RETHINKING DATA AUGMENTATION FOR IMPROVING TRANSFERABLE TARGETED ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diverse input patterns induced by data augmentations prevent crafted adversarial perturbations from over-fitting to white-box models, hence improving the transferability of adversarial examples for non-targeted attacks. Nevertheless, current data augmentation methods usually perform unsatisfactory for transferable targeted attacks. In this paper, we revisit the commonly used data augmentation method - DI, which is originally proposed to improve non-targeted transferability and discover that its unsatisfactory performance in targeted transferability is mainly caused by the unreasonable restricted diversity. Besides, we also show that directly increasing the diversity of input patterns offers better transferability. In addition, our analysis of attention heatmaps suggests that incorporating more diverse input patterns into optimizing perturbations enlarges the discriminative regions of the target class in the white-box model. Therefore, these generated perturbations can activate discriminative regions of other models with high probabilities. Motivated by this observation, we propose to optimize perturbations with a set of augmented images that have various discriminative regions of the target class in the white-box model. Specifically, we design a data augmentation method, which includes multiple image transformations that can significantly change discriminative regions of the target class, to improve transferable targeted attacks by a large margin. On the ImageNet-compatible dataset, our method achieves an average of 92.5% targeted attack success rate in the ensemble transfer scenario, shedding light on transfer-based targeted attacks.

1 INTRODUCTION

Data augmentations have been widely used in current training paradigms of deep neural networks to improve the generalizability of learned models (Cubuk et al., 2019; 2020). It is also found that data augmentations mitigate the over-fitting of surrogate white-box models, which are oftentimes used to generate highly transferable non-targeted adversarial examples (Xie et al., 2019; Dong et al., 2019) to fool black-box models into incorrect predictions (Goodfellow et al., 2014). Specifically, previous works indicate that loss-preserving transformations provide an alternative visual representation of images, and models adopting augmented images as input can be considered as augmented models (Lin et al., 2019). As a result, integrating loss-preserving transformations into one model derives multiple augmented models, which can be attacked simultaneously to significantly improve non-targeted transferability (Dong et al., 2018). For example, three widely used and effective methods, Diverse Input (DI) (Xie et al., 2019), Translation-invariant (TI) (Dong et al., 2019) and Scale-invariant (SI) (Lin et al., 2019), restrict transformations within a small range to stabilize loss values. Note that DI which is originally proposed to defend against adversarial examples is also a loss-preserving transformation that preserves the model performance of benign images (Xie et al., 2017).

Nevertheless, while loss-preserving transformations offer considerable performance gains in transferable non-targeted attacks, it is unsuitable in transferable targeted attacks (Naseer et al., 2021; Zhao et al., 2021). The major challenge is that targeted transferability aims to fool models into the prediction of a target class rather than producing incorrect predictions as in the non-targeted scenario. To address it, one way is to precisely learn the target feature distribution and drive adversarial examples towards this distribution (Inkawhich et al., 2019; 2020a; Naseer et al., 2021), because decision boundaries of the target class share the same center among different models (Liu et al., 2016). However, these approaches require additional datasets for training auxiliary networks

so as to capture the target feature distribution. To mitigate this issue, Zhao et al. (2021) revisits iterative transferable non-targeted attacks without additional datasets and attributes the failure of these methods to unconvergence induced by a few iterations in iterative attacks. However, increasing the number of iterations in transferable non-targeted attacks attains limited benefit. Specifically, Zhao et al. (2021) directly applies DI into transferable targeted attacks. However, the assumption that loss-preserving transformations can be used as an augmented model (Lin et al., 2019) may not transfer to transferable targeted attacks. Unlike non-targeted attacks, targeted attacks are confronted with the gradient vanishment problem caused by the CE loss (Zhao et al., 2021). When the CE loss is close to 0, the gradient tends to vanish. It motivates us to address this problem by incorporating more diverse input patterns into iterative attacks. Therefore, we aim to study the influence of imposing more diverse input patterns into transferable targeted attacks and explore what attributes of image transformation can improve the target transferability of adversarial examples.

In this paper, we explore the effect of data augmentations in transferable targeted attacks with the help of DI. Specifically, we remove the restriction that the size difference between original images and randomly resized images should within a small range. We increase the size difference for optimizing perturbations on more diverse input patterns. Compared to the original version of DI, DI with a larger size difference postpones the arrival of gradient vanishment and achieves better performance. It demonstrates that this delay provides more useful gradients for optimizing perturbations. To further understand the improvement of transferability caused by unlimited DI, we analyze the difference of discriminative regions among different models by visualizing their attention heatmaps of the target class. We demonstrate that diversified augmented images have different discriminative regions in white-box models, and perturbations crafted on various discriminative regions can cover discriminative regions of the target class in other black-box models with a high probability, resulting in better target transferability. Moreover, we calculate the Intersection over Union (IoU) of discriminative regions w.r.t the target class between the original and augmented images to evaluate each image transformation. Following RandAugment (Cubuk et al., 2020), we propose a data augmentation method, which contains image transformations with low IoU, to improve transferable targeted attacks. Comprehensive experiments indicate that including multiple image transformations can eliminate over-fitting to the white-box models, achieving much better performance under the single-model transfer and ensemble transfer scenarios. We briefly summarize our primary contributions as follows:

- We provide fresh insights of DI into transferable targeted attacks. DI with a high size difference optimizes perturbations on diverse discriminative regions, resulting in improving target transferability.
- We propose the attention-deviation transformation that significantly changes discriminative regions. We also utilize the IoU of discriminative regions between the original and augmented images for the quantization of attention difference.
- Inspired by RandAugment, we propose a data augmentation method to boost the targeted attack success rates by combining multiple image transformations.

Overall, we hope the change of attention heatmaps caused by data augmentation can facilitate a better understanding of why transferable targeted attacks occur.

2 Background

2.1 Transferable Non-targeted Attacks

Let f_θ denote a white-box surrogate model, parameterized by θ which produces probabilities of all classes. We also use x to represent the benign image, y as the corresponding ground-truth label. Iterative Fast Gradient Sign Method (I-FGSM) (Kurakin et al., 2018) can be formulated as:

$$\begin{aligned} x_0^{adv} &= x, g_{i+1} = \nabla_x \mathcal{L}(f_\theta(x_i^{adv}), y) \\ x_{i+1}^{adv} &= Clip_{x,\epsilon}\{x_i^{adv} + \alpha \cdot sign(g_{i+1})\}, \end{aligned} \tag{1}$$

where x_i^{adv} denotes the adversarial example at the i-th iteration, $\mathcal{L}$ is the classification loss (*e.g.*, Cross Entropy), α is the step size, $Clip_{x,\epsilon}$ restricts perturbations centered on x with a radius ϵ. This iterative perturbation optimization leads to over-fit to the white-box model hence generating

adversarial examples with low transferability. To mitigate the effect of over-fitting, MI (Dong et al., 2018) integrates an additional momentum term into I-FGSM for stably updating perturbations:

$$g_0 = \mathbf{0}, \quad g_{i+1} = \mu \cdot g_i + \frac{\nabla_x \mathcal{L}(f_\theta(x_i^{adv}), y)}{||\nabla_x \mathcal{L}(f_\theta(x_i^{adv}), y)||_1}, \tag{2}$$

where μ is the decay factor. Different from MI, DI (Xie et al., 2019) optimizes perturbations on diverse input patterns:

$$g_{i+1} = \nabla_x \mathcal{L}(f_\theta(\mathcal{T}(x_i^{adv}, p, u)), y), \tag{3}$$

where $\mathcal{T}$ applies random resizing and padding with the probability p and an upper bound u that determines the size difference between original and resized images. u is set to 330 by default in MI. In subsequent analysis, we adjust the upper bound u to generate various augmented images with different strengths of image distortion. TI (Dong et al., 2019) applies image translation to evade over-fitting to discriminative regions of the white-box model. The discriminative regions are represented by the attention maps. Afterwards, they convert this data augmentation into convolving the gradients with a kernel W based on the translation-invariant property:

$$g_{i+1} = W * \nabla_x \mathcal{L}(f_\theta(x_i^{adv}), y). \tag{4}$$

In addition, other works attempt to design advanced gradient calculation methods (Lin et al., 2019; Wu et al., 2020a; Wang & He, 2021) to avoid over-fitting, and destroy critical features of predictions (Wu et al., 2020b; Wang et al., 2021) that may be shared among different models.

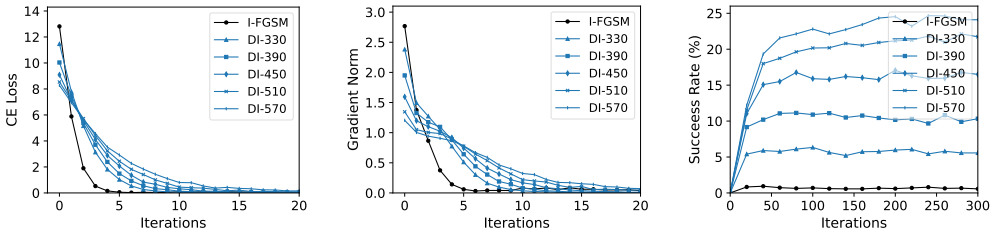
2.2 Transferable Targeted Attacks

Instead of maximizing the classification loss between the adversarial prediction and the ground-truth label in transferable non-targeted attacks, targeted attacks minimize the loss between the adversarial prediction and the targeted class y_t. Therefore, I-FGSM in transferable targeted attacks can be formulated as:

$$\begin{aligned} x_0^{adv} &= x, g_{i+1} = \nabla_x \mathcal{L}(f_\theta(x_i^{adv}), y_t) \\ x_{i+1}^{adv} &= Clip_{x,\epsilon}\{x_i^{adv} - \alpha \cdot sign(g_{i+1})\}. \end{aligned} \tag{5}$$

However, existing works in transferable non-targeted attacks perform unsatisfactory in transferable targeted attacks. Zhao et al. (2021) finds that the unreasonable limited number of iterations restricts the perturbation optimization in transferable targeted attacks. With a large number of iterations, DI, TI and MI achieve better target transferability. However, the problem of gradient vanishment of the Cross Entropy (CE) loss arises along with the large iterations. To address this problem, they propose to utilize the logit value of the targeted class as the classification loss $\mathcal{L}$. This simple Logit loss achieves better performance and consistently outperforms the Po+Trip loss (Li et al., 2020a) which utilizes the Poincaré distance to address the decreasing gradient problem of CE. Different from these two methods that concentrate on designing new loss functions, Wei et al. (2022) optimizes perturbations on global and local inputs for improving universality and target transferability. By adopting the "crop" operation, it can generate more diverse input patterns hence helps improve the target transferability. Different from Wei et al. (2022), this paper provides a comprehensive analysis on how data augmentation methods influence the target transferability. Based on the analysis, this paper also proposes a new data augmentation strategy to boost target transferability.

Apart from designing a new loss function for boosting target transferability, there is another line of work that focuses on training auxiliary networks to capture the feature distribution of the target class. For example, the Feature Distribution Attack (FDA) (Inkawhich et al., 2020b) utilizes a set of training data to train a tiny classifier, which predicts whether features extracted from the white-box model belong to the target class or not. In the process of attacking, FDA maximizes the probability of the tiny classifier to generate adversarial perturbations. It follows the similar idea in Activation Attack (AA) (Inkawhich et al., 2019) that disturbing intermediate features can transfer among different models. As an extension of FDA, FDA^N+xent (Inkawhich et al., 2020a) incorporates tiny classifiers into multiple layers and combines the CE loss as part of the optimization objective, resulting in better performance. In addition to training auxiliary classifiers, Transferable targeted perturbations (TTP) (Naseer et al., 2021) trains a generator for directly crafting adversarial examples that share similar intermediate features with the target samples. Yang et al. (2021) integrates a conditional class vector into the generator for multi-target class training. However, these methods require additional training datasets to train auxiliary networks. In this paper, we demonstrate that data augmentation with a large iteration can achieve much better performance.



(a) CE loss (b) Gradient magnitude (c) Transfer success rates

Figure 1: (a) and (b) show CE loss and gradient magnitude of attacks, respectively. (c) shows the targeted attack success rate on black-box models.

3 Data Augmentation in Transferable Targeted Attacks

In this section, we first delve into the data augmentation in transferable targeted attacks through an empirical study of DI with various values of the upper bound u. We demonstrate that more diverse input patterns yield better performance. Then, we provide fresh insights into transferable targeted attacks by considering the attention heatmaps of the target class in augmented images.

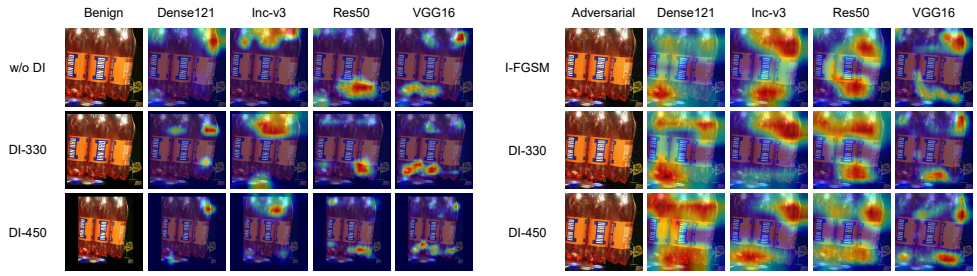
3.1 More Diverse Input Patterns in DI

Through combining the gradients of the loss function w.r.t parameters from various augmented images, data augmentation can generate more generic parameters so as to mitigate over-fitting and improve the generalization of networks (Simonyan & Zisserman, 2014; He et al., 2016b; Krizhevsky et al., 2017). From the viewpoint that adversarial perturbations can be regarded as optimized parameters (Lin et al., 2019), the perturbations optimized on diverse augmented inputs tend to be generalized to different models. Therefore, recent studies apply image transformations in transferable untargeted attacks to generate perturbations with high transferability (Xie et al., 2019; Dong et al., 2019; Lin et al., 2019). However, they are not suitable for transferable targeted attacks. One reason is that few iterations limit the convergence of attacks (Zhao et al., 2021). Despite enlarging the number of iterations can attain fairly high performance, it also reaches a plateau quickly since the CE loss is over-fitted to limited input patterns. The above observation motivates us to incorporate more diverse input patterns into transferable targeted attacks. Specifically, we lift restrictions of the upper bound in DI, which is originally set as a small number ($u = 330$) to prevent network performance from degrading on benign images (Xie et al., 2017). We explore DI with different upper bound u for including more diverse input patterns. The experiments are conducted on the ImageNet-compatible dataset with a DenseNet121 as the white-box model (Huang et al., 2017). These generated adversarial examples are limited with $\epsilon = 16/255$ and used to attack black-box models (ResNet50, VGGNet16, Inception-v3).

Fig.1(a) and Fig.1(b) show the attack curves of CE loss and gradient magnitude, respectively. As can be seen, I-FGSM decreases sharply within 5 iterations, while DI with a larger upper bound decreases slowly. In addition, when the CE loss is close to 0, the corresponding gradient tends to vanish, which is also observed in (Li et al., 2020b). These vanished gradients may lead to useless perturbation optimizations. To handle this problem, Zhao et al. (2021) replaces the CE loss with the logit output of the target class. In this paper, we argue that incorporating more diverse input patterns into attacks can also mitigate the gradient vanishing phenomenon. Fig.1(c) presents the targeted attack success rate (TASR) of DI with upper bound u on black-box models. We find that DI with a larger upper bound requires more iterations to converge, and the most diverse DI-570 outperforms DI-330 by a large margin. This empirical study indicates that more diverse input patterns of DI can mitigate the over-fitting problem, hence generate more transferable adversarial perturbation.

3.2 Analysis on Attention Heatmaps of DI

To further understand the effect of diverse input patterns in DI, we visualize attention heatmaps of the target class in different models for original and augmented images in Fig.2(a). Besides, the attention heatmaps for adversarial images generated by I-FGSM, DI-330 and DI-450 are also visualized in Fig.2(b). Note that attention heatmaps present the discriminative regions for each model.



(a) Benign images (b) Adversarial images

Figure 2: The attention heatmaps w.r.t the target class of four models (Densetnet121, Inception-v3, Resnet50 and VGG16) for (a) benign images and (b) adversarial images generated from densetnet121. Grad-CAM (Selvaraju et al., 2017) is utilized for visualization.

From Fig.2(a), we observe that four models have varied attention heatmaps of the target class for each image. It suggests that these models utilize different discriminative regions for the target prediction. As an adversarial example generated by one model may be highly related to the discriminative region of this model, it makes it hard to transfer to attack other models with different discrimination regions. This observation is also discussed in TI (Dong et al., 2019). Besides, we also find the attention heatmaps among original and augmented images are different for each model, which indicates that data augmentation can change the discriminative regions of models. Fig.2(b) further illustrates that optimizing perturbations on more diverse discriminative regions can enlarge the discriminative regions of the target class. Therefore, crafted perturbations with a large discriminative region can activate discriminative regions of other models with high probabilities. In addition, the area of discriminative regions becomes larger when increasing the upper bound u of DI, which suggests that the distortion magnitude of image transformations may be positively correlated with the area of discriminative regions in crafted perturbations, resulting in a higher targeted attack success rate.

To further illustrate the divergence of discriminative regions among different distortion magnitudes of DI, we report the Intersection over Union (IoU) of attention heatmaps between original and augmented images on randomly sampled 5,000 images from the ImageNet validation set. We binarize attention heatmaps by a threshold value of 0.5 to calculate IoU. As shown in Fig.3, the median IoU of 5,000 images decreases as the upper bound of DI increases. It suggests that one image transformation with a high distortion magnitude can significantly change discriminative regions of the target class. This motivates us to utilize the IoU as a metric to explore other image transformation methods for generating more diverse input patterns with different discriminative regions.

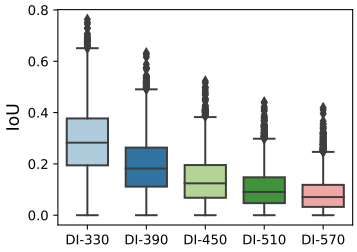


Figure 3: IoU of DI with different u.

In summary, the above analysis provides fresh insights into the performance improvements of data augmentations. Specifically, we empirically explain the success of DI under a large upper bound and find that diversified discriminative regions of augmented images drive optimized perturbations towards enlarging the area of the discriminative region w.r.t the target class. Besides, we propose the IoU metric to evaluate other image transformations for improving target transferability.

4 METHODOLOGY

For transferable targeted attacks, previous research has been devoted to utilizing additional networks to capture feature distributions of targeted classes (Inkawhich et al., 2019; 2020a;b; Naseer et al., 2021), or design a new loss function to avoid the phenomenon of gradient vanishment (Zhao et al., 2021). However, little attention is paid to data augmentation. In this section, we provide a comprehensive analysis of various image transformation methods used in training models based on the above proposed IoU metric. We then propose a data augmentation method to eliminate the over-fitting problem and improve target transferability.

4.1 EVALUATING IMAGE TRANSFORMATIONS

AutoAugment (Cubuk et al., 2019) and RandAugment (Cubuk et al., 2020) are effective data augmentation methods to train models. Based on the transformations they apply, we explore the following transformations:

- `solarize`
- `color`
- `contrast`
- `brightness`
- `sharpness`
- `shear-x`
- `shear-y`
- `translate-x`
- `translate-y`
- `flip`
- `crop`
- `rotate`

Where the first five transformations are used to alter the visual effects of images, the remaining transformations change the shape and size of images. Among them, "crop" and "rotate" convert images into more diverse patterns. We name the above two category data augmentation methods as visual and positional transformations, respectively. Intuitively, positional transformations may change the position of discriminative regions more significantly than visual transformations.

To support this hypothesis, we calculate IoUs of the discriminative regions w.r.t the target class between original and augmented images for each transformation. Following RandAugment, we linearly split the parameters of transformations into several magnitudes (0 to 10). However, in order to incorporate more diverse input patterns into image transformations, we randomly sample the parameters that are smaller than the given magnitude. Hence, a higher magnitude means adjusting images more significantly. The detailed parameters used in each transformation are provided in Appendix B. Fig.4 plots the average IoU values of each transformation at different magnitudes on 5,000 images randomly sampled from the ImageNet validation set. We observe that visual transformations attain higher IoU values than positional transformations under different magnitudes. Thus, image transformations with little change in discriminative regions may be useless for improving transferability. To support this hypothesis, we further conduct transferable targeted attacks with each transformation (see Appendix C). The improvement of visual transformations is small and fluctuating. In addition, "shear-x/y", "translate-x/y", and "flip" achieve lower IoUs, because each transformation change images in a single direction. In contrast, "crop" and "rotate" can generate more diverse input patterns. More significant diversity leads to better generalization of generated perturbations, as shown in Appendix C. This evaluation provides a set of image transformations that can deviate from discriminative regions of the original images, improving target transferability.

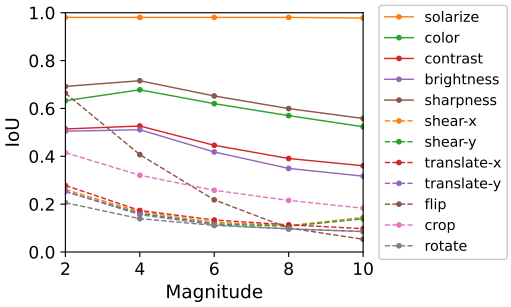


Figure 4: IoU of image transformations with different magnitudes.

4.2 DATA AUGMENTATION

Motivated by the above analysis, we resort to data augmentation for improving transferable targeted attacks. Specially, we introduce the attention-deviation transformation, defined as follow:

Definition 1. ***Attention-deviation transformation.*** *Given an image x with its target class t, if there exists an image transformation $T(\cdot)$ with the magnitude M that attains a low IoU of discriminative regions w.r.t t between x and $T(x)$, and generate diverse input patterns when the parameter of transformation is less than M, we term $T(\cdot)$ as a attention-deviation transformation.*

As discussed in Sec.3.2, the attention-deviation transformation provides various different discriminative regions in optimizing perturbations like DI, resulting in the expansion of discriminative regions of the white-box model so as to overlap specific discriminative regions of other black-box models. It is different from loss-preserving transformation Lin et al. (2019), which explores image transformations that are invariant to the outputs of models. As shown in Sec.4.1, we discover that positional transformations significantly alter discriminative regions. Thus, these positional transformations can be served as the attention-deviation transformation. Given the multiple transformations and driven by RandAugment, we propose a data augmentation method, which integrates different transformations to improve target transferability. Specifically, for each iteration, we randomly

sample N transformations from the given set of transformations T_{set}, and randomly generate the parameters of current transformations that are less than the magnitude M. Therefore, we have two hyper-parameters N and M in the data augmentation method. Algorithm is shown in Appendix D.

To illustrate the effect of tiny and huge changed transformations, we propose two versions of the proposed method. Tiny-Augment (T-Aug) defines $T_{set} = \{$"shear-x/y", "translate-x/y", "flip"$\}$, and Huge-Augment (H-Aug) uses T_{set} as $\{$"crop", "rotate", "DI"$\}$. Through generating diverse input patterns by these transformations, T-Aug and H-Aug enable alleviating over-fitting to the white-box model and activate discriminative regions of the target class on black-box models.

5 Experiments

In this section, we first conduct transferable targeted attacks to validate the effectiveness of the proposed data augmentation methods in single-model transfer and ensemble transfer scenarios. Then we evaluate the importance of each transformation in T-Aug or H-Aug.

5.1 Setup

Following (Zhao et al., 2021), we adopt four models with different architectures: ResNet50 (He et al., 2016a), DenseNet121 (Huang et al., 2017), VGGNet16 (Simonyan & Zisserman, 2014) with batch normalization Ioffe & Szegedy (2015) and Inception-v3 (Szegedy et al., 2016) and the ImageNet-compatible dataset[1] to conduct experiments. The NIPS 2017 Competition on Adversarial Attacks and Defenses firstly introduce this dataset, which contains 1,000 images and corresponding target classes for transferable targeted attacks. We restrict perturbations by ℓ_∞ norm with $\epsilon = 16/255$, and set the step size as $\alpha = 2/255$, the number of iterations as 300. We evaluate attack performance by the percentage of adversarial examples that the black-box models successfully classifies as the target class, which is termed as Targeted Attack Success Rate (TASR). We use an NVIDIA GeForce RTX 3090 with 24GB of memory to conduct experiments.

5.2 Comparison to State of the Art

We compare our method with DI and DI-TM, where the default parameters of TI and MI are directly used here, the upper bound and the transformation probability of DI are set as 330 and 1.0, respectively. DI with the probability of 1.0 leads to generating diverse input patterns in each attack iteration. For our method, under using DenseNet121 as the white-box model, a grid search is performed to find the hyper-parameters N and M on 200 images randomly sampled from the ImageNet-compatible dataset. Results are shown in Appendix E. Finally, we use $N = 5, M = 2$ for T-Aug, $N = 2, M = 8$ for H-Aug. We perform experiments using the CE and logit loss functions, respectively. Note that we overlook Po+Trip (Li et al., 2020b) because it has worse performance than Logit.

Single-model transfer. Table 1 reports the results when using one model as the white-box model. We can observe that T-Aug and H-Aug consistently outperform DI and DI-TM by a large margin under the 100th and 300th iterations. In particular, when using T-Aug to attack DenseNet121 from ResNet50, the crafted adversarial examples achieve 87.1% TASR on average. However, DI-TM has better performance than our method in some cases at the 20th iteration, because more diverse input patterns in our methods require a lot of iterations to converge. In addition, T-Aug and H-Aug lead alternatively in different attack scenarios but have similar performance in most cases. It demonstrates that more transformations with a lower magnitude in T-Aug are equivalent to fewer transformations with a higher magnitude in H-Aug. Therefore, combining them together may become redundant and degrade performance. Appendix F shows that the combination of T-Aug and H-Aug achieve worse performance than themselves. Besides, TM can improve the performance of both DI and our methods. It suggests that our methods are complementary to TM. We also find that integrating our methods into the Logit loss still obtains better performance than DI-TM with the Logit loss. Specifically, the Logit loss improves TASR from 11.3% to 23.1% when using T-Aug-TM to attack Inception-v3 from VGG16. However, the Logit loss degrades the performance of our methods in

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

Table 1: TASR (%) of several methods with the CE and Logit losses under ℓ_∞ norm with $\epsilon = 16/255$. We show TASRs with 20/100/300 iterations. TM is the combination of TI and MI. The best results are in bold.

Attack	White-box Model: Res50			White-box Model: Dense121		
	→Dense121	→VGG16	→Inc-v3	→Res50	→VGG16	→Inc-v3
DI	16.0/19.2/19.0	16.8/18.2/18.2	0.1/0.2/0.3	8.1/8.4/8.3	8.6/7.7/8.2	0.1/0.2/0.0
T-Aug	**20.1/76.1/90.6**	13.5/62.7/79.2	**2.0/18.3/27.1**	**9.5**/46.8/59.3	9.2/43.7/56.6	**2.0/13.5/18.2**
H-Aug	16.9/74.3/89.8	**22.3/81.3/92.8**	1.5/16.3/24.0	8.3/**53.8/71.2**	**17.3/74.2/87.6**	1.7/12.4/**18.2**
DI-TM	**27.1**/39.7/44.3	**18.9**/27.6/29.4	2.2/3.4/4.1	**12.9**/16.7/18.4	8.1/10.6/10.6	1.7/2.2/3.2
T-Aug-TM	17.1/**82.0/95.2**	10.0/63.0/82.8	**3.9/34.1/57.2**	9.9/52.0/69.6	7.1/41.7/58.3	**4.3/26.7**/40.2
H-Aug-TM	13.0/77.9/93.7	15.5/**80.7/94.7**	2.2/29.6/55.4	7.6/**57.0/82.2**	**13.7/72.3/90.7**	2.3/22.6/**42.7**

Attack	White-box Model: VGG16			White-box Model: Inc-v3		
	→Res50	→Dense121	→Inc-v3	→Res50	→Dense121	→VGG16
DI	0.2/0.2/0.1	0.0/0.0/0.0	0.0/0.0/0.0	**0.7**/0.7/0.9	0.2/0.8/1.1	0.4/0.5/1.0
T-Aug	**2.4**/10.8/14.7	**3.1**/18.0/24.1	**0.1**/1.6/**2.0**	0.4/4.1/7.3	0.6/9.4/13.8	0.6/4.3/7.3
H-Aug	2.1/**11.8/17.3**	1.9/**19.2/26.5**	**0.1/1.8**/1.5	0.5/**10.2/18.5**	**1.3/17.0/33.0**	**1.8/16.6/32.4**
DI-TM	0.6/0.6/0.5	0.4/0.3/0.4	0.0/0.0/0.0	0.8/1.8/2.4	0.8/2.4/2.9	0.7/1.3/1.8
T-Aug-TM	**2.5/18.1**/31.6	**4.3/28.7**/46.5	**0.4/6.7/11.3**	0.7/5.2/12.0	0.8/11.8/23.9	0.6/4.0/9.5
H-Aug-TM	1.4/**18.1/37.5**	2.9/25.2/**49.9**	0.2/4.4/10.5	**1.1/11.6/31.1**	**1.8/20.0/50.3**	**0.8/17.2/44.1**

(a) Attacks with the **CE** loss

Attack	White-box Model: Res50			White-box Model: Dense121		
	→Dense121	→VGG16	→Inc-v3	→Res50	→VGG16	→Inc-v3
DI	**24.3**/50.1/57.5	**24.0**/51.4/57.5	0.6/1.2/1.4	**13.8**/28.8/28.9	15.5/31.8/33.0	0.6/1.5/1.2
T-Aug	16.9/**69.5/83.6**	12.3/59.9/74.4	**2.0/21.7/33.5**	9.8/**49.8**/65.1	8.3/46.4/61.7	**2.6/15.3/25.7**
H-Aug	13.2/64.5/81.7	17.9/**71.6/83.7**	1.1/16.0/27.6	7.5/48.9/**67.6**	**15.6/65.1/77.6**	0.9/13.5/20.6
DI-TM	**30.4**/64.4/71.8	**22.6**/55.1/62.8	2.7/7.1/9.6	**16.1**/39.3/43.7	**13.5**/33.0/38.1	2.1/7.1/7.7
T-Aug-TM	15.3/**70.3/87.2**	9.8/58.1/79.1	**3.0/35.0/58.6**	9.0/**52.3**/71.6	7.9/44.0/64.9	**4.4/29.3/45.7**
H-Aug-TM	11.0/65.7/84.6	12.7/**70.1/86.0**	2.2/29.3/54.4	6.6/49.2/**74.5**	10.1/**63.3/79.2**	2.6/23.3/42.8

Attack	White-box Model: VGG16			White-box Model: Inc-v3		
	→Res50	→Dense121	→Inc-v3	→Res50	→Dense121	→VGG16
DI	1.2/3.6/3.1	0.7/3.4/4.0	0.0/0.0/0.1	0.4/1.5/1.6	0.5/2.1/3.1	0.4/1.7/3.1
T-Aug	**4.0/30.5/44.4**	**4.9/37.6/53.1**	**0.5/5.7/9.4**	0.8/5.1/9.5	0.6/8.3/16.5	0.5/5.3/9.6
H-Aug	2.9/21.3/34.4	2.8/29.0/44.9	0.1/3.3/5.2	**0.8/10.1/20.7**	**0.9/16.9/32.5**	**1.9/17.1/32.1**
DI-TM	3.0/9.6/11.3	3.2/12.0/13.7	0.1/0.6/0.7	0.9/2.0/2.8	1.1/3.3/5.0	0.6/2.2/3.9
T-Aug-TM	**4.0/33.5/56.5**	**5.4/41.0/63.2**	**0.3/11.5/23.1**	1.0/6.0/14.5	1.1/10.3/27.3	0.8/4.7/13.6
H-Aug-TM	2.2/24.2/46.8	2.7/30.2/57.2	0.3/8.0/18.4	**1.3/11.9/29.8**	**1.8/20.2/48.2**	**1.6/17.9/41.8**

(b) Attacks with the **Logit** loss.

some cases. It can be explained by the fact that the Logit loss originally proposed to increase the logit value of the target class can also improve that of other classes (See Appendix G).

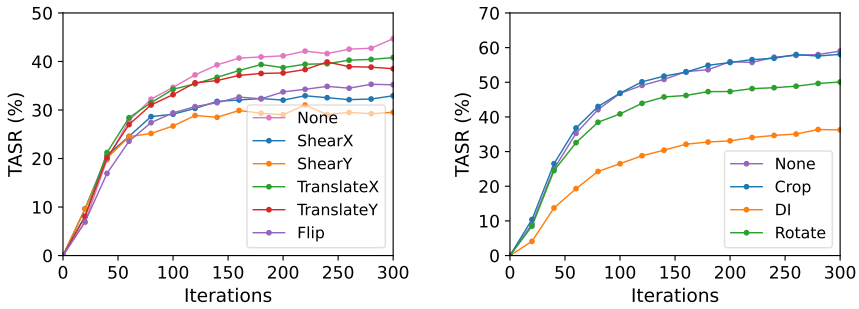
Ensemble transfer. Adversarial examples crafted from multiple white-box models tend to attack other models with a high probability (Dong et al., 2018). Therefore, we select one model as the black-box model, and the remaining models as the white-box models. Table 2 reports the results, which consistently illustrate the effectiveness of T-Aug and H-Aug on different loss functions. For example, H-Aug-TM with the CE loss achieves an average 92.5% TASR. However, the performance improvement of our methods with the Logit loss is inferior. This is because the unbounded Logit loss relies too much on white-box models with high logit values.

5.3 Importance of Each Transformation

The proposed T-Aug and H-Aug can significantly improve target transferability among different models. The number of randomly selected transformations $N = 5$ and the magnitude $M = 2$ in T-Aug suggest that integrating all tiny transformations with low distortion achieves similar performance as the huge transformations with high distortion. To further evaluate the effect of each

Table 2: TASR (%) of attacking one black-box model in ensemble transferable attacks. We report TASR at the 300th iteration. The equal weights are assigned to white-box models. The best results are in bold.

Ensemble Attack	Loss	Black-box Model				Average
		Res50	Dense121	VGG16	Inc-v3	
DI-TM	CE	31.1	55.2	51.6	16.1	38.5
DI-TM-LI	CE	42.8	55.2	67.9	24.2	47.5
T-Aug-TM	CE	93.4	**98.2**	92.1	**82.9**	91.6
H-Aug-TM	CE	**95.1**	**98.2**	**98.4**	78.5	**92.5**
DI-TM	Logit	70.2	82.3	82.2	29.1	65.9
DI-TM-LI	Logit	75.9	83.0	84.1	35.4	69.6
T-Aug-TM	Logit	**89.2**	**91.6**	87.4	**77.6**	**86.4**
H-Aug-TM	Logit	89.1	91.3	**92.3**	70.7	85.8



(a) T-Aug (b) H-Aug

Figure 5: Average TASR (%) deteriorates when an image transformation is excluded in (a) T-Aug and (b) H-Aug. "None" denotes no removal in the set of transformations.

transformation, we conduct experiments on T-Aug and H-Aug with random removal of an image transformation using DenseNet121 as the white-box model. The attacks utilize CE as the loss function and are based on I-FGSM without TI and MI. We set $N = 4$ for T-Aug. Fig.5 reports the average TASR of attacking other models. For T-Aug, performance degradation occurs when removing any of the image transformations. It suggests that each image transformation of T-Aug produces a marked effect on target transferability. Among these transformations, deleting "shear-x/y" reduces performance more significantly due to the more obvious changes in image shape. For H-Aug, we observe that removing "crop" or "rotate" achieves similar performance with including all transformations. However, including either of them with 'DI' together achieves higher performance than using "DI" individually. It demonstrates that the magnitude $M = 8$ of H-Aug can generate input patterns with high changes of discriminative regions using "DI" and either of "crop, rotate". Besides, deleting "DI" leads to significant degradation of TASR. It suggests that DI without limited upper bounds is most helpful among these transformations.

6 Conclusion

In this paper, we provide an exhaustive study on DI with unrestricted upper bounds. More diverse input patterns introduced by a larger upper bound would improve target transferability. Through visualizing discriminative regions, we explain the high performance of DI is caused by the fact that diversified discriminative regions of augmented images drive the crafted adversarial perturbations towards covering more discriminative regions. Therefore, we aim to utilize multiple image transformations to further improve transferable targeted attacks. We propose to exploit the IoU metric of discriminative regions between original and augmented images for filtering widely used image transformations. Based on selected transformations, we introduce two data augmentation methods in terms of tiny and huge changed transformations, named T-Aug and H-Aug. The experimental results demonstrate the effectiveness of T-Aug and H-Aug regardless of the used loss functions. In the future, we will adaptive adjust the magnitude M for each transformation to include all transformations together.

REFERENCES

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 113–123, 2019.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

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

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4312–4321, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016b.

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

Nathan Inkawhich, Wei Wen, Hai Helen Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7066–7074, 2019.

Nathan Inkawhich, Kevin Liang, Binghui Wang, Matthew Inkawhich, Lawrence Carin, and Yiran Chen. Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability. *Advances in Neural Information Processing Systems*, 33:20791–20801, 2020a.

Nathan Inkawhich, Kevin J Liang, Lawrence Carin, and Yiran Chen. Transferable perturbations of deep feature distributions. *arXiv preprint arXiv:2004.12519*, 2020b.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 638–646, 2020a.

Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 641–649, 2020b.

Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. On generating transferable targeted perturbations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7708–7717, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

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

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

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

Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7619–7628, 2021.

Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Incorporating locality of images to generate targeted transferable adversarial examples. *arXiv preprint arXiv:2209.03716*, 2022.

Dongxian Wu, Yisen Wang, Shutao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. *ArXiv*, abs/2002.05990, 2020a.

Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R. Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1158–1167, 2020b.

Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1161–1170, 2020c.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.

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

Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Boosting transferability of targeted adversarial examples via hierarchical generative networks. *arXiv preprint arXiv:2107.01809*, 2021.

Zhengyu Zhao, Zhuoran Liu, and Martha Larson. On success and simplicity: A second look at transferable targeted attacks. *Advances in Neural Information Processing Systems*, 34:6115–6128, 2021.

A ATTENTION HEATMAPS OF ORIGINAL LABELS

We visualize attention heatmaps of the original class on original and augmented images. As shown in Fig.A.1, the discriminative regions of each model are similar among various augmented images. This is because these models are trained on augmented images with the original class. Besides, different models have comparable attention for each image, similar to the finding in (Wu et al., 2020c). In contrast, attention heatmaps of the target class vary among augmented images. It suggests that transferable non-targeted attacks drive perturbations towards deviating discriminative regions of the ground-truth class, while transferable targeted attacks aim to highly activate the discriminative regions of the target class.

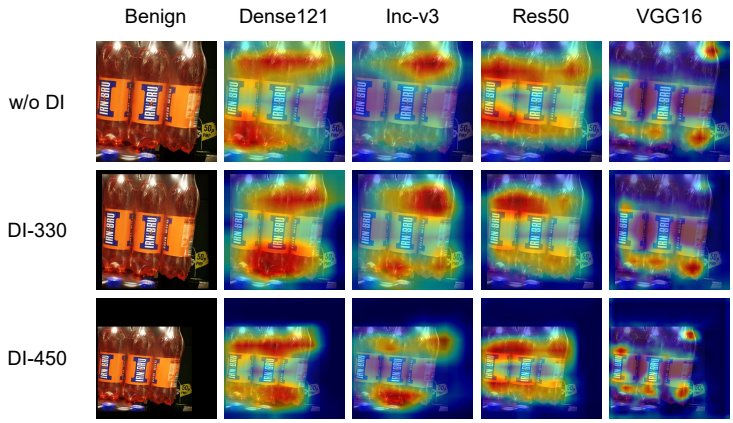


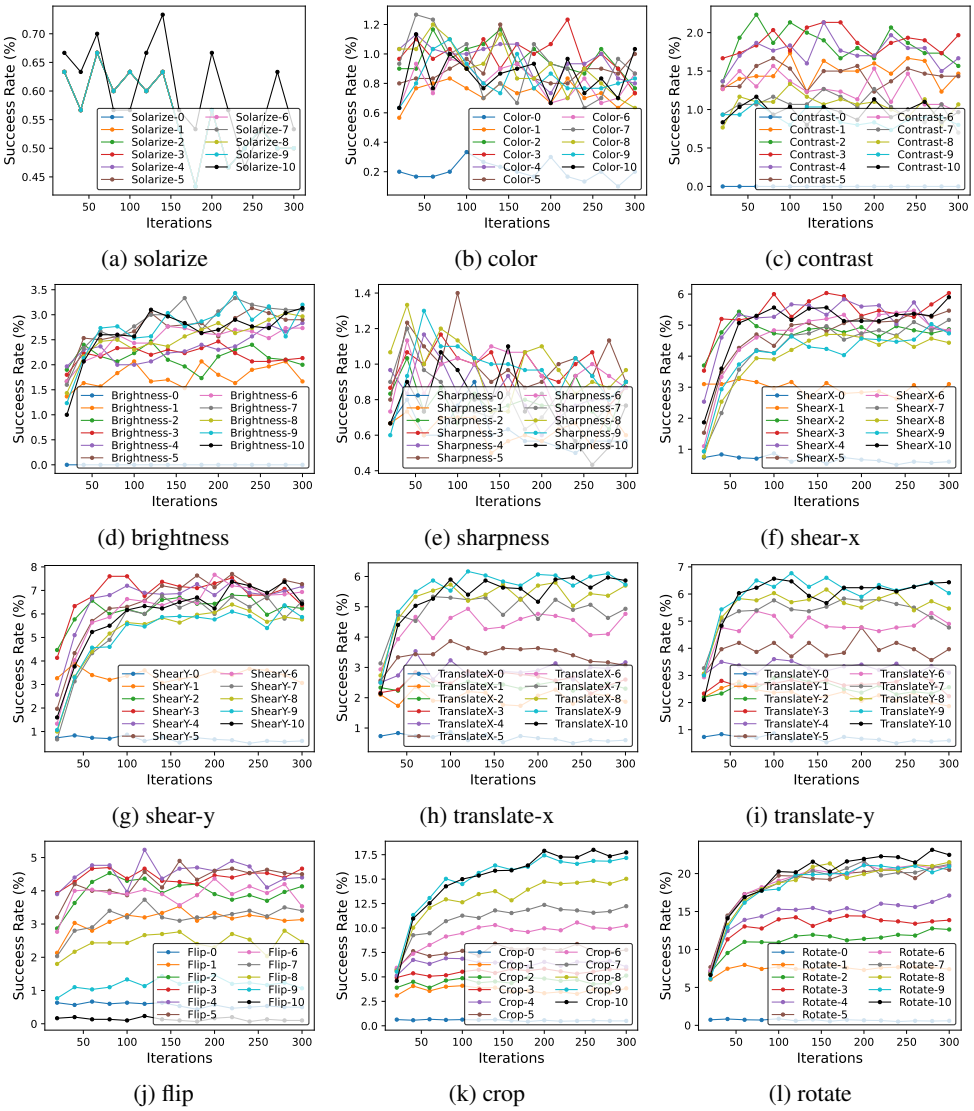
Figure A.1: The attention heatmaps the original class of Densetnet121, Inception-v3, Resnet50 and VGG16 models for augmented images.

B DETAILED PARAMETERS IN EACH TRANSFORMATION

Following RandAugment Cubuk et al. (2020), we linearly split the parameters of transformations into several magnitudes (0 to 10). The parameter intervals are shown in Table B.1. For a given magnitude M, we first calculate the corresponding value of the specific parameter by the parameter interval. We then randomly select one value smaller than the specified parameter as the parameter of image transformation to be performed. This method enables randomness of image transformations. Note that the parameter of "flip" is the probability of horizontally and vertically flip the given image. The parameter of "crop" is the lower bound for the crop area, the upper bound of "crop" is set to 1.0 by default, a subsequent resizing operation resizes the crop to the original size. The parameter of "DI" is the upper bound u and its probability p is set as 1.0 by default.

Table B.1: Parameters of each transformation. "Signed" denotes to randomly convert the parameter into the negative number.

Transformation	Lower distortion	Upper distortion	Signed
solarize	255	0	False
color	0	5	False
contrast	0	5	False
brightness	0	5	False
sharpness	0	5	False
shear-x/y	0	180	True
translate-x/y	0	299	True
flip	0	1	False
crop	1	0	False
rotate	0	180	True
DI	299	700	False



(a) solarize (b) color (c) contrast

(d) brightness (e) sharpness (f) shear-x

(g) shear-y (h) translate-x (i) translate-y

(j) flip (k) crop (l) rotate

Figure C.1: TASR (%) of each transformation with different magnitudes. We use DenseNet121 as the white-box model, and report the performance of attacking other three black-box models.

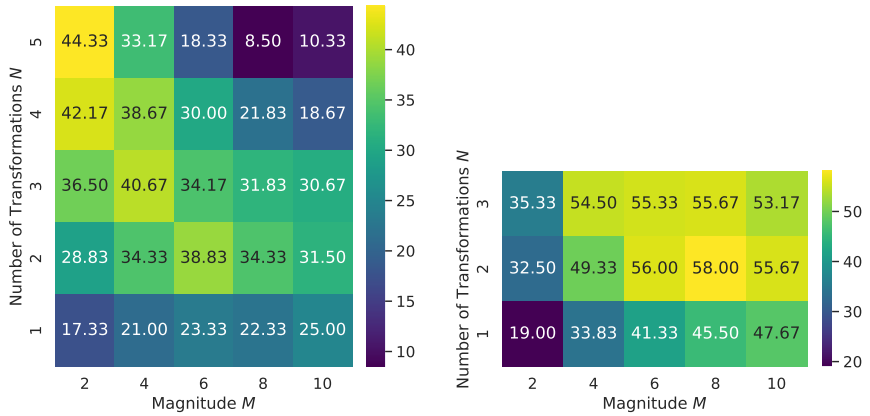
C TRANSFERABLE TARGETED ATTACKS WITH EACH TRANSFORMATION

We conduct transferable targeted attacks for each transformation based on I-FGSM with the CE loss, and use DenseNet121 as the white-box model. Fig.C.1 show the attack curves of I-FGSM under various image transformations. For visual transformations (Fig.C.1(a)-C.1(e)), we observe that these transformations with different magnitudes have fluctuated performance. It empirically suggests that visual transformations with tiny changes in discriminative regions are unsuitable for transferable targeted attacks. For positional transformation without "flip" (Fig.C.1(f)-C.1(i),C.1(k),C.1(l)), we observe that each of them improves target transferability. However, "shear-x/y" and "translate-x/y" lead to inferior improvement due to the monotonic input patterns. In contrast, "crop" and "rotate" with abundant changes achieve much higher performance. Besides, "flip" (Fig.C.1(j)) attains the worst performance at the magnitude 10. This is because that only the horizontal and vertical flipped image is used to optimize perturbations when the probability is 1.0. Overall, positional transformations with more diverse input patterns are helpful in improving target transferability than visual transformations, while they require more iterations to converge. Despite each tiny changed transformation has limited improvement on target transferability, including them together can generate more diverse input patterns.

Algorithm 1 Data Augmentation for improving transferable targeted attacks

Input: the classification loss function $\mathcal{L}$, white-box model f_θ, benign image x, targeted class y_t.
Parameter: The perturbation budget ϵ, iteration number I, step size α, a set of image transformations T_{set}, the number of transformation N, the magnitude M
Output: The adversarial example x^{adv}.

1: Initialize x_0^{adv} by x.
2: **for** $i = 0$ to $I - 1$ **do**
3: Generate T_{use} by randomly selecting N transformations from T_{set}.
4: Traverse each transformation in T_{use} with the magnitude M to generate x_i^{aug} from x_i^{adv}.
5: $g_{i+1} = \nabla_x \mathcal{L}(f_\theta(x_i^{aug}), t)$
6: $x_{i+1}^{adv} = Clip_{x,\epsilon}\{x_i^{adv} - \alpha \cdot sign(g_{i+1})\}$
7: **end for**
8: **return** x_I^{adv}



(a) T-Aug (b) H-Aug

Figure E.1: TASR (%) of (a) T-Aug and (b) H-Aug with different number of iterations N and magnitudes M. We use DenseNet121 as the white-box model, and report the average performance of the 300th iteration on attacking other three black-box models. 200 images randomly selected from the ImageNet-compatible dataset are used here.

D Algorithm of the Proposed Method

Algorithm 1 illustrates the data augmentation methods for improving transferable targeted attacks. According to the set of image transformations T_{set} provided, this algorithm can be divided into T-Aug and H-Aug. For each iteration, we optimize perturbations on a diverse input pattern generated from randomly selected N image transformations with a predefined magnitude M. The hyper-parameters N and M reduce the search space, following RandAugment (Cubuk et al., 2020). Our algorithm is easily combined with MI and TI, and is suitable for different loss functions.

E Hyper-parameters N and M

We search for the optimal number of image transformations N and magnitude M on a subset of the ImageNet-compatible dataset. To illustrate that these optimal parameters are shared in different white-box models, we conduct experiments when using DenseNet121 as the white-box model, and report the average TASR (%) of attacking other black-box models with T-Aug and H-Aug. For T-Aug, we set $N \in \{1, 2, 3, 4, 5\}, M \in \{2, 4, 6, 8, 10\}$, while for H-Aug, we set $N \in \{1, 2, 3\}, M \in \{2, 4, 6, 8, 10\}$. Fig.E.1(a) and E.1(b) report the results for T-Aug and H-Aug, respectively. Fig.E.1(a) shows the relative gain in TASR across increasing N when $M = 2$, and similar trends across increasing M when $N = 1$. It suggests that a larger N or M can improve the diverseness of input patterns. However, when $M > 2$ or $N > 2$, TASR increases first and then decreases. It demonstrates that excessive diversity can harm target transferability. This is be-

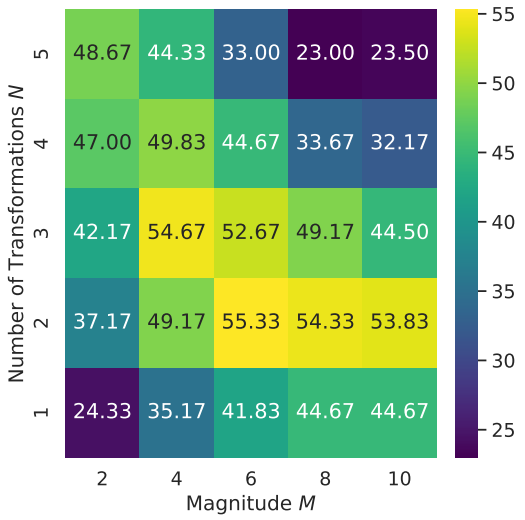


Figure F.1: TASR (%) of TH-Aug with different number of iterations N and magnitudes M.

cause optimizing perturbations is overwhelmed by excessive input patterns. Based on the reported TASR, we select $N = 5$, $M = 2$ for T-Aug. From Fig.E.1(b), we observe that the H-Aug with $N = 2, M = 8$ gains more improvement than others. We use these settings to conduct subsequent experiments.

F THE COMBINATION OF T-AUG AND H-AUG

Table F.1: TASR (%) of TH-Aug with the CE and Logit losses under ℓ_∞ norm with $\epsilon = 16/255$. The set of image transformations in TH-Aug is the combination of tiny and huge changed transformations. We show TASRs with 20/100/300 iterations. TM is the combination of TI and MI.

Attack	White-box Model: Res50			White-box Model: Dense121		
	→Dense121	→VGG16	→Inc-v3	→Res50	→VGG16	→Inc-v3
TH-Aug	12.4/73.5/88.6	10.4/71.6/89.1	0.6/12.2/17.6	8.1/51.0/70.7	9.0/64.8/80.3	0.4/8.9/13.8
TH-Aug-TM	8.6/65.2/83.9	8.3/62.7/82.9	1.7/16.8/38.6	6.7/47.6/67.2	7.0/53.5/71.2	1.8/17.2/30.0
Attack	White-box Model: VGG16			White-box Model: Inc-v3		
	→Res50	→Dense121	→Inc-v3	→Res50	→Dense121	→VGG16
TH-Aug	1.2/8.6/13.0	0.8/14.0/19.3	0.0/0.4/1.0	0.5/7.5/16.2	0.3/12.2/26.4	0.7/9.2/20.6
TH-Aug-TM	1.1/10.3/24.0	1.5/14.9/33.6	0.2/2.2/5.6	0.7/7.6/17.4	0.9/13.1/29.1	0.6/7.4/20.5

(a) Attacks with the **CE** loss

Attack	White-box Model: Res50			White-box Model: Dense121		
	→Dense121	→VGG16	→Inc-v3	→Res50	→VGG16	→Inc-v3
TH-Aug	11.1/66.2/82.3	9.0/62.9/81.4	0.9/12.6/23.4	7.3/50.5/68.2	8.7/58.6/74.1	0.7/11.1/18.5
TH-Aug-TM	7.7/56.6/78.0	6.9/55.0/76.9	1.5/17.2/38.6	6.3/43.4/68.0	7.2/49.1/69.0	1.0/17.4/34.1
Attack	White-box Model: VGG16			White-box Model: Inc-v3		
	→Res50	→Dense121	→Inc-v3	→Res50	→Dense121	→VGG16
TH-Aug	1.4/19.4/33.6	2.0/25.3/41.8	0.0/2.0/4.4	0.4/7.5/17.6	0.6/11.8/26.5	0.8/10.9/22.1
TH-Aug-TM	1.4/13.9/35.4	1.3/18.6/41.8	0.2/3.7/9.3	1.1/7.6/17.1	0.6/11.6/29.1	0.6/7.8/19.6

(b) Attacks with the **Logit** loss

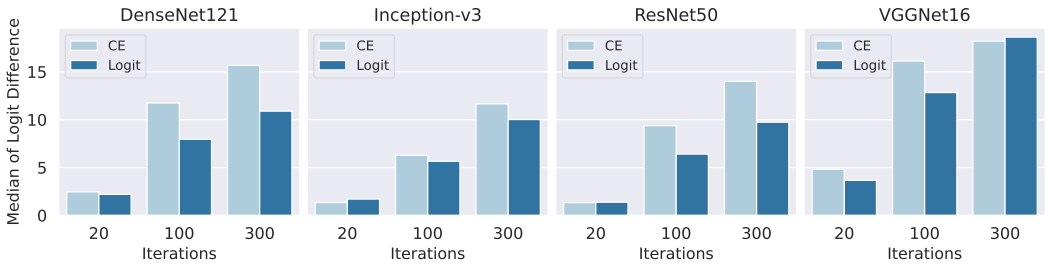
.

Let TH-Aug denote the combination of T-Aug and H-Aug. The set of image transformation T_{set} is the union of tiny and huge changed transformations. We set $N = 2, M = 6$ in TH-Aug, as

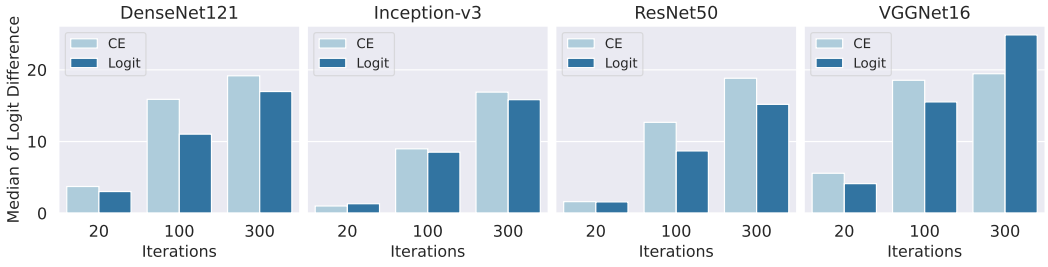
shown in Fig.F.1. Table F.1 shows the performance of TH-Aug in single-model transfer scenarios. As can be seen, TH-Aug performs worse than T-Aug and H-Aug. Suppose that the distortion of one image transformation of H-Aug is equivalent to the distortion of combining multiple image transformations of T-Aug Therefore, when selecting an image transformation from tiny and huge changed transformations respectively in one iteration, the image distortion of TH-Aug may be equal to that of T-Aug with larger values of N and M, resulting in lower performance, shown in the upper right of Fig.E.1(a).

G THE PROBLEM OF THE LOGIT LOSS

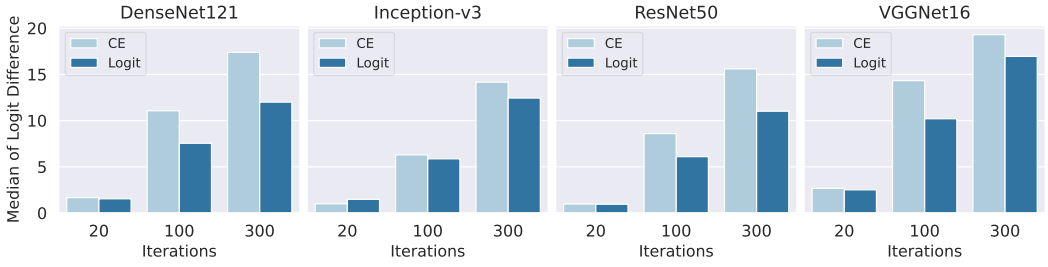
Despite the Logit loss avoids gradient vanishment in iterative attacks by increasing the logit value of the target class, it may also improve logit values of other classes. Therefore, we design a median logit difference metric to explore why Logit has lower performance than CE in some scenarios. Specifically, for an adversarial example, we calculate the difference between the logit value of the target class and the highest logit value of the remaining classes. Then we calculate the median logit difference among all adversarial examples. A higher median logit difference means that the target class is more dominant than other classes in models' predictions. As shown in Fig.G.1, the CE loss attains a higher median logit difference than the Logit loss when using DenseNet121 or ResNet50 as the white-box model. In contrast, when using Inception-v3 or VGGNet-16 as the white-box model, the CE loss and the Logit loss have similar values of the median logit difference. Based on this observation, we can obtain the conclusion that the Logit loss is not applicable in different models when combining with our data augmentation methods. The results of Table 1 also prove this point. One simple and possible solution is that increasing the logit value of the target class while decreasing the highest logit value of other classes, like the C&W attack (Carlini & Wagner, 2017). We leave it to future work.



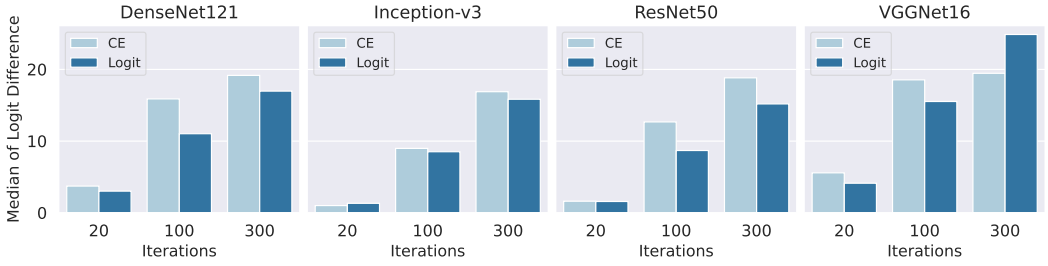
(a) T-Aug



(b) H-Aug



(c) T-Aug-TM



(d) H-Aug-TM

Figure G.1: The median value of the difference between the logit value of the target class and the highest logit value of the remaining classes on adversarial examples crafted by (a) T-Aug, (b) H-Aug, (c) T-Aug-TM and (d) T-Aug-TM. For each method, we compare the median logit different of CE loss with that of the Logit loss. The median logit differences with 20/100/300 iterations are reported. The whole ImageNet-compatible dataset is used here.