original image and its transformed versions.

Xie *et al*. first introduced the DI method [10], which involves random resizing and padding, and has shown a significant boost in transferability. The Resized DI (RDI) approach [43] rescales the DI-transformed image back to the original size, creating a reduced-scale version. The Scale-Invariant (SI) technique [25] computes the average gradient with respect to images scaled linearly by powers of two, repeated five times. Admix [24] blends the original image with those from other classes. The Spectrum Simulation Attack (SSA) [44] introduces noise within the frequency domain. The Structure Invariant Attack (SIA) [11] and Block Shuffle & Rotation (BSR) [12] employ random transformations and shuffling & rotation operations on image blocks, respectively. The Deformation-Constrained Warping (DeCoWa) method [13] applies elastic deformation coupled with an internal min-max optimization step.

3 A Closer Look at Transformations in TTA

In this section, we delve into the impact of image transformations on targeted transferability. We also conduct experimental evaluations to determine if the existing consensuses can serve as universal metrics to understand the role of transformations and to guide method design.

3.1 Analysis

We begin with a comprehensive evaluation to assess the current state of simple TTAs. This evaluation encompasses 30 different combinations involving six transformation scenarios—{No transformation, DI [10], ODI [34], SIA [11], BSR [12], and T-Aug [26]} and five loss functions—{$\mathcal{L}_{\text{CE}}$, $\mathcal{L}_{\text{CE}}^{\text{MLC}}$ [22], $\mathcal{L}_{\text{Logit}}$ [19], $\mathcal{L}_{\text{Po+Trip}}$ [21], and $\mathcal{L}_{\text{Logit}}$ with Self-Universality (SU) strategy [27]}. In line with the established protocol [19], we employ the TMI attack as the baseline:

$$\begin{aligned} \boldsymbol{x}_0^{adv} &= \boldsymbol{x}, \quad \boldsymbol{g}_0 = \boldsymbol{0}, \\ \boldsymbol{g}_{i+1} &= \mu \cdot \boldsymbol{g}_i + \frac{\boldsymbol{W} * \nabla_{\boldsymbol{x}} \mathcal{L}(f(\mathcal{T}(\boldsymbol{x}_i^{adv}), y_t))}{\left\| \boldsymbol{W} * \nabla_{\boldsymbol{x}} \mathcal{L}(f(\mathcal{T}(\boldsymbol{x}_i^{adv}), y_t)) \right\|_1}, \\ \boldsymbol{x}_{i+1}^{adv} &= \Pi_\epsilon(\boldsymbol{x}_i^{adv} + \alpha \cdot \text{sign}(\boldsymbol{g}_{i+1})). \end{aligned} \tag{5}$$

Here, μ denotes the momentum decay factor [8], and $\boldsymbol{W}$ represents a pre-defined kernel for gradient smoothing [9]. For our study, we extend the number of attack iterations to $T = 900$, in keeping with the increased diversity afforded by the most

Fig. 4. Targeted transferability resulted by various transformation and loss function combinations. The quality of transformation-induced features is identified as more pivotal than softening prediction distribution in enabling targeted transferability.

advanced methods, surpassing the DI method [19]. Targeted AEs are crafted using a ResNet-50 surrogate model [45], and ***black-box transferability*** is evaluated by the variance in confidence scores of black-box models with respect to the target class y_t between the adversarial (generated by $\mathcal{T}$-TMI with $\mathcal{L}$) and benign samples:

$$T_{\mathcal{T},\mathcal{L}} = \mathbb{E}_{\boldsymbol{x}\in\mathcal{X},g\in\mathcal{G}}\left(g(\boldsymbol{x}^{adv}_{\mathcal{T},\mathcal{L}},t) - g(\boldsymbol{x},t)\right), \tag{6}$$

where $g(\boldsymbol{x},t)$ is the confidence score to the target class, and the empirical expectation is computed over 14 black-box models and 1,000 samples, each with a unique target label, to ensure statistical significance. Further experimental details are provided in Section 5.1.

The results are presented in Fig. 4, where improvements over the baseline methods are only noticeable when transformations are applied. These advanced loss functions—whether they work by softening the prediction distribution [22], [23] or by employing alternative objectives [19], [21], [27]—seem to be tailored to leverage the features diversified by transformations, yet they do not address why these transformations are so crucial to targeted transferability. Additionally, with certain transformations like ODI, even the standard cross-entropy loss can achieve performance on par with more complex loss functions. This further highlights the intrinsic importance of transformation techniques in enhancing feature diversity over merely softening prediction distributions.

Based on the insights gained, we wish to underscore the innate ability of transformations to mitigate the issue of gradient vanishing in simple TTAs. In essence, when transformations are applied to perturbations, they can reduce their adversarial effectiveness [46], [47], which has been recognized as a key defensive strategy. Coupled with the gradient vanishing phenomenon described by Eq. (4), the advantages of transformations can be seen as a *dual* facilitation: *In each iteration, they introduce diversified features (ϕ in Eq. (4)) while simultaneously reducing the adversarial utility (i.e., , diminishing p_t), thus significantly enhancing gradient utilization.* To date, researchers have primarily concentrated on the *outcome*, *i.e.*, , the variation in $\boldsymbol{p}$, to aid min-max optimization processes [42], [48], [49]. However, in the following section, we will demonstrate the inadequacy of focusing solely on the prediction distribution. This realization prompts us to investigate the *cause*. Specifically, we aim to determine which transformations can effectively induce and support the attack on features that are more inherent, invariant, and transferable within images—a question that has been largely neglected in prior studies. Previous research has often relied on empirical intuition or general agreement when devising transformation methods, without robust analytical frameworks.

3.2 Failures of Existing Consensuses

In this subsection, we scrutinize the prevailing consensuses commonly referenced in the development of transformation methods for transfer attacks. Our objective is to ascertain their robustness as metrics reflecting targeted transferability and to determine their utility in guiding the design of transformations for TTAs. To achieve this, we assess the black-box transferability of various transformation techniques under both default and expanded hyperparameter configurations. We then examine the presence of a positive (or linear) correlation between their transferability and the quantitative measures derived from the established consensus, both across different methods and within individual methods. The consensuses under investigation and their corresponding measures are introduced as follows:

- **Diversity**. Initially, researchers emphasized the need for transformations that preserve loss [9], [10], [24], [25], ensuring that the surrogate model's output remains relatively constant post-transformation. The intensity of transformations was thus limited to a narrow scope to maintain this characteristic. However, more recent advancements have shifted towards utilizing gradient averaging over multiple transformed samples (20×) to achieve stable gradient updates [11]–[13], [44]. This shift implies that increased diversity might lead to improved transferability. We quantify ***diversity*** by calculating the average cross-entropy loss values of the surrogate model across 50,000 images from the ImageNet-1k validation set [50].
- **Attention Deviation**. Dong *et al*. first pointed out that discrepancies in discriminative regions between models and defensive strategies can hinder adversarial transferability [9]. They advocated for the use of translation transformations to extend the reach of adversarial perturbations to wider attention regions. Following this, other studies have also utilized attention as a universal objective for crafting transferable attacks [51], [52]. More recently, BSR [12] and H-Aug [26] have set the new SOTAs for untargeted and targeted attacks by disrupting attention through image block shuffling and random attention-deviation transformations, respectively. We adopt the ***attention deviation*** metric proposed by Wei *et al*. [26], which computes the Intersection over Union (IoU) of Grad-CAM [53] activations concerning the target label between original and transformed images. The results are averaged over 1,000 images with distinct target labels.
- **Gradient Magnitude**. A significant portion of research on simple TTAs has been devoted to addressing the issue of gradient vanishing, particularly in studies focusing on loss functions [19], [22], [23], [27]. The magnitude of gradients has often been used as an indicator to gauge the efficacy of different approaches. It is hypothesized that larger gradient magnitudes could lead to more efficient updates of perturbations, aligning with our analysis of the *outcome* from the dual effects of transformations. To test this hypothesis, we examine the gradient magnitudes produced by various transformations under an identical loss function. Specifically, the ***gradient magnitude*** is measured by the mean ℓ_2-norm of gradients over 900 $\mathcal{T}$-TMI iterations and is averaged across 1,000 images with distinct target labels.

All attacks are executed using the $\mathcal{L}_{\text{CE}}^{\text{MLC}}$ loss function,

Fig. 5. Scatter diagrams depicting relationships between ***black-box transferability*** and (a) ***diversity***, (b) ***attention deviation***, and (c) ***gradient magnitude***. As indicated by the Pearson Correlation Coefficients (PCCs) in the top left corner of each figure, all three metrics exhibit weak correlation to transferability in the cross-method case, impeding their application in identifying what transformations are effective and designing one from scratch.

which has demonstrated effectiveness in our evaluations. The results are presented in Fig. 5. Within the intra-method analyses—where the relationship between different hyperparameter configurations of individual method is evaluated—a positive or linear relationship may be discernible. Notably, the expanded parameters for the DI method expose the constraints of experimental setups in previous research.* Nevertheless, this only suggests that these metrics can provide empirical guidance for parameter scaling within specific transformations. Importantly, there is a significant confusion of these metrics across different methods, which complicates the task of developing new transformation methods from the ground up.

4 The Proposed Method

This section delineates the proposed method in a systematic manner. It integrates the motivation derived from previous analyses, delineates two plausible strategies for gauging the impact of transformation methods on targeted transferability, and culminates in the introduction of a powerful yet efficient transformation technique to augment simple TTAs.

4.1 Motivation

In our research, we fervently promote the investigation into robust blind estimators for transformation-induced targeted transferability. To this end, we aspire to establish design principles for formulating transformation techniques. Our inquiry specifically targets questions such as *which transformations prove most effective* and *how to determine the synergistic effects of combining various transformations*. We anticipate that our findings will identify transformations that facilitate highly efficient simple TTAs. Put succinctly, our goal is to surpass the efficacy of more resource-intensive attacks while preserving efficiency and the ability to handle arbitrary (sample, label) pairs. This is accomplished without resorting to additional data or retraining models, thereby striving for an optimal balance between performance and practicality. Moreover, the estimators in question should conform to the prerequisites of black-box transfer attacks, thereby ensuring their applicability to a broader spectrum of data beyond natural imagery. To begin, we focus on basic transformations and subsequently investigate their potential synergistic enhancements. The specifics of the basic transformations evaluated and their respective intensities $\mathcal{S}$ are meticulously defined and expounded upon in Appendix A.

Reflect on the trends elucidated in Figs. 4 and 5, which indicate that employing more powerful transformations or an increased scope of transformed images (by amplifying the hyperparameters of the given method) leads to stronger targeted transferability. We juxtapose these findings with those of resource-intensive attacks, recognizing that both are predominantly influenced by data diversity. To illustrate this, we replicate the SOTA generative method, M3D [18], with varying quantities of training data. The effects of training data volume on resource-intensive attacks (as shown in Fig. 1a) bear a striking parallel to the influence of transformation strength (in both cross-method and intra-method scenarios) on simple attacks—both a paucity of transformation and insufficient training samples culminate in negligible transferability. Conversely, more robust transformations and a greater variety of transformed images, as well as an increased volume of training data, lead to marked enhancements. These insights propel us to initiate our exploration with an examination of AEs' ***self-transferability*** in relation to basic transformations.

4.2 Mirroring Black-box Transferability via Self-Transferability

To elucidate, we conjecture that various transformed versions of a single image may serve as surrogates for multi-class data, thereby aiding the optimization of multi-to-single perturbations that better capture the invariant features. In certain studies, this approach has been construed as integration of transformation layers with the surrogate model, effectively emulating a range of black-box models [41], [54]. To substantiate our hypothesis, we define ***self-transferability*** as the mean efficacy of targeted adversarial perturbations across varying transformation intensities:

$$ST_{\widehat{\mathcal{T}}} = \mathbb{E}_{\boldsymbol{x}\in\mathcal{X}, s\in\mathcal{S}} \left(f(\widehat{\mathcal{T}}_s(\boldsymbol{x}_{\mathcal{T}}^{adv}), t) - f(\widehat{\mathcal{T}}_s(\boldsymbol{x}), t) \right), \tag{7}$$

where s represents the intensity level of a basic transformation $\widehat{\mathcal{T}}$. Hereafter, we omit the loss function from the discussion, as the $\mathcal{L}_{\text{CE}}^{\text{MLC}}$ loss [22] is adopted by default. We generate AEs employing extant transformations with the TMI attack and conduct a PCC test to compare their black-box transferability, *i.e.*, , Eq. (6), with self-transferability, *i.e.*, , Eq. (7). It is important to note that methods with high correlation are deliberately excluded in the

Fig. 6. Scatter diagrams depicting relationships between ***black-box transferability*** and ***self-transferability*** to (a) rotation and (b) scaling transformations.

Fig. 7. PCCs calculated between the ***black-box transferability*** against 14 black-box models and the ***self-transferability*** against various basic transformations, based on AEs obtained from 12 existing attacks.

cross-method analysis. For instance, the (R)DI and H-Aug are not considered when examining the scaling transformation, and both BSR and H-Aug are omitted for rotation.

In Fig. 6, we present scatter plots for rotation and scaling transformations, revealing notably tighter correlations in the cross-method scenario compared to previous metrics. The self-transferability of AEs against these basic transformations is an effective reflection of the trends observed in their black-box transferability, applicable to both cross-method and intra-method contexts. An exception is observed where intensifying the DI transformation initially leads to an increase in self-transferability for rotation, which then diminishes—this represents the sole outlier. The PCCs for all scenarios are detailed in Fig. 7, from which we can discern the following:

1) The self-transferability of AEs to most geometric transformations demonstrates a consistently high-quality linear correlation with their black-box transferability. For example, attacking features at varying frequency-domain distortions (which often induce color shifts) [44], elastic deformations [13], and random block-wise transformations [11]—despite their seemingly disparate natures—uniformly impact the exploitation of features across scales. This correlation persists even with more complex transformations and a range of geometric ones, such as rotation, shear, and translation. This not only highlights the potency of geometric transformations in eliciting transferable features but also emphasizes their intrinsic similarities. Given these similarities, it is more strategic to target a few highly effective basic transformations to optimize overall performance, rather than indiscriminately combining overlapping operations. This calls for a deeper exploration into which specific transformations, when leveraged by TTAs, contribute most significantly to black-box transferability.
2) The self-transferability to color transformations shows a weak correlation with black-box transferability in cross-method comparisons, but a robust correlation in the intra-method cases of SSA and SIA. This can be ascribed to the fact that both transformations pertain to the manipulation of image color. This finding suggests the importance of analysis to pinpoint complementary color transformations, as their potential has often been underestimated in existing methods.

The insights delineated above establish a foundation for our method design, which is predicated on focusing on one or more particularly impactful basic transformations and then pursuing additional synergistic enhancements.

Fig. 8. Alignment between the surrogate model and black-box models, which reflects the black-box transferability resulted by transformations. The vanilla gradient update, *i.e.*, TMI, quickly misaligns the surrogate from black-box models and leads to poor transferability.

4.3 Assessing Effectiveness via Surrogate Self-Alignment

In this section, we probe deeper into comparing the effectiveness of attacking basic transformations on the black-box transferability of AEs. This investigation is challenging due to the prohibition of direct evaluations on black-box models. To circumvent this, we suggest an approach rooted in the concept of model alignment. The underlying principle is straightforward: models trained on identical distributions exhibit a degree of inherent alignment. AEs that capitalize on this alignment between the surrogate and black-box models are more likely to transfer. However, the persisting discrepancies [55] between models, along with the overfitting of gradient-based optimization, sharply misalign the surrogate model with the black-box model on the distribution of AEs. In Fig. 8, we illustrate the average alignment metric between the surrogate and black-box models (defined subsequently in Eq. (8)) throughout the attack iterations. It becomes apparent that direct utilization of the surrogate model's gradients can lead to substantial misalignment with the black-box model, culminating in poor transferability. Although transformations can mitigate this misalignment, it prompts the question: *how can one blindly assess the efficacy of these basic transformations in facilitating this mitigation?*

Our solution leverages what we term ***surrogate self-alignment*** induced by basic transformations. To simplify our analysis, we first examine the similarity between representations of individual samples. An illustration is depicted in Fig. 9. Given the inherent alignment between models f and g, we can reasonably postulate that the cosine similarity of their representations, $\text{CosSim}(\boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\psi}(\boldsymbol{x}))$, is significantly greater than zero. More informally, we may suggest that $\boldsymbol{\phi}(\boldsymbol{x}) \approx \boldsymbol{\psi}(\boldsymbol{x})$.

Fig. 9. The colored numbers indicate the cosine similarity between representations generated by the surrogate model (ResNet-50) and the black-box model (ConvNeXt), where the AE $\boldsymbol{x}^{adv}$ is created based on the TMI algorithm. On the clean sample, these two models exhibit a high degree of alignment. However, AEs with weak transferability have limited impact on the black-box model, which suggests that we can introduce an image transformation $\mathcal{T}$ to make $\boldsymbol{\phi}(\mathcal{T}(\boldsymbol{x}^{adv}))$ more closely approximate $\boldsymbol{\phi}(\boldsymbol{x})$ (***self-alignment***), thereby enhancing the approximation to $\boldsymbol{\psi}(\boldsymbol{x}^{adv})$ (***alignment***) and thereby improving transferability.

For an AE $\boldsymbol{x}^{adv}$, crafted using a weakly-transferable algorithm like TMI that has a limited impact on the black-box model g, we hypothesize that $\boldsymbol{\psi}(\boldsymbol{x}^{adv}) \approx \boldsymbol{\psi}(\boldsymbol{x}) \approx \boldsymbol{\phi}(\boldsymbol{x})$, with all three representations being markedly different from $\boldsymbol{\phi}(\boldsymbol{x}^{adv})$. This difference becomes more pronounced with each iteration of the attack. Image transformations can enhance attack effectiveness by harnessing the mutual alignment between f and g, with the aim of making $\boldsymbol{\psi}(\boldsymbol{x}^{adv})$ closer to $\boldsymbol{\phi}(\mathcal{T}(\boldsymbol{x}^{adv}))$. In line with the principles of black-box attacks, we suggest an extension of this strategy such that $\boldsymbol{\phi}(\mathcal{T}(\boldsymbol{x}^{adv})) \approx \boldsymbol{\phi}(\boldsymbol{x})$ (self-alignment).

We then expand the concept of similarity between individual representations to encompass the mutual alignment across the entire dataset and multiple black-box models. Let $\Phi = \{\boldsymbol{\phi}(\boldsymbol{x}_i)\}_{i=1}^b$, $\Phi_{\widehat{\mathcal{T}}_s} = \{\boldsymbol{\phi}(\widehat{\mathcal{T}}_s(\boldsymbol{x}_i))\}_{i=1}^b$, and $\Psi = \{\boldsymbol{\psi}(\boldsymbol{x}_i)\}_{i=1}^b$ denote the sets of representations produced by models f, $f \circ \widehat{\mathcal{T}}_s$, and g for a set of samples $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^b$, respectively. The superscript $(\cdot)^{adv}$ indicates the corresponding representations for the adversarial counterparts of $\mathcal{X}$, which are crafted using the TMI algorithm and model f. To quantify the mutual alignment between two sets of representations, such as Φ and Ψ, we measure the overlap of their respective k-nearest neighbors [56] throughout the dataset. This is formalized as:

$$\text{Alignment}\,(\Phi, \Psi) = \mathbb{E}_{g \in \mathcal{G}, (\boldsymbol{\phi}, \boldsymbol{\psi}) \in (\Phi, \Psi)} \frac{1}{k} \left| d_{\text{knn}}(\boldsymbol{\phi}, \Phi \setminus \boldsymbol{\phi}) \cap d_{\text{knn}}(\boldsymbol{\psi}, \Psi \setminus \boldsymbol{\psi}) \right|, \tag{8}$$

where d_{knn} returns the set of indices corresponding to its k-nearest neighbors, with k set to 100 for a dataset comprising $b = 1,000$ images.

Table 1 presents the alignment metrics between models for various combinations. The results are consistent with those illustrated in Fig. 9, demonstrating that the alignment scores for $\text{Alignment}(\Phi, \Psi)$ and $\text{Alignment}(\Psi, \Psi^{adv})$ are not only comparable but also significantly exceed those of the other two scenarios. This substantiates the efficacy of our approach, leveraging the self-alignment, $\text{Alignment}(\Phi^{adv}_{\widehat{\mathcal{T}}_s}, \Phi)$, as a proxy for desired alignment, $\text{Alignment}(\Phi^{adv}_{\widehat{\mathcal{T}}_s}, \Psi^{adv})$, to blindly gauge the impact of different transformations on the transferability of attacks. Fig. 10 further delineates the desired alignment and self-alignment metrics for the basic transformations under study at varying intensities. These metrics closely mirror each other, reflecting both the individual transformation trends at distinct intensities and the comparative magnitudes across transformations, thereby corroborating our analysis. *Notably, scaling not only yields a significantly enhanced surrogate self-alignment compared to other geometric transformations but also demonstrates a wider effective intensity range, highlighting its potential as a key transformation for our strategy.* To corroborate these preliminary assessments and findings, we conduct comparative experiments among these basic transformations in Section 5.2.1.

TABLE 1
Mutual alignment between surrogate model and multiple black-box models.

Case	Alignment
$\text{Alignment}(\Phi, \Psi)$	0.42
$\text{Alignment}(\Psi, \Psi^{adv})$	0.32
$\text{Alignment}(\Psi^{adv}, \Phi^{adv})$	0.12
$\text{Alignment}(\Phi^{adv}, \Phi)$	0.10

(a) *Alignment* (b) *Self-Alignment*

Fig. 10. The metrics of ***alignment*** and ***self-alignment*** exhibit a high degree of similarity. Notably, the scaling transformation demonstrates distinctive effectiveness, simultaneously inducing high alignment strength and a large effective range of transformation intensities.

Fig. 11. The proposed S^{4}ST encompasses the *Base* operation, which is based on scaling, complemented by two orthogonal enhancements—*Aug* and *Block*.

4.4 S^{4}ST

Drawing on the insights and findings from our research, we introduce a novel transformation methodology focused on scaling, named Strong, Self-transferable, Fast, and Simple Scale Transformation (S^{4}ST), designed to enhance simple TTAs. Figure 11 illustrates the schematic overview of S^{4}ST, with detailed explanations to follow.

4.4.1 Overview

The essence of the S^{4}ST framework is founded on the basic scaling transformation, informed by our preceding analysis, and seeks synergistic benefits with complementary basic operations. In alignment with recent advancements, our framework also integrates strategies for image block manipulation. To accommodate block-wise scaling and alleviate the computational demands associated with upscaled inputs, we have refined the scaling transformation to preserve dimensional consistency. As a result, the S^{4}ST framework is segmented into three distinct components: S^{4}ST-*Base*, S^{4}ST-*Aug*, and S^{4}ST-*Block*, each reflecting one of the three primary motivations identified.

4.4.2 S^{4}ST-Base

We define the intensity of the S^{4}ST-*Base* with the variable r, which is distinct from the definition of s in Appendix A. Consider a sample $\boldsymbol{x}$ with dimensions $H \times W$ and an intensity $r > 1$. With a probability p_r, S^{4}ST-*Base* selects a relative scale factor r' from a uniform distribution $\mathcal{U}(\frac{1}{r}, r)$. If $r' < 1$, the sample $\boldsymbol{x}$ is downscaled to dimensions $r'H \times r'W$, followed by random zero-padding to restore its original size. In contrast, if $r' > 1$, a square patch of dimensions $\frac{1}{r'}H \times \frac{1}{r'}W$ is randomly cropped from the image and then upscaled to the initial dimensions.

Scaling transformations are also utilized in existing approaches, such as DI [10] and RDI [43]. The innovations of our work compared to these methods are threefold: *1)* Our black-box condition analysis singles out scaling as the foremost

Fig. 12. PCCs between the self-transferability to scaling and other basic transformations, derived from 11 intensity levels using S^{4}ST-*Base*. This evaluation did not involve any black-box models.

effective basic transformation other than empirical choice; *2)* We ascertain that an extended range of scale transformations can lead to further enhancements; *3)* We introduce a versatile scaling transformation that includes both enlargement and reduction, unlike previous methods that typically focus on one direction. These advancements enable our S^{4}ST-*Base* to stand out for its combined effectiveness and efficiency. A comprehensive comparison among variations of the scaling transformation is presented in Section 5.2.2.

4.4.3 S^{4}ST-Aug

Guided by the insights gained from Fig. 7, we aimed to integrate a suite of complementary basic transformations into the S^{4}ST-*Base* to further boost its efficacy. With a probability p_{aug}, S^{4}ST-*Aug* initially applies a randomly chosen transformation from an equitable selection pool of basic transformations to the entire input image. To compile this pool, we execute S^{4}ST-*Base* across a spectrum of intensities $r \in [1.02, 1.04, 1.06, 1.08, 1.1, 1.2, 1.4, 1.6, 1.8, 2.0, 2.2]$. We then evaluate the linear correlation of self-transferability for this array of AEs over a range of basic transformations, with the results displayed in Fig. 12. We select five basic transformations that demonstrate the lowest PCCs with ST_{Scaling}. Additionally, we omit the translate and crop transformations due to their significant overlap with S^{4}ST-*Base*. Following the parameterization in *torchvision*[1], the transformation pool for S^{4}ST-*Aug* includes the following basic transformations with their empirically determined parameters:

- **Flip** randomly flips the image vertically or horizontally.
- **Brightness/Contrast/Saturation** randomly alters the `brightness/contrast/saturation` factor from 1 to $s' \sim \mathcal{U}(0, 2)$.
- **Hue** randomly adjusts the `hue` factor from 0 to $s' \sim \mathcal{U}(-0.5, 0.5)$.

4.4.4 S^{4}ST-Block

Recent research has underscored the value of block-wise transformations in enhancing adversarial transferability. With S^{4}ST-*Block*, we advance this concept by randomly dividing the image into m blocks and applying S^{4}ST-*Base* to each block, where $m = m_h \times m_w$ indicates partitioning the image into m_h and m_w segments along the coordinate axes (not specifying horizontal or vertical orientation in order).

We acknowledge that the design of the block-wise transformation is heavily inspired by prior works such as SIA [11] and BSR [12]. It is important to reiterate that our design is rooted in the analysis conducted within this study, as opposed to being purely based on intuition or empirical choice, in accordance with the shortcomings of existing paradigms discussed in Section 3.2. This analytical approach yields notable differences between our proposed method and pre-existing ones: *1)* Our method solely concentrates on block-wise scaling, deduced from our self-alignment estimation concerning the efficacy of various basic transformations, unlike SIA's arbitrary transformations for each image block or BSR's shuffling and rotation operations; *2)* The selection of complementary transformations in our approach is guided by correlation analyses of self-transferability under black-box scenarios (as depicted in Fig. 12), rather than the intuitive selections employed by SIA and BSR. Comparisons against these methods are provided in Section 5.3.

5 Experiments

In this section, we detail the experimental setup and conditions. We scrutinize the parameter configurations and the ablation of components for our proposed method, showcasing its novel aspects in contrast to existing techniques. This is followed by extensive experiments designed to compare our method with the SOTA transformation-based simple TTAs and the resource-intensive approaches.

5.1 Experimental Settings

Dataset. Our experiments are conducted on the ImageNet-Compatible dataset[2], a benchmark commonly utilized for evaluating TTA performance [19], [27], [34]. Released for the NIPS 2017 adversarial competition, it encompasses 1,000 images with dimensions of 299×299, each paired with a targeted label. For our evaluated models, we resize the images to 224×224.

Models. In line with prior studies, we predominantly generate AEs using ResNet-50 [45] and transfer them to a variety of dissimilar victim models. These include CNN architectures—{MobileNet-v2 (MNv2) [57], EfficientNet-b0 (EN) [58], ConvNeXt-small (CNX) [59], Inception-v3 (INv3) [60], Inception-v4 (INv4) [61], Inception ResNet-v2 (IRv2), and Xception (Xcep) [62]} and vision transformers—{ViT [63], SwinT [64], MaxViT [65], Twins [66], PiT [67], TNT [68], and DeiT [69]}. We source their pre-trained weights from the *torchvision*[3] and *timm* [70] libraries. Supplemental black-box transfer experiments are performed on defense models and real-world APIs, with further details provided in subsequent sections.

Baseline. The TMI attack is employed as the baseline. The step size and perturbation budget are configured to 2/255 and 16/255, respectively. The momentum accumulation decay factor is set at 1 [8], and a 5×5 Gaussian kernel [9] is used for TI. We adopt $\mathcal{L}_{\text{CE}}^{\text{MLC}}$ [22] with $T = 900$ as the attack criterion.

Competitors. Among simple TTAs, we compare our S^{4}ST transformation with established methods like DI [10], RDI [43], SI [25], Admix [24], SSA [44], ODI [34], SIA [11], BSR [12], T-Aug [26], and H-Aug [26]. For resource-intensive TTAs, we select Dominant Feature Attack (DFA) [14], TTP [15], and M3D [18] as competitors. We adhere to the implementations detailed in their respective publications and the *TransferAttack*[4] code base.

Metrics. We gauge the efficacy of different methods by their average targeted success transfer rate (tSuc), i.e., the proportion of successful transfers across the dataset from the surrogate to the victim models. Additionally, we evaluate the time required by different transformations to craft AEs. All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU using the PyTorch [71] framework. For consistency, the batch size is set to 10 across all experiments, ensuring that no method surpasses

1. https://github.com/pytorch/vision/tree/main/torchvision/transforms

2. https://github.com/cleverhans-lab/cleverhans/tree/master/cleverhans_v3.1.0/examples/nips17_adversarial_competition/dataset

3. https://github.com/pytorch/vision/tree/master/torchvision/models

4. https://github.com/Trustworthy-AI-Group/TransferAttack

the memory threshold, thereby maintaining fairness in the comparison of time consumption metrics.

5.2 Analysis on S^{4}ST

5.2.1 Scaling versus Other Basic Transformations

Our evaluation commences with the selection rationale for the scaling transformation, evaluating the judiciousness of transformation choices through the lens of the proposed self-alignment evaluation. To this end, we generate adversarial samples employing each basic transformation under consideration. With a probability of 0.9, given a transformation strength s, each transformation is executed with an intensity $s' \in \mathcal{U}(0, s)$, selecting a random variant for the surrogate model to process and compute gradients. The outcomes illustrated in Fig. 13 reveal that scaling transformation stands out, achieving an unparalleled average tSuc of 58.0% at $s = 0.8$. This performance is nearly 10% superior to that of the shear transformation (48.8%) and the perspective transformation (48.4%). Notably, the shear transformation shows substantial effectiveness at lower intensities; however, its efficiency markedly wanes at higher intensities. Furthermore, it is observed that employing color transformations in isolation offers marginal benefits for the attack. Yet, subsequent experimentation will demonstrate that when synergized as auxiliary augmentations to scaling, color transformations can significantly bolster targeted transferability.

Fig. 13. Results by applying various basic transformations with incremental intensities. The scaling transformation exhibits a distinctive superior enhancement in targeted transferability.

5.2.2 Comparison among Scaling's Variations

We perform comparative evaluations of the (R)DI methods under both unidirectional (*i.e.*, limited to either enlargement or reduction) and bidirectional conditions (*i.e.*, integrating both enlargement and reduction), with the findings presented in Fig. 14. Specifically, under the hyperparameter configurations from prior TTA research, the DI and RDI methods attain an average tSuc of only 22.4% and 27.3%, respectively. Our research reveals that minor adjustments to their parameters can elevate these figures to 36.6% and 41.7%. By synergistically applying these methods in both the upscale and downscale directions, we obtain a significantly enhanced average tSuc of 56.8% at $r = 2.4$, which, while slightly lower than the 57.2% achieved by the simple scaling transformation, is substantially surpassed by our S^{4}ST-*Base* approach, which further boosts the outcome to 61.6%. This highlights the superiority of our method in both unlocking and augmenting the effectiveness of the basic scaling transformation. Fig. 15 delineates the time consumption of these methods, highlighting our approach's computational efficiency. The incremental time investment required by S^{4}ST-*Base* compared to the baseline method (at $r = 1.0$) is minimal, and it markedly surpasses the competitors.

(a) Unidirectional comparison (b) Bidirectional comparison

Fig. 14. Our analysis uncovers significant untapped potential in existing scaling-related methods. We illustrate that prior achievements can be substantially enhanced by extending the scaling range and incorporating bidirectional scaling. Our tailored refinements to the S^{4}ST-*Base* further amplify this effect.

Fig. 15. In addition to its effectiveness, our S^{4}ST also exhibits superior efficiency. The additional time required by our S^{4}ST is trivial, significantly outstripping the simple scaling and the integration of RDI [43] and DI [10] methods in terms of computational speed.

5.2.3 Effectiveness of S^{4}ST Components

Building upon the S^{4}ST-*Base* framework, we delve into the benefits conferred by the proposed complementary augmentations (S^{4}ST-*Aug*) and block-wise scaling (S^{4}ST-*Block*). We maintain the probability of scaling transformations at $p_r = 0.9$ and progressively increase the number of image blocks m and the probability of applying complementary augmentations p_{Aug}. The outcomes are depicted in Fig. 16. For S^{4}ST-*Block*, an escalation in the number of image blocks correlates with a marked enhancement in attack efficacy, attaining its zenith at $m = 6\,(2 \times 3)$. A further increase in blocks does not yield improved convergence within the configured iterations for the attack. In the case of S^{4}ST-*Aug*, a higher frequency of applying complementary augmentations is linked with a steady amplification in performance. Additionally, the integration of five distinct types of complementary transformations, particularly when paired with scaling (for $r \geq 1.2$), results in substantially improved outcomes compared to their solitary application (at $r = 1.0$). This emphasizes the soundness of our approach, which synergizes a primary transformation with complementary enhancements.

5.2.4 Hyperparameter Setting and Component Ablation

To ascertain the most effective hyperparameter configuration for the given number of attack iterations, we utilize Bayesian opti-

(a) Block-wise scaling (b) Complementary augmentations

Fig. 16. Further enhancements to S^{4}ST-*Base* achieved by (a) S^{4}ST-*Block* and (b) S^{4}ST-*Aug*, both of which demonstrate significant improvements.

mization to concurrently search four parameters: $[p_r, r, p_{aug}, m]$. This is chosen to circumvent the potential suboptimal results that may arise from independently optimizing parameters for each component. The optimization endeavor targets the maximization of the average tSuc on a randomly selected subset of the ImageNet-Compatible dataset (50 images), consisting of 10 initial random explorations and a total of 100 trials. The resulted hyperparameter setting is $p_r = 0.9, r = 1.7, p_{aug} = 1.0, m = 6$. Furthermore, we detail the ablation study outcomes for the components of S^{4}ST with this configuration, as illustrated in Tables 2 and 3. Remarkably, S^{4}ST-*Aug* and S^{4}ST-*Block* exhibit satisfying synergy and collectively contribute to a notable increase in performance, achieving an 83.0% average tSuc. Additionally, the pool for S^{4}ST-*Aug*, which consists of five transformations, demonstrates a degree of complementarity and robustness, with the exclusion of any single transformation having a minimal impact on the overall performance.

TABLE 2
Results by different combinations of the S^{4}ST components.

S^{4}ST-*Base*	S^{4}ST-*Aug*	S^{4}ST-*Block*	$[p_r, r, p_{aug}, m]$	tSuc	Time
✓			$[0.9, 1.7, 0., 1]$	58.2	1.24
✓	✓		$[0.9, 1.7, 1., 1]$	76.8	1.40
✓		✓	$[0.9, 1.7, 0., 6]$	74.3	1.31
✓	✓	✓	$[0.9, 1.7, 1., 6]$	83.0	1.45

TABLE 3
Results by excluding complementary augmentations from S^{4}ST.

S^{4}ST	-Flip	-Contrast	-Brightness	-Saturation	-Hue
83.0	80.8	82.6	82.7	82.7	82.6

5.3 Comparison Against Existing Transformations

This subsection presents comparisons of the proposed S^{4}ST with existing transformation methods across various settings.

5.3.1 Single Surrogate Attacks

Table 4 reports the single surrogate model (ResNet-50) evaluation. Herein the transformed copies per iteration for SSA, SIA, BSR, and DeCoWa are limited to 1. We also set $m_1 = 1, m_2 = 5$ for Admix to ensure a fair comparison with SI. The results demonstrate that the proposed S^{4}ST exhibits superior performance, not only achieving significantly higher average Suc than the previous SOTA in simple TTAs, surpassing H-Aug by 14.2% and ODI by 24.9%, but also leading in terms of efficiency, with computational time only 25.7% of H-Aug and 28.4% of ODI. In fact, as can be seen in conjunction with Fig. 14, even the simple scaling transformation and our S^{4}ST-*Base* are sufficient to achieve transferability on par with all methods except for T/H-Aug. And the further enhancements through block-wise scaling and complementary augmentations synergistically contribute to the final superior performance. This evaluation is extended to alternative surrogate models (without further tuning the parameters), with results reported in Table 5, and various attack iterations in Fig. 2. Our method consistently outperforms the others by a significant margin.

TABLE 5
Results by alternative surrogate models.

Surrogate	DenseNet-121 [72]		VGG16-bn [73]		ViT [63]	
Attack	tSuc	Time	tSuc	Time	tSuc	Time
TMI [8], [9]	1.1	1.86	0.8	1.66	0.0	3.99
+ODI [34]	46.2	5.57	38.3	5.26	31.1	7.70
+SIA [11]	49.9	2.39	33.7	2.30	22.1	4.63
+DeCoWA [13]	43.6	3.89	46.5	3.66	14.6	8.30
+BSR [12]	47.9	2.03	32.1	1.95	27.6	4.27
+T-Aug [26]	49.3	2.01	48.1	1.75	27.2	4.06
+H-Aug [26]	59.3	4.51	48.3	7.00	-	-
+S^{4}ST (Ours)	**74.6**	2.15	**65.6**	1.83	**64.8**	4.05

The best results are in **bold**.

5.3.2 Ensemble-based Attacks and Against Robust Victims

We further extend our comparison to scenarios involving the generation of AEs based on an ensemble of surrogate models and evaluations using robust victim models. For the selection of ensemble surrogate models, we distinguish between the commonly used self-ensemble in untargeted transfer attacks [11]–[13], [44], where the average gradient over multiple transformed copies is computed in a single iteration, and cross-model ensemble [8], which involves calculating the loss function using the average logits from different models to obtain gradients. In line with the evaluation protocols of previous research [15], [18], we incorporate robust victim models for evaluation, including Augmix [74], stylized training (SIN and SIN-IN) [75], adversarial training (AT) [2], [76], and Ensemble AT [3]. Table 6 presents the results for tSuc and untargeted transfer success rate (uSuc) under the aforementioned settings.

Firstly, the proposed S^{4}ST transformation significantly outperforms existing methods across all scenarios. Moreover, upon closer examination, it is evident that S^{4}ST benefits the most from increased computational effort and a larger number of surrogate models—whether it is averaging over multiple variants of a single surrogate model or computing gradients across models. Notably, it achieves performance close to that of the cross-model ensemble in the self-ensemble case, highlighting its potential for practical application. This is particularly relevant since obtaining multiple distinct surrogate models is generally challenging in the black-box scenario. Additionally, adversarially trained models with larger perturbation budgets exhibit significantly improved defensive capabilities, yet they remain susceptible to erroneous (untargeted) predictions induced by targeted AEs.

TABLE 4
Comparison between input transformations based on ResNet-50 as the surrogate model.

Attack	tSuc	Time	MNv2	EN	CNX	INv3	INv4	IRv2	Xcep	AvgC	ViT	Swin	MaxViT	Twins	PiT	TNT	DeiT	AvgT
TMI [8], [9]	0.8	1.24	2.0	1.5	2.4	0.6	0.7	0.7	0.9	1.3	0.2	0.9	0.4	0.1	0.1	0.2	0.2	0.3
+Admix [24]	2.1	6.27	5.9	4.7	6.3	0.9	1.6	2.4	2.5	3.5	0.3	2.0	1.0	0.5	1.1	0.6	0.2	0.8
+SI [25]	4.4	6.11	8.5	9.6	8.4	5.7	5.3	6.3	7.0	7.3	1.1	2.4	1.5	1.2	2.5	1.3	0.6	1.5
+SSA [44]	19.8	1.43	35.0	32.0	35.1	22.5	30.4	35.1	33.0	31.9	4.8	11.1	7.1	7.4	10.8	7.2	6.0	7.8
+DI [10]	22.4	1.44	29.6	33.7	48.8	15.4	35.4	38.1	28.3	32.8	5.9	17.1	10.8	11.1	19.1	11.6	8.3	12.0
+RDI [43]	27.3	1.24	37.1	42.9	47.4	25.8	39.8	45.6	36.9	39.4	11.0	21.7	13.2	13.6	22.3	15.0	10.2	15.3
+ODI [34]	58.1	5.11	73.2	77.1	75.3	64.3	75.7	77.0	73.2	73.7	37.6	54.1	37.9	38.4	54.5	43.0	31.7	42.5
+SIA [11]	58.3	1.55	82.5	81.3	84.3	51.1	74.2	76.2	72.0	74.5	30.8	58.5	42.6	39.3	53.2	40.7	29.8	42.1
+DeCoWA [13]	58.4	2.78	74.9	78.0	77.4	56.7	74.4	77.7	73.7	73.3	40.6	58.4	36.1	36.6	55.1	44.1	33.3	43.5
+BSR [12]	60.0	1.84	81.8	80.3	81.4	62.6	79.3	78.1	76.6	77.2	32.2	56.8	38.4	39.6	55.7	44.2	33.5	42.9
+T-Aug [26]	65.5	1.33	82.1	82.7	81.9	72.4	84.1	84.4	82.8	81.5	41.7	61.9	47.5	45.2	64.2	50.2	36.4	49.6
+H-Aug [26]	68.8	5.64	81.8	82.4	85.9	69.7	86.0	82.8	84.7	81.9	43.1	73.4	56.7	59.0	63.0	54.1	40.4	55.7
+S^{4}ST	**83.0**	1.45	**93.5**	**93.9**	**93.9**	**85.9**	**93.0**	**92.9**	**93.1**	**92.3**	**63.0**	**86.3**	**74.8**	**73.8**	**80.1**	**75.1**	**63.2**	**73.8**

The best and the second best results are in **bold** and underlined, respectively.

TABLE 6
Results of attacking robust models (tSuc/uSuc) and ensemble-based attacks.

Attack	Augmix [74]	Stylized [75]		Adversarial training (AT) [2]				Ensemble AT [3]			Avg.	AvgC	AvgT
		SIN	SIN-IN	$\ell_2 = .1$	$\ell_2 = .5$	$\ell_\infty = .5$	$\ell_\infty = 1.$	INv3_{ens3}	INv3_{ens4}	IRv2_{ens}			
Surrogate: ResNet-50													
TMI [8], [9]	3.4/55.8	0.5/51.6	33.7/83.4	0.2/42.4	0.0/28.6	0.0/24.3	0.0/19.1	0.0/17.8	0.0/17.9	0.0/12.4	3.8/35.3	1.3	0.3
+ODI [34]	78.5/94.6	45.2/87.6	96.5/99.4	65.6/88.8	3.5/44.5	6.4/46.2	0.6/28.2	14.4/51.4	5.8/40.1	3.2/27.6	32.0/60.8	73.7	42.5
+SIA [11]	89.1/96.7	39.8/84.6	99.6/**100.0**	64.1/87.5	1.5/41.5	3.7/40.2	0.1/27.2	5.7/37.7	1.7/32.5	0.6/19.9	30.6/56.8	74.5	42.1
+DeCoWA [13]	82.8/95.7	45.2/87.9	97.7/99.8	72.2/91.7	4.0/48.5	8.2/49.0	0.7/30.2	16.8/58.2	8.8/49.0	4.0/34.1	34.0/64.4	73.3	43.5
+BSR [12]	86.5/96.1	48.0/87.0	98.9/99.9	69.6/89.9	2.8/44.9	5.5/46.9	0.3/25.7	11.0/48.7	3.8/38.5	2.1/27.6	32.9/60.5	77.2	42.9
+T-Aug [26]	87.4/97.1	44.2/88.6	97.6/99.8	71.8/90.0	3.7/46.8	7.0/47.2	0.3/28.8	19.5/57.1	8.0/46.1	4.6/33.4	34.4/63.5	81.5	49.6
+H-Aug [26]	80.7/94.5	34.8/81.0	93.4/97.8	62.0/86.7	2.8/43.9	5.6/44.4	0.3/27.9	13.0/49.6	5.4/39.0	2.5/26.6	30.1/59.1	81.9	55.7
+S^4ST (Ours)	**94.9/98.8**	**64.1/92.3**	**99.8**/99.9	**89.2/96.7**	**13.7/54.5**	**21.2/57.5**	**2.2/31.0**	**42.3/71.0**	**20.7/55.4**	**12.4/40.1**	**46.1/69.7**	**92.3**	**73.8**
Surrogate: ResNet-50 × 10 (Self-Ensemble)													
$\text{TMI}_{Self\text{-}Ens}$	3.4/55.8	0.5/51.6	33.7/83.4	0.2/42.4	0.0/28.6	0.0/24.3	0.0/19.1	0.0/17.8	0.0/17.9	0.0/12.4	3.8/35.3	1.3	0.3
+ODI [34]	92.2/98.2	66.9/93.3	98.9/**99.9**	83.2/94.4	7.5/51.1	12.5/52.2	1.1/30.6	27.8/65.6	14.0/53.3	7.0/39.8	41.1/67.8	87.0	63.6
+SIA [11]	92.7/98.2	47.0/86.0	**99.9/99.9**	74.1/90.4	2.0/43.7	4.5/44.1	0.5/27.0	7.6/43.0	2.8/35.2	1.0/24.2	33.2/59.2	81.4	53.1
+DeCoWA [13]	88.1/96.9	56.4/93.4	97.8/99.8	78.0/92.8	4.7/47.8	8.5/50.2	0.8/31.0	18.7/59.9	9.3/49.6	4.4/36.2	36.7/65.8	78.4	52.8
+BSR [12]	91.3/97.8	54.4/88.7	99.5/**99.9**	74.0/91.2	2.8/44.6	6.3/47.9	0.5/27.1	11.6/52.0	5.6/42.6	2.6/31.3	34.9/62.3	82.6	51.5
+T-Aug [26]	92.6/98.3	56.2/91.1	98.6/99.8	81.1/93.5	5.7/51.1	10.8/51.0	0.5/29.5	24.6/62.4	11.5/50.4	6.5/37.0	38.8/66.4	87.3	62.6
+H-Aug [26]	92.8/98.2	58.4/90.5	98.1/99.8	81.4/94.7	8.0/49.4	13.5/51.7	0.6/30.0	29.8/65.0	14.2/51.7	6.9/37.6	40.4/66.9	90.4	73.3
+S^4ST (Ours)	**98.3/99.8**	**79.8/96.1**	**99.9/99.9**	**94.3/98.5**	**25.0/63.8**	**35.3/65.8**	**3.5/34.4**	**64.0/83.9**	**41.5/70.9**	**28.7/60.0**	**57.0/77.3**	**95.8**	**83.7**
Surrogate: ResNet-50 + ResNet-152 + DenseNet-121 + VGG16-bn (Cross-Model Ensemble)													
TMI_{Ens}	28.0/69.1	5.0/56.6	67.8/89.5	5.4/49.4	0.2/28.6	0.1/26.2	0.0/21.0	0.0/19.7	0.0/19.4	0.0/12.2	10.7/39.2	23.0	9.7
+ODI [34]	93.0/97.4	61.8/88.6	98.9/99.8	79.7/92.7	5.6/48.7	10.0/49.8	0.9/30.1	20.8/55.7	10.1/42.9	5.6/30.5	38.6/63.6	91.8	73.8
+SIA [11]	98.4/99.5	68.5/92.5	99.8/**99.9**	90.2/97.1	6.9/50.4	13.0/52.8	1.0/30.8	23.5/57.4	9.9/43.1	5.4/30.2	41.7/65.4	95.9	83.9
+DeCoWA [13]	97.7/99.0	77.0/96.2	99.4/**99.9**	93.4/98.0	13.4/57.7	21.7/59.8	2.4/35.4	51.9/79.7	32.5/65.6	17.7/49.0	50.7/74.0	95.2	82.2
+BSR [12]	98.1/99.6	73.0/94.2	99.8/**99.9**	89.8/96.8	7.3/52.0	15.4/55.9	1.0/31.3	35.0/67.1	14.0/51.3	8.4/37.1	44.2/68.5	95.4	78.7
+T-Aug [26]	97.6/99.4	67.7/93.9	99.7/**99.9**	90.3/97.4	10.5/52.5	16.5/57.4	1.4/33.3	48.8/75.4	24.3/56.9	14.7/44.8	47.2/71.1	96.3	83.1
+H-Aug [26]	97.1/99.1	62.9/91.2	99.4/**99.9**	88.1/96.0	8.6/52.2	14.8/54.6	1.1/31.5	39.4/69.0	17.1/52.6	9.1/39.6	43.8/68.6	96.5	85.3
+S^4ST (Ours)	**99.3/99.9**	**81.6/96.3**	**99.9/99.9**	**96.6/99.0**	**27.2/64.4**	**39.3/68.6**	**4.5/36.3**	**70.8/87.1**	**47.0/72.5**	**30.5/56.9**	**59.7/78.1**	**98.6**	**92.5**

The best results are in **bold**. uSuc is the ratio (in %) of the false predictions caused by the targeted AE by the original correct predictions.

TABLE 7
Comparison with resource-intensive attacks under the 10-Targets (all-source) setting.

Attack	CNNs						Vision transformers						Robust mechanisms					
	CNX	INv3	INv4	IRv2	Xcep	Avg.	ViT	Swin	PiT	TNT	DeiT	Avg.	Augmix	SIN	SIN-IN	$\text{AT}_{\ell_2=.5}$	$\text{AT}_{\ell_\infty=.5}$	Avg.
Resource-intensive TTAs																		
DFA [14]	25.3	1.6	15.0	16.3	13.8	14.4	0.7	3.1	1.6	1.7	3.9	2.2	19.9	7.8	56.4	0.3	0.3	16.9
BSR+DFA [12], [14]	46.0	16.2	37.8	37.5	35.1	34.5	4.5	9.8	6.9	7.3	7.0	7.1	33.0	17.2	49.9	0.6	0.9	20.3
TTP [15]	54.7	38.4	56.6	57.9	60.6	53.6	30.0	35.6	24.4	28.6	11.4	26.2	68.7	27.8	84.1	2.9	5.4	37.8
M3D [18]	**84.0**	**65.9**	**83.5**	**85.9**	**84.4**	**80.7**	**60.8**	**67.7**	**62.0**	**62.2**	**41.0**	**58.7**	**88.0**	**62.9**	**94.6**	**5.2**	**9.2**	**52.0**
Simple TTAs																		
TMI [8], [9]	5.4	1.1	1.9	2.1	2.1	2.5	0.8	1.7	1.1	1.0	0.7	1.1	4.4	1.4	37.1	0.4	0.4	8.7
+ODI [34]	74.3	62.6	73.6	75.1	70.8	71.3	37.6	53.3	55.5	42.8	32.2	44.3	78.1	45.2	96.8	6.1	9.6	47.2
+SIA [11]	87.5	53.9	78.5	79.9	73.6	74.7	34.6	64.2	58.1	47.8	35.6	48.1	90.6	44.1	**99.7**	3.6	5.9	48.8
+DeCoWA [13]	81.1	55.3	73.0	79.1	69.9	71.7	43.3	58.1	60.1	46.0	33.9	48.3	86.0	46.9	98.3	7.4	11.5	50.0
+BSR [12]	83.5	62.3	79.4	80.4	73.9	75.9	35.6	59.8	60.6	48.6	34.3	47.8	88.0	52.1	98.8	5.2	8.9	50.6
+T-Aug [26]	81.1	70.8	83.7	85.1	79.7	80.1	43.1	64.6	66.4	51.6	37.1	52.6	87.4	44.5	98.7	6.6	10.6	49.6
+H-Aug [26]	86.3	69.0	84.2	82.8	82.5	81.0	42.9	72.3	65.2	57.2	40.9	55.7	80.8	36.4	92.7	5.5	8.7	44.8
+S^4ST (Ours)	**96.0**	**86.2**	**94.5**	**95.0**	**94.7**	**93.3**	**69.1**	**88.0**	**84.9**	**80.1**	**66.5**	**77.7**	**95.8**	**68.1**	**99.5**	**18.1**	**28.3**	**62.0**

The best results in respective groups are in **bold**.

5.4 Comparison Against Resource-Intensive Attacks

We compare input transformation-empowered simple attacks with the resource-intensive ones, including generative methods—{TTP [15] and M3D [18]} and universal attacks—{DFA [14] and BSR+DFA}. It is computationally forbidden to train these methods to attack the massive target labels specified by the ImageNet-Compatible dataset. Therefore, we follow the 10-Targets (all-source) setting [15] that averages tSuc on 10 target labels—{`Grey-Owl`, `Goose`, `Bulldog`, `Hippopotamus`, `Cannon`, `Fire-Truck`, `Model-T`, `Parachute`, `Snowmobile`, `Street sign`} against the whole dataset. We use the pre-trained generators provided by the authors for TTP and M3D, which were trained over 50,000 additional images from the ImageNet train set. We also allow 50,000 images from the ImageNet validation set for DFA training. Results are reported in Table 7. The results indicate that the transfer attack capabilities of simple TTAs based on existing transformation methods are inferior to resource-intensive approaches, particularly when transferring to vision transformers and robust models. However, our S^4ST-TMI surpasses the current SOTA resource-intensive methods, M3D, with an average performance improvement of 13.9% across three types of target models. It also offers the flexibility to handle any (sample, label) pair and the advantages of not requiring model training or extensive data (as illustrated in Fig. 1), thereby establishing a new SOTA for TTAs.

5.5 Transferring Against Real-World API

Evaluation against the commercially deployed models in the real world is performed to verify the practicality of our method. Fig. 17 provides several examples obtained from the Google Cloud Vision API (object detection)[5] API. And we further follow the quantitative evaluation setting of M3D to perform transfer attacks on 5 labels over 100 images against the label detection[6] API. tSuc here is measured if the target label appears in the returned label list. Semantically similar labels are treated as the same one. We compare S^4ST-TMI and M3D as well as provide the

5. https://cloud.google.com/vision/docs/object-localizer
6. https://cloud.google.com/vision/docs/labels

Fig. 17. Examples of successful attacks on the Google Cloud Vision API (object detection), are better viewed by zooming in.

Fig. 18. Targeted AEs, bolstered by our S^4ST transformation, unveil the potential mechanism for enhancing targeted transferability through image transformations—by strengthening the manipulation of objects, textures, and backgrounds within individual images to generate coherent semantics that align with the intended target label.

TABLE 8
Results of attacking Google Cloud Vision API.

Target class	Similar classes	M3D [18]	S^4ST-TMI	S^4ST-TMI$_{Ens}$
Grey-Owl	Owl, Screech owl, great grey owl	23	41	62
Goose	geese and swans	39	50	51
Bulldog	Dog, Companion dog	46	68	91
Parachute	Parachute	28	78	85
Street sign	Signage, Street sign	39	59	77
	Average	35.0	59.2	73.2

results of S^4ST-TMI$_{Ens}$ in Table 8. It shows that the average tSuc of our method is over 24.2% compared to M3D, and utilizing ensemble surrogates further achieves an average tSuc up to 73.2%. These impressive results indicate our S^4ST-empowered attacks are promising in assessing the vulnerability of models with unknown training data distribution and data processing schemes.

5.6 Visual Inspections

Fig. 18 provides examples of targeted adversarial samples. It demonstrates that our S^4ST significantly enhances the infusion of the target label's semantics into AEs, a process that is further augmented when attacking an ensemble of surrogate models. For instance, in the first set of examples, the bird crista is subtly altered to resemble the windows of a church loft; in the sixth set, the background among the flock of geese is imbued with the texture of a conch. These observations are consistent with previous findings on unbounded perturbations [19] and data-agnostic perturbations [14], [15], but they further underscore the manipulation of objects, textures, and backgrounds within individual images. More effective transformations can facilitate this process, revealing potential avenues for enhancing targeted transferability through transformations. However, this also reveals an intrinsic tension between stronger transferability and imperceptibility, which may provide a clue for distinguishing targeted AEs from untargeted ones and clean images, thereby aiding specializing defense strategies.

6 Conclusion

In this paper, we elucidate the crucial role of image transformations in TTAs. Moreover, we propose leveraging the self-transferability of AEs against basic transformations as an indicator of the black-box transferability induced by different complex transformations. This approach assists in revealing

the sufficiency of attacking on basic geometric transformations-induced features for facilitating targeted transferability. Furthermore, we introduce the surrogate self-alignment to blindly assess the efficacy of various basic transformations, revealing the superior effectiveness of the simple scaling transformation. Leveraging these insights, which are accessible under black-box conditions, we have developed a novel S^4ST transformation. This transformation sets a new benchmark in the realm of TTAs across multiple scenarios, significantly outperforming existing methods in both effectiveness and efficiency.

Appendix A
Details of Evaluated Basic Transformations

Our study investigates a range of basic transformations that are routinely utilized in training neural networks. To assess self-transferability and self-alignment, we strive to quantify the intensity of all transformations using a unified factor, denoted as $s \in [0, 1]$. We establish the intensity factor set $\mathcal{S}$ as an arithmetic sequence comprising 100 equally spaced samples within the interval $[0, 1]$. For transformations with multiple directions, we compile all variations at the same intensity to compute the average outcomes. For example, regarding rotation, we consider images rotated by an identical angle in both the clockwise and counterclockwise directions. Two special cases are translate and shear, for which we combine the variations in both horizontal and vertical directions and both the positive and negative angles. To encapsulate their multidirectional nature, we uniformly sample 100 grid points (s_1, s_2) within the $[0, 1] \times [0, 1]$ space, yielding four variants at the same intensity level. We ensure a sufficiently broad range of variation for each transformation to secure more precise outcomes.

These transformations are executed using *torchvision*[7], conforming to its parameter definitions. For an input $\boldsymbol{x}$ with dimensions $H \times W$, we succinctly describe the evaluation settings as follows.

- **Rotation** returns two images respectively rotated by $\texttt{angle} = s \times 180°$ (clockwise) and $\texttt{angle} = -s \times 180°$ (counterclockwise) degrees.
- **Scaling** returns two images respectively scaled to the $\texttt{size}$ of $(1 + 1.5s)H \times (1 + 1.5s)W$ (enlargement) and $\frac{1}{(1+1.5s)}H \times \frac{1}{(1+1.5s)}W$ (reduction).
- **Shear** returns four images respectively sheared by $(s_1 \times 90°, s_2 \times 90°)$, $(-s_1 \times 90°, s_2 \times 90°)$, $(s_1 \times 90°, -s_2 \times 90°)$, and $(-s_1 \times 90°, -s_2 \times 90°)$ degrees parallel to the x-axis and y-axis.
- **Perspective** returns one image applied perspective transformation with $\texttt{distortion_scale} = s$.
- **Flip** returns two images respectively flipped vertically and horizontally.
- **Crop** returns a square patch of $\boldsymbol{x}$ with an area of $(1 - 0.95s) \times H \times W$.
- **Translate** returns four images respectively translated by $(s_1 \times H, s_2 \times W)$, $(-s_1 \times H, s_2 \times W)$, $(s_1 \times H, -s_2 \times W)$, and $(-s_1 \times H, -s_2 \times W)$ pixels along the x-axis and y-axis.
- **Solarize** returns one image with pixels values $x_{i,j} \geq (1 - s)$ been inverted to $1 - x_{i,j}$.
- **Hue** returns two images with the $\texttt{hue}$ factor jittered to $-0.5s$ and $0.5s$, respectively.
- **Brightness/Contrast/Saturation** returns two images with the $\texttt{brightness/contrast/saturation}$ factor jittered to $(1 + 4s)$ and $\frac{1}{1+4s}$, respectively.

7. https://github.com/pytorch/vision/tree/main/torchvision/transforms

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in International Conference on Learning Representations, 2014. 1

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in International Conference on Learning Representations, 2018. 1, 3, 11, 12

[3] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in International Conference on Learning Representations, 2018. 1, 11, 12

[4] Z. Yang, Q. Xu, W. Hou, S. Bao, Y. He, X. Cao, and Q. Huang, "Revisiting auc-oriented adversarial training with loss-agnostic perturbations," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023. 1

[5] X. Jia, Y. Zhang, X. Wei, B. Wu, K. Ma, J. Wang, and X. Cao, "Improving fast adversarial training with prior-guided knowledge," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024. 1

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in International Conference on Learning Representations), 2015. 1

[7] J. Gu, X. Jia, P. de Jorge, W. Yu, X. Liu, A. Ma, Y. Xun, A. Hu, A. Khakzar, Z. Li et al., "A survey on transferability of adversarial examples across deep neural networks," TMLR, 2024. 1

[8] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2018. 1, 2, 4, 9, 11, 12

[9] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4312–4321. 1, 2, 4, 5, 9, 11, 12

[10] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2730–2739. 1, 2, 3, 4, 5, 8, 9, 10, 11

[11] X. Wang, Z. Zhang, and J. Zhang, "Structure invariant transformation for better adversarial transferability," in Proceedings of the IEEE/CVF International Conference on Computer Vision, October 2023, pp. 4607–4619. 1, 4, 5, 7, 9, 11, 12

[12] K. Wang, X. He, W. Wang, and X. Wang, "Boosting adversarial transferability by block shuffle and rotation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 1, 4, 5, 9, 11, 12

[13] Q. Lin, C. Luo, Z. Niu, X. He, W. Xie, Y. Hou, L. Shen, and S. Song, "Boosting adversarial transferability across model genus by deformation-constrained warping," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 4, pp. 3459–3467, Mar. 2024. 1, 4, 5, 7, 11, 12

[14] C. Zhang, P. Benz, T. Imtiaz, and I. S. Kweon, "Understanding adversarial examples from the mutual influence of images and perturbations," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2020. 1, 4, 9, 12, 13

[15] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "On generating transferable targeted perturbations," in Proceedings of the IEEE/CVF International Conference on Computer Vision, October 2021, pp. 7708–7717. 1, 4, 9, 11, 12, 13

[16] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1765–1773. 1

[17] Z. Wang, H. Yang, Y. Feng, P. Sun, H. Guo, Z. Zhang, and K. Ren, "Towards transferable targeted adversarial examples," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2023, pp. 20 534–20 543. 1, 4

[18] A. Zhao, T. Chu, Y. Liu, W. Li, J. Li, and L. Duan, "Minimizing maximum model discrepancy for transferable black-box targeted attacks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2023, pp. 8153–8162. 1, 2, 4, 6, 9, 11, 12, 13

[19] Z. Zhao, Z. Liu, and M. Larson, "On success and simplicity: A second look at transferable targeted attacks," in Advances in Neural Information Processing Systems, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 6115–6128. 1, 2, 3, 4, 5, 9, 13

[20] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in International Conference on Learning Representations, 2017. 1

[21] M. Li, C. Deng, T. Li, J. Yan, X. Gao, and H. Huang, "Towards transferable targeted attack," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2020. 1, 3, 4, 5

[22] J. Weng, Z. Luo, S. Li, N. Sebe, and Z. Zhong, "Logit margin matters: Improving transferable targeted adversarial attack by logit calibration," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 3561–3574, 2023. 1, 3, 4, 5, 6, 9

[23] X. Sun, G. Cheng, H. Li, L. Pei, and J. Han, "On single-model transferable targeted attacks: A closer look at decision-level optimization," IEEE Transactions on Image Processing, vol. 32, pp. 2972–2984, 2023. 1, 2, 3, 5

[24] X. Wang, X. He, J. Wang, and K. He, "Admix: Enhancing the transferability of adversarial attacks," in Proceedings of the IEEE/CVF International Conference on Computer Vision, October 2021, pp. 16 158–16 167. 2, 4, 5, 9, 11

[25] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in International Conference on Learning Representations, 2020. 2, 4, 5, 9, 11

[26] Z. Wei, J. Chen, Z. Wu, and Y.-G. Jiang, "Rethinking data augmentation for improving transferable targeted attacks," 2023. [Online]. Available: https://openreview.net/forum?id=go0P5gsBE2 2, 4, 5, 9, 11, 12

[27] ——, "Enhancing the self-universality for transferable targeted attacks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2023, pp. 12 281–12 290. 2, 4, 5, 9

[28] G. Ortiz-Jiménez, A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard, "Optimism in the face of adversity: Understanding and improving deep learning through adversarial robustness," Proceedings of the IEEE, vol. 109, no. 5, pp. 635–659, 2021. 3

[29] L. Fowl, M. Goldblum, P.-y. Chiang, J. Geiping, W. Czaja, and T. Goldstein, "Adversarial examples make strong poisons," Advances in Neural Information Processing Systems, vol. 34, pp. 30 339–30 351, 2021. 3

[30] S. Chen, G. Yuan, X. Cheng, Y. Gong, M. Qin, Y. Wang, and X. Huang, "Self-ensemble protection: Training checkpoints are good data protectors," in International Conference on Learning Representations, 2023. 3

[31] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, Q. A. Chen, K. Fu, and Z. M. Mao, "Adversarial sensor attack on lidar-based perception in autonomous driving," in Proceedings of the 2019 ACM SIGSAC conference on computer and communications security, 2019, pp. 2267–2281. 3

[32] Y. Deng, X. Zheng, T. Zhang, C. Chen, G. Lou, and M. Kim, "An analysis of adversarial attacks and defenses on autonomous driving models," in 2020 IEEE international conference on pervasive computing and communications (PerCom). IEEE, 2020, pp. 1–10. 3

[33] S. Aneja, L. Markhasin, and M. Nießner, "Tafim: Targeted adversarial attacks against facial image manipulations," in European Conference on Computer Vision. Springer, 2022, pp. 58–75. 3

[34] J. Byun, S. Cho, M.-J. Kwon, H.-S. Kim, and C. Kim, "Improving the transferability of targeted adversarial examples through object-based diverse input," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2022, pp. 15 244–15 253. 3, 4, 9, 11, 12

[35] A. Kurakin, I. Goodfellow, S. Bengio et al., "Adversarial examples in the physical world," in International Conference on Learning Representations, 2017. 3

[36] N. Inkawhich, K. Liang, B. Wang, M. Inkawhich, L. Carin, and Y. Chen, "Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability," in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 20 791–20 801. 4

[37] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of big data, vol. 6, no. 1, pp. 1–48, 2019. 4

[38] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in 2018 international interdisciplinary PhD workshop (IIPhDW). IEEE, 2018, pp. 117–122. 4

[39] L. Taylor and G. Nitschke, "Improving deep learning with generic data augmentation," in 2018 IEEE symposium series on computational intelligence (SSCI). IEEE, 2018, pp. 1542–1547. 4

[40] H. Yuan, Q. Chu, F. Zhu, R. Zhao, B. Liu, and N. Yu, "Automa: Towards automatic model augmentation for transferable adversarial attacks," IEEE Transactions on Multimedia, vol. 25, pp. 203–213, 2021. 4

[41] Z. Yuan, J. Zhang, and S. Shan, "Adaptive image transformations for transfer-based adversarial attack," in European Conference on Computer Vision. Springer, 2022, pp. 1–17. 4, 6

[42] R. Zhu, Z. Zhang, S. Liang, Z. Liu, and C. Xu, "Learning to transform dynamically for better adversarial transferability," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2024, pp. 24 273–24 283. 4, 5

[43] J. Zou, Z. Pan, J. Qiu, X. Liu, T. Rui, and W. Li, "Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting," in European Conference on Computer Vision. Springer International Publishing, 2020, pp. 563–579. 4, 8, 9, 10, 11

[44] Y. Long, Q. Zhang, B. Zeng, L. Gao, X. Liu, J. Zhang, and J. Song, "Frequency domain model augmentation for adversarial attack," in European Conference on Computer Vision. Springer International Publishing, 2022, pp. 549–566. 4, 5, 7, 9, 11

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2016. 5, 9

[46] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," in International Conference on Learning Representations, 2018. 5

[47] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," in International Conference on Learning Representations, 2018. 5

[48] J. Wang, T. Zhang, S. Liu, P.-Y. Chen, J. Xu, M. Fardad, and B. Li, "Adversarial attack generation empowered by min-max optimization," in Advances in Neural Information Processing Systems, vol. 34. Curran Associates, Inc., 2021, pp. 16 020–16 033. 5

[49] W. Wu, Y. Su, M. R. Lyu, and I. King, "Improving the transferability of adversarial samples with adversarial transformations," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2021, pp. 9024–9033. 5

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255. 5

[51] S. Chen, Z. He, C. Sun, J. Yang, and X. Huang, "Universal adversarial attack on attention and the resulting dataset damagenet," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 4, pp. 2188–2197, 2020. 5

[52] Z. Wang, H. Guo, Z. Zhang, W. Liu, Z. Qin, and K. Ren, "Feature importance-aware transferable adversarial attacks," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 7639–7648. 5

[53] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626. 5

[54] Y. Li, S. Bai, Y. Zhou, C. Xie, Z. Zhang, and A. Yuille, "Learning transferable adversarial examples via ghost networks," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 11 458–11 465, Apr. 2020. 6

[55] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," NeurIPS, 2006. 7

[56] M. Huh, B. Cheung, T. Wang, and P. Isola, "Position: The platonic representation hypothesis," in Forty-first International Conference on Machine Learning. 8

[57] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2018. 9

[58] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proceedings of the International Conference on Machine Learning, vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114. 9

[59] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2022, pp. 11 976–11 986. 9

[60] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2016. 9

[61] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, Feb. 2017. 9

[62] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 2017. 9

[63] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words:

Transformers for image recognition at scale," in International Conference on Learning Representations, 2021. 9, 11

[64] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision, October 2021, pp. 10 012–10 022. 9

[65] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," in European Conference on Computer Vision. Springer International Publishing, 2022, pp. 459–479. 9

[66] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," in Advances in Neural Information Processing Systems, vol. 34. Curran Associates, Inc., 2021, pp. 9355–9366. 9

[67] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, October 2021, pp. 11 936–11 945. 9

[68] K. Han, A. Xiao, E. Wu, J. Guo, C. XU, and Y. Wang, "Transformer in transformer," in Advances in Neural Information Processing Systems, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 15 908–15 919. 9

[69] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in Proceedings of the 38th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, July 2021, pp. 10 347–10 357. 9

[70] R. Wightman, "Pytorch image models," https://github.com/rwightman/pytorch-image-models, 2019. 9

[71] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. 9

[72] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 2017. 11

[73] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, 2015. 11

[74] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple method to improve robustness and uncertainty under data shift," in International Conference on Learning Representations, 2020. 11, 12

[75] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." in International Conference on Learning Representations, 2019. 11, 12

[76] H. Salman, A. Ilyas, L. Engstrom, A. Kapoor, and A. Madry, "Do adversarially robust imagenet models transfer better?" in Advances in Neural Information Processing Systems, vol. 33. Curran Associates, Inc., 2020, pp. 3533–3545. 11