Enhancing the Self-Universality for Transferable Targeted Attacks

Zhipeng Wei[1,2], Jingjing Chen[1,2†], Zuxuan Wu[1,2], Yu-Gang Jiang[1,2]
[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing
zpwei21@m.fudan.edu.cn, {chenjingjing, zxwu, ygj}@fudan.edu.cn

Abstract

In this paper, we propose a novel transfer-based targeted attack method that optimizes the adversarial perturbations without any extra training efforts for auxiliary networks on training data. Our new attack method is proposed based on the observation that highly universal adversarial perturbations tend to be more transferable for targeted attacks. Therefore, we propose to make the perturbation to be agnostic to different local regions within one image, which we called as self-universality. Instead of optimizing the perturbations on different images, optimizing on different regions to achieve self-universality can get rid of using extra data. Specifically, we introduce a feature similarity loss that encourages the learned perturbations to be universal by maximizing the feature similarity between adversarial perturbed global images and randomly cropped local regions. With the feature similarity loss, our method makes the features from adversarial perturbations to be more dominant than that of benign images, hence improving targeted transferability. We name the proposed attack method as Self-Universality (SU) attack. Extensive experiments demonstrate that SU can achieve high success rates for transfer-based targeted attacks. On ImageNet-compatible dataset, SU yields an improvement of 12% compared with existing state-of-the-art methods. Code is available at https://github.com/zhipeng-wei/Self-Universality.

1. Introduction

It has been demonstrated in recent works that adversarial examples have the properties of transferability, which means an adversarial example generated on one white-box model can be used to fool other black-box models [3, 15, 27, 30, 33]. The existence of transferability brings convenience to performing black-box attacks, hence raising security concerns for deploying deep models in real-world applications [14, 22, 28, 35]. Consequently, considerable research attention has been spent on improving the transferability of adversarial examples for both non-targeted and targeted attacks [4, 29, 36].

Compared to non-targeted attacks, transfer-based targeted attacks are inherently much more challenging since the goal is to fool deep models into predicting the specific target class. The major difficulty of transfer-based targeted attacks is caused by the fact that the gradient directions from a source image to a target class are usually different among different DNNs [16]. Hence, transfer-based attack methods designed for non-targeted attacks typically work poorly for targeted attacks. To increase the transferability, previous studies make efforts in aligning the feature of the generated adversarial example with the feature distributions of the targeted class, which are learned from class-specific auxiliary networks [7, 8] or generative adversarial networks [18]. However, these works assume that the training dataset is available and require extra training efforts for auxiliary networks, making it hard to apply in real-world scenarios.

This paper investigates the problem of transfer-based targeted attacks. Specifically, we propose a new method that improves the transferability of adversarial examples in a more efficient way, i.e., without any training efforts for auxiliary networks to learn the feature distributions of the targeted class. Our method is proposed based on the observation that more universal perturbations yield better attack success rates in targeted attacks. To this end, our goal is to enhance the universality of the generated adversarial perturbations, in order to improve its targeted transferability. Note that existing universal adversarial perturbation (UAP) attacks [17] require optimizing the perturbations on an abundant of images to achieve universality, which is not applicable in our setting. To get rid of using extra data and make transfer-based targeted attacks as convenient as non-targeted attacks, we propose to make the perturbation to be agnostic to different local regions within one image, which we called as self-universality. Then our method optimizes the self-universality of adversarial pertur-

†Corresponding author.

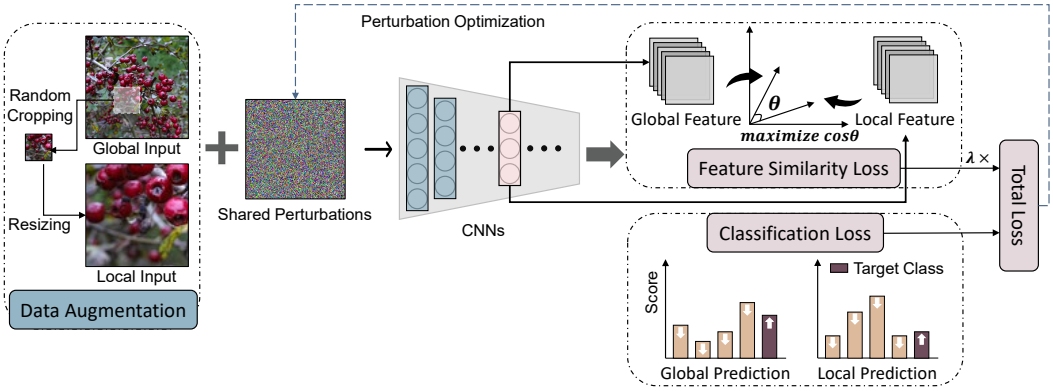


Figure 1. Overview of the proposed SU attack. The random cropping is applied to the given benign image to generate the local image patch. After cropping, the local patch is resized to the shape of the benign image. Then both benign and local adversarial images with the shared perturbations are input to a surrogate white-box CNN model. Finally, the gradients obtained from the classification loss and the feature similarity loss are used to optimize perturbations.

bations instead. To be specific, in addition to classification loss, our Self-Universality (SU) attack method introduces a feature similarity loss that maximizes the feature similarity between adversarial perturbed global images and randomly cropped local regions to achieve self-universality. In this way, our method makes the features from adversarial perturbations to be more dominant than that of benign images, hence improving targeted transferability.

Figure 1 gives an overview of the proposed Self-Universality (SU) attack. SU firstly applies random cropping on benign images to obtain local cropped patches. Then it resizes local patches to the same size with benign images. Consequently, global and local inputs with shared perturbations are input to the white-box model. Finally, adversarial perturbations are updated by minimizing the classification loss (*e.g.*, Cross Entropy) between inputs and the target class and maximizing the feature similarity loss (*e.g.*, Cosine Similarity) of adversarial intermediate features between local and global inputs. Benefiting from satisfying the prediction of the target class between global and local inputs and approximating adversarial intermediate features between the two, the proposed SU attack can generate perturbations with self-universality, thereby improving the cross-model targeted transferability. We briefly summarize our primary contributions as follows:

- Through experiments, we find that highly universal adversarial perturbations tend to be more transferable for targeted attacks, which brings new insight into the design of transfer-based targeted attack methods.
- Based on the finding, we propose a novel Self-Universality (SU) attack method that enhances the universality of adversarial perturbations for better targeted transferability without the requirement for extra data.
- We conduct comprehensive experiments to demonstrate that the proposed SU attack can significantly improve the cross-model targeted transferability of adversarial images. Notably, SU can be easily combined with other existing methods.

2. Related Work

In this section, we review existing works on non-targeted transferable attacks as well as targeted transferable attacks.

2.1. Non-targeted Transferable Attacks

Non-targeted transferable attacks are based on the Iterative-Fast Gradient Sign Method (I-FGSM) [11], which iteratively calculates the gradient of the classification loss with respect to the input and updates perturbations along the direction of maximizing the loss. The multiple-step update of I-FGSM leads to more wrong predictions of white-box models. However, the generated adversarial examples are hard to attack other black-box models. This phenomenon is called over-fitting to the white-box model [10]. Subsequently, several works perform data augmentation or advanced gradient calculation to overcome the over-fitting problem. Data augmentation creates various input patterns, which generate adversarial perturbations with a generic pattern. For the input images, Diverse Input (DI) [33] performs random resizing and padding, Scale-invariant method (SIM) [13] applies scale transformation, Admix [24] extends SIM by integrating images from other labels. However, these data augmentation approaches neglect to consider more diverse input patterns (*e.g.*, random cropping in this paper), and focus on the loss-preserving transformation [13]. This is because the gradient direction fluctuates violently under more diverse input patterns without specified target classes in non-target attacks. Advanced gradient calculation pays attention to changing gradient calculation methods or designing a new loss function. For changing

gradient calculation, the momentum term [3] and the Nesterov accelerated gradient [13] can be incorporated with I-FGSM to stabilize gradients. Translation-Invariant (TI) [4] smooths gradients through a predefined convolution kernel in order to mitigate the over-fitting problem. Skip Gradient Method (SGM) [31] modifies the path of gradient back-propagation by skipping residual modules. Variance Tuning [23] utilizes gradients from data points around previous data to avoid over-fitting. For designing a new loss function, Attention-guided Transfer Attack (ATA) [32] and Feature Importance-aware Attack (FIA) [26] disrupt important features that are likely to be utilized in other black-box models. ATA applies Grad-CAM [19] to represent attention weights of features. Different to ATA, FIA computes averaged gradients of features among various augmented inputs as attention weights. However, these new loss functions are restricted to global structures. In contrast, this paper calculates the feature similarity loss between global and local structures of images.

2.2. Targeted Transferable Attacks

The transfer-based targeted attacks are more challenging than the non-targeted attacks due to the requirement of fooling the model into predicting the specified target class. Previous efforts have been devoted into training class-specific auxiliary networks [7, 8] or generative adversarial networks (GANs) [18], in order to learn feature distributions of the targeted class. Feature Distribution Attack (FDA) [8] utilizes the intermediate features of the training dataset from the white-box model to train a binary classifier, which predicts the probability that the current feature belongs to the targeted class. Then, through learning class-wise and layer-wise feature distributions, FDA generates targeted transferable adversarial perturbations by maximizing the probability outputted from the trained binary classifier. Subsequently, $FDA^N + xent$ [7] incorporates the CE loss and multi-layer information with FDA, achieving better performance. Transferable targeted perturbations (TTP) [18] trains a generator to synthesize perturbations that seek to have similar features with targeted samples in the latent space of a pretrained discriminator. This series of FDA works and TTP improve the performance of targeted transferable attacks, but require additional training efforts for auxiliary networks on the training dataset.

To improve the efficiency of targeted transferable attacks, recent works pay more attention to iterative attacks based on I-FGSM. Activation Attack (AA) [9] replace the classification loss with a euclidean distance loss between features of adversarial and targeted images. However, AA depends on the selected examples of the targeted class and achieves low performance on the large resolution image dataset (*e.g.* ImageNet [2]). Different to AA, Po+Trip [12] utilizes Poincaré distance metric to solve the noise curing. They also design a triplet loss for driving adversarial examples away from the original class. However, Po+Trip evaluates the attack performance only on easy transfer scenarios, and also achieves poor performance on the targeted scenario used in the Logit attack [36]. Logit [36] attributes the poor performance of I-FGSM based methods to the restricted number of iterations. Thus, Logit enlarges the number of iterations for ensuring the convergence of attacks. In addition, to overcome the vanishing gradient caused by the large iterations, Logit utilizes gradients of the targeted logit output with respect to inputs to update perturbations. Different to them, our method leverages the global and local structures to generate adversarial examples with self-universality, hence improving targeted transferability. Note that the proposed SU attack can be easily combined with existing methods.

3. Methodology

3.1. Preliminary

Let f represent the white-box surrogate model, v be the black-box victim model, $x \in \mathcal{X} \subset \mathbf{R}^{H\times W\times C}$ be the benign image with the ground-truth label $y \in \mathcal{Y} = \{1, 2, ..., K\}$, where f and v are trained with the same training dataset, H, W, C denote the number of height, width and channels respectively, and K is the number of classes. We use $f(x), v(x)$ to be the prediction over the set of classes $\mathcal{Y}$. Given a specified targeted class y_t, the targeted transferable attacks aim to generate adversarial examples x_{adv} from the white-box model f that satisfy $v(x_{adv}) = y_t$. Following previous works, we enforce a L_∞-norm constraint on perturbations, which can be formulated as $||x_{adv} - x||_\infty = ||\delta||_\infty \leq \epsilon$, where δ is the perturbation, ϵ denotes a constant of the norm constraint. Given a classification loss J (*e.g.* CE loss) of f, the targeted attack of I-FGSM can be formulated as:

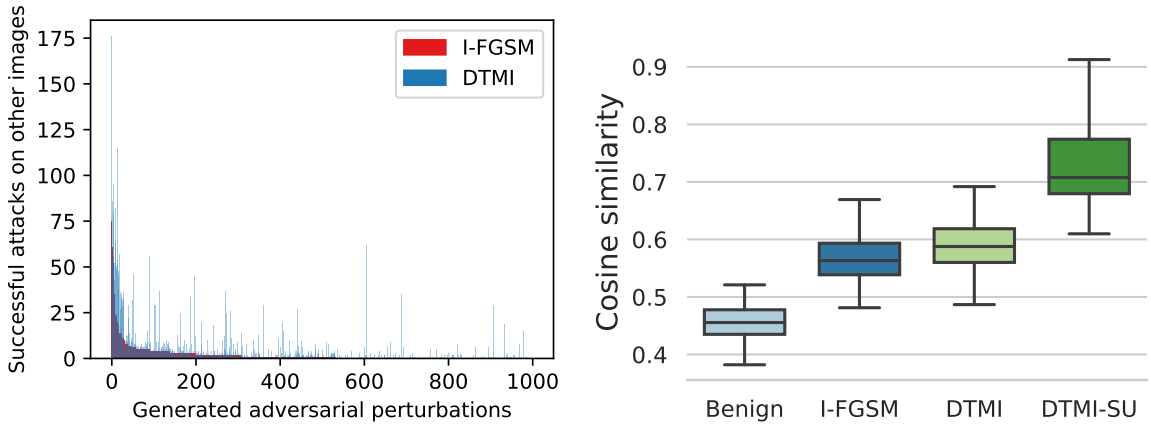
$$\delta_0 = 0, g_0 = 0, \tag{1}$$

$$g_{i+1} = \nabla_\delta J(f(x + \delta_i), y_t), \tag{2}$$

$$\delta_{i+1} = \delta_i - \alpha * sign(g_{i+1}), \tag{3}$$

$$\delta_{i+1} = Clip_{x,\epsilon}(\delta_{i+1}), \tag{4}$$

where δ_i, g_i denote the adversarial perturbation and the gradient of the i-th iteration, $i = [0, ..., I-1]$ and I is the maximum number of iterations; α is the step size; the function $Clip_{x,\epsilon}(\cdot)$ projects δ to the vicinity of x for satisfying the L_∞-norm constraint. Specifically, I-FGSM first initializes δ_0 and g_0 as 0 (Eq.1). Then it calculates the gradients of the loss function with respect to the perturbation (Eq.2), and drives the prediction of f on the adversarial example towards the targeted class by minimizing $J(f(x + \delta_i), y_t)$ (Eq.3). Finally, it restricts the adversarial perturbation with the L_∞-norm constraint (Eq.4). Eq.2, 3, 4 are performed alternately until the maximum iteration is reached. Through



(a) Universal evaluation

(b) Feature similarity analysis

Figure 2. (a) For each targeted perturbation generated by I-FGSM or DTMI, we calculate the number of successful targeted attacks caused by adding this perturbation to benign images other than the original optimized image. (b) We average the cosine similarity of intermediate features among different benign images or universal adversarial images that are generated by adding the same perturbation to these benign images.

this greedy update, I-FGSM is easily caught in local maxima and over-fitting to f, which results in performing poorly for targeted attacks. Therefore, previous work [36] utilizes the combination of DI [33], TI [4] and MI [3] as the baseline, since it could achieve more stable performances with a large number of iteration compared to I-FGSM. For convenience, we abbreviate the name of this combination as DTMI. DTMI replaces Eq.2 with

$$g_{i+1} = \mu \cdot g_i + \frac{W \cdot \nabla_\delta J(f(T(x+\delta_i, p)), y_t)}{||W \cdot \nabla_\delta J(f(T(x+\delta_i, p)), y_t)||_1}, \quad (5)$$

where $T(x+\delta_i, p)$ perform random resizing and padding with a probability p in DI, W is the pre-defined convolution kernel in TI, μ is a decay factor in MI.

3.2. Universality of Targeted Perturbations

Transferable targeted attacks aim to generate cross-model perturbations that drive the prediction of different models towards predicting the same specified class. It requires perturbations generated from one white-box model are model-agnostic, i.e., universal to different models. A simple way to meet this requirement is to optimize perturbations on massive models. However, it may be impractical in the real world, because the number of white-box models is usually limited. In contrast, targeted universal adversarial perturbations (UAP) [17] lead to a specific prediction y_t of DNNs for almost all images. [34] attributes this phenomenon to the dominant features produced by Targeted UAP. Specifically, let δ_u denote targeted UAP, they conclude that the predictive logit vectors of δ_u have a higher linear correlation with that of $x+\delta_u$, while the linear correlation between x and $x+\delta_u$ is low. It suggests that the features of δ_u represent the feature distribution of the specific targeted class and dominate the prediction of DNNs. Therefore, targeted UAP could transfer to attack other models trained on the same dataset, because these models share the similar feature distribution of the target class. In summary, universality may correlate to transferability in targeted attacks. Motivated by the above analysis, we delve into the divergence between perturbations crafted by I-FGSM and DTMI from the perspective of universality.

To further verify the findings in the above analysis, we conduct experiments to show the correlation between universality and targeted transferability. We optimize targeted perturbations on the ImageNet-compatible dataset (dataset details are in Section 4.1) by I-FGSM and DTMI using DenseNet121 as the white-box model. Here CE loss is utilized as classification loss. Given a set of benign images $\Phi = \{x^1, ..., x^m, ..., x^M\}$, we generate the perturbation for each image in the set. Denote δ^m as the generated perturbation for x^m. To evaluate the universality of δ^m, we add δ^m to all benign images other than x^m, and calculate the number of successful targeted attacks on DenseNet121 by:

$$\sum_{j=1, j\neq m}^{M} \begin{cases} 1, & \text{if } f(x^j + \delta^m) = y_t, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

The number of successful targeted attacks is used as the metric to evaluate the universality of δ^m. A higher number denotes better universality.

Figure 2a shows the comparison of universality between each perturbation crafted by I-FGSM and DTMI. For better visualization, we sort perturbations in descending order based on their universality reported in I-FGSM. Notably, DTMI attains higher universality than I-FGSM for each perturbation. This is because, aided by the diverse input pat-

terns provided by DI, DTMI makes the perturbations to be universal to these input patterns. Since DTMI achieves better targeted transferability than I-FGSM [36], we can basically draw a conclusion that there is a relatively positive correlation between universality and targeted transferability.

To provide insights on how the targeted perturbation affects the image features, we compare the average cosine similarity of intermediate features between benign images and adversarial images that generated by adding one specific targeted perturbation to these benign images. Specifically, these benign images can be denoted by Φ^m, which is the set of x^j satisfying $f(x^j + \delta^m) = y_t$. Adversarial images can be denoted by $\{x^j + \delta^m\}$ for $x_j \in \Phi^m$. Figure 2b shows the changes in the feature similarity after adding the same perturbation δ^m. It can be observed that after adding this perturbation, the image features become more similar. This is because the targeted perturbations drive features from different benign images towards the feature distribution of the target class. In other words, compared to benign images, targeted perturbations produce more dominant features. In addition, from Figure 2b, we can also find that compared to I-FGSM, targeted perturbations generated by DTMI make the image features more similar. Since DTMI achieves better targeted transferability than I-FGSM, the result basically indicates that perturbations with highly targeted transferability will produce more dominant features and enjoy higher universality. The above analysis suggests that adversarial examples with high universality tend to be more transferable in targeted attacks.

3.3. Self-Universality (SU) Attack

To enhance the universality of adversarial perturbations for transferable targeted attacks, we propose the Self-Universality (SU) attack method. Instead of optimizing the perturbations on different images to achieve universality, our SU method optimizes the perturbation to be agnostic to different local regions within one image, which is called self-universality. In this way, our method is able to generate the universal perturbations without extra data. To achieve self-universality, our method generates perturbations by incorporating randomly cropped local regions within one image into the iterative attacks. To create local input patterns, we consider the random cropping, which crops images into a local image patch by the scale parameter $s = \{s_l, s_{int}\}$. Where s_l denotes the lower bound for the area of the random cropped images, and s_{int} is the interval value between the lower and upper bounds, thus $s_l + s_{int}$ is the upper bound. After cropping images, we resize them to the same shape as benign images. Let $Loc(x, s)$ be the random cropping and the resizing operations*. In addition, motivated by Figure 2b, we propose a feature similarity loss to maximize the cosine similarity of intermediate features between the adversarial global and local inputs. Under this design, SU could generate perturbations with more dominant features. This loss can be formulated by $CS(f_l(x+\delta_i), f_l(Loc(x, s)+\delta_i))$, where $f_l(\cdot)$ means to extract features from the l-th layer of the white-box model f, the function $CS(\cdot, \cdot)$ calculates cosine similarity score of features between adversarial global and local inputs. In summary, the proposed SU replaces Eq.2 with:

$$\begin{aligned} g_{i+1} =& \nabla_\delta(J(f(x+\delta_i), y_t) + J(f(Loc(x, s)+\delta_i), y_t) \\ & - \lambda \cdot CS(f_l(x+\delta_i), f_l(Loc(x, s)+\delta_i))), \end{aligned} \tag{7}$$

where λ weights the contribution of the classification loss and the proposed similarity loss.

In this way, the perturbation optimization of SU can improve the self-universality of perturbations, and the maximization of the cosine similarity can align intermediate features between adversarial global and local inputs for improving the dominance of feature representation of perturbations. Following [36], we integrate SU with the baseline DTMI to further boost targeted transferability, which we named DTMI-SU. We show DTMI-SU in Algorithm 1. In the end, SU can generate adversarial perturbations with high dominant features by incorporating local images, as shown in Figure 2b, hence improving targeted transferability.

Algorithm 1 DTMI-SU attack

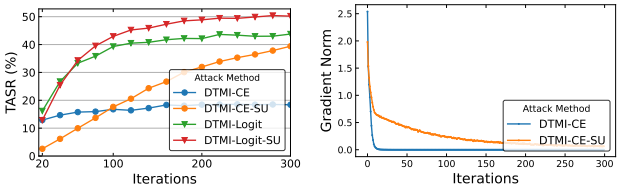
Input: the classification loss function J, white-box model f, benign image x, targeted class y_t.
Parameter: The perturbation budget ϵ, iteration number I, step size α, scale parameter $s = \{s_l, s_{int}\}$, weighted parameter λ, and DTMI parameters $T(\cdot, p)$, W, μ.
Output: The adversarial example x_{adv}.

Initialize δ_0 and g_0 by Eq.1
for $i = 0$ to $I - 1$ **do**
  Random cropping and resizing: $\hat{x} = Loc(x, s)$
  DI: $x' = T(x + \delta_i, p)$; $\hat{x}' = T(\hat{x} + \delta_i, p)$
  Calculate gradients: $g_{i+1} = \nabla_\delta(J(f(x'), y_t) + J(f(\hat{x}'), y_t) - \lambda \cdot CS(f_l(x'), f_l(\hat{x}')))$
  $g_{i+1} = \mu \cdot g_i + \frac{W \cdot g_{i+1}}{||W \cdot g_{i+1}||_1}$
  Update and Clip δ_{i+1} by Eq.3, 4
end for
return $x + \delta_I$

*We perform it by RandomResizedCrop in torchvision and ignore the parameter of the random aspect ratio.

Model	Layer 1	Layer 2	Layer 3	Layer 4
Res50	Block-1	Block-2	Block-3	Block-4
Den121	Block-1	Block-2	Block-3	Block-4
VGG16	ReLU-4	ReLU-7	ReLU-10	ReLU-13
Inc-v3	Conv-4a	Mix-5d	Mix-6e	Mix-7c

Table 1. Intermediate layers for feature extraction. "Block" means the basic block in ResNet50 and DenseNet121. "ReLU-k" denotes the k-th ReLU layer.



(a) TASR (%) (b) Gradient norm

Figure 3. (a) TASR (%) of attacking ResNet50 from DenseNet121. (b) Gradient norm of performing DTMI-CE and DTMI-CE-SU when using DenseNet121 as the surrogate model.

4. Experiment

4.1. Experimental Settings

4.1.1 Dataset and models

Following [36], we attack four diverse classifier architectures: ResNet50 [5], DenseNet121 [6], VGGNet16 [20], and Inception-v3 [21] with the ImageNet-compatible dataset[†]. The architectures of these four models are more diverse and challenging than the ones used in [12]. The used dataset is first introduced by the NIPS 2017 Competition on Adversarial Attacks and Defenses. It consists of 1,000 images and corresponding labels for targeted attacks.

4.1.2 Attack setting

We use the Targeted Attack Success Rate (TASR) to evaluate the targeted attack on one black-box model, which is the rate of adversarial examples that are successfully classified as the targeted class by the black-box model. Hence, the methods with higher TASR can generate adversarial examples with high targeted transferability. Following [36], we set the maximum perturbation as $\epsilon = 16$, the step size as $\alpha = 2$, the maximum number of iterations as $I = 300$, For each model, we select one layer 1/2/3/4 from shallow to deep layers (as shown in Table 1) to extract features.

4.2. Performance Comparison

Following [36], we adopt the combination of DI, TI and MI (DTMI) as the baseline. The parameters of DTMI are the same as [36]. We then integrate the proposed SU method with several baselines: CE loss with DTMI (DTMI-CE) and Logit loss [36] with DTMI (DTMI-Logit). We exclude Po+Trip [12] in the comparison because it has a worse performance than the Logit loss.

For SU, the weighted parameter λ is set as 10^{-3}, the scale parameters s is set as $(0.1, 0)$, and the layer 3 is used to extract features. These parameters will be discussed in Section 4.3.

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

Single-model transferable attacks. We conduct single-model transferable experiments by selecting one model as the white-box model to attack the other three black-box models. Table 2 shows the results. We have the following observations. First, the proposed SU method achieves higher TASR than DTMI-CE and DTMI-Logit by a large margin in almost all cases. It suggests that the proposed SU is suitable for different classification loss functions. Second, the attacks using ResNet50 and DenseNet121 as white-box models outperform that using VGGNet16 and Inception-v3, which is also observed in [8, 9, 36]. It may be caused by the skip connection structure, which mitigates the gradient vanishing/exploding problem in a large number of iterations and emphasizes features of lower layers by backpropagating gradients on both residual modules and skip connections. These features tend to transfer among different models [30, 31]. Third, SU performs worse than other methods when using a small number of iterations (*e.g.* $I = 20$). This is because SU requires a large iteration for convergence, illustrated in Figure 3a. Especially for the CE loss, DTMI achieves the convergence at $I = 20$, while DTMI-SU steadily improves performance as the iteration increases, since SU can eliminate the vanishing gradient caused by CE, as shown in 3b. We also perform experiments with a smaller perturbation budget $\epsilon = 8$, which are shown in Appendix. It follows a similar trend as Table 2. These results jointly demonstrate the effectiveness of SU. In addition, we visualize adversarial examples in Appendix. These adversarial examples, which humans can correctly understand, lead to the specific target prediction of DNNs.

Ensemble model transferable attacks. As mentioned above, the targeted perturbations generated from a set of models are more model-agnostic, and thus can obtain high cross-model transferability. Hence, we select one black-box model to attack, and use the other three models as white-box models. Following [36], we simply assign equal weights to all white-box models. Results are shown in Table 3. As can be seen, compared to single-model transferable attacks, ensemble attacks achieve much better performance. It demonstrates that the model-agnostic perturbations can easily transfer to attack other models. In addi-

Attack	White-box Model: Res50			White-box Model: Dense121		
	→ Dense121	→ VGG16	→ Inc-v3	→ Res50	→ VGG16	→ Inc-v3
DTMI-CE	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
DTMI-CE-SU	6.2/27.8/**54.2**	3.0/20.2/**45.4**	0.2/4.5/**10.1**	2.6/17.6/**39.4**	1.4/12.6/**32.4**	0.2/4.8/**10.8**
DTMI-Logit	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
DTMI-Logit-SU	23.8/63.9/**75.5**	16.6/55.9/**66.9**	2.0/8.3/**11.6**	12.8/42.9/**50.2**	9.3/37.2/**45.2**	1.8/7.5/**10.4**
Attack	White-box Model: VGG16			White-box Model: Inc-v3		
	→ Res50	→ Dense121	→ Inc-v3	→ Res50	→ Dense121	→ VGG16
DTMI-CE	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
DTMI-CE-SU	0.2/2.1/**2.8**	0.2/2.1/**3.2**	0.0/0.2/**0.2**	0.4/1.2/**2.9**	0.2/1.4/**5.0**	0.1/0.8/**2.5**
DTMI-Logit	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
DTMI-Logit-SU	3.4/11.7/**13.9**	3.3/13.8/**16.1**	0.2/0.9/**0.8**	0.6/2.3/**4.3**	0.5/3.8/**7.4**	0.3/1.9/**4.4**

Table 2. TASR (%) of all black-box models under four attack scenarios using ResNet50, DenseNet121, VGGNet16 and Inception-v3 as white-box models, respectively. We conduct these experiments three times and report average TASR with 20/100/300 iterations, the standard deviation is shown in Appendix. The best results with 300 iterations are in bold.

Ensemble Attack	Black-box Model				Average
	Res50	Dense121	VGG16	Inc-v3	
DTMI-CE	31.1	55.2	51.6	16.1	38.5
DTMI-CE-SU	**55.7**	**65.0**	**68.2**	**29.3**	**54.5**
DTMI-Logit	70.2	82.3	82.2	29.1	65.9
DTMI-Logit-SU	**75.3**	**82.9**	**84.2**	**34.5**	**69.2**

Table 3. TASR (%) of one black-box model in ensemble transfer attacks. TASR with 300 iterations is reported. The best results are in bold.

Attack	Dense121	VGG16	Inc-v3
DTMI-SI/+SU	85.7/**87.2**	69.0/**71.8**	35.8/**41.6**
DTMI-Adm./+SU	89.1/**89.4**	75.7/**79.1**	42.1/**47.1**
DTMI-EMI/+SU	71.0/**79.0**	64.6/**82.4**	5.0/**14.8**
ODI-TMI/+SU	89.9/**92.8**	81.0/**91.7**	66.9/**72.0**

Table 4. Average TASR (%) of different combinational attacks. We use ResNet50 as the white-box model, and report the results with 300 iterations. Logit is used here.

tion, the proposed SU further boosts cross-model targeted transferability. However, the improvement of TASR gained from DTMI-Logit-SU is not so significant. This can be explained by the fact that the loss function value of Logit increases infinitely, while the value of cosine similarity is bounded. Thus, it impedes the maximization of feature similarity. Nevertheless, methods that are combined with SU can promote targeted transferability in all cases.

Combination with existing methods To further boost transferable targeted attacks, we combine our method with other attacks, including SI [13], Admix [24], EMI [25], and ODI [1]. Table 4 reports the results. The attack methods combined with the proposed SU achieve better performance than other attack methods. In particular, compared to ODI, SU achieves an average TASR gain of 6%. Such remarkable improvements demonstrate that the proposed SU can be easily combined with other existing methods and further improve targeted transferability. In terms of computational efficiency, SU only requires two forward propagations for global and local inputs per iteration, which is lower than SI, Admix, and EMI (which require ≥ 5 forward propagations). To compare SU with ODI, we report the computation time (sec) required to generate an adversarial example. DTMI-SU and DTMI-ODI require 2.3 and 2.6 sec, respectively. It indicates that SU requires less computation time than previous methods.

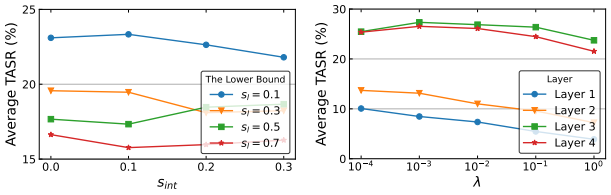
4.3. Ablation Studies

In this section, we utilize CE as the classification loss to investigate the effects of each component and various hyper-parameters. In addition, we investigate the effects of different regions of random cropping.

Effect of components. The evaluations are conducted on four attack scenarios using ResNet50, DenseNet121, VGGNet16 and Inceptionv3 as white-box models, respec-

Local	Feature Similarity Loss	Averaged TASR
-	-	9.8
✓	-	10.9
✓	✓	15.6

Table 5. Average TASR (%) of black-box models for our proposed method with different component combinations. The classification loss is set as CE. '✓' indicates that the component is used while '-' indicates that it is not used. TASR is averaged among four attack scenarios using ResNet50, DenseNet121, VGGNet16 and Inceptionv3 as white-box models, respectively.



(a) Scale parameter (s_l, s_{int}) (b) Layer l and weight λ

Figure 4. Average TASR with different values of hyper-parameters in the attack. The classification loss is set as CE.

tively. We split SU into the local inputs and the feature similarity loss. Table 5 shows the results of different component combinations. As can be seen, both components improve targeted transferability. And the improvement of TASR gained from feature similarity loss is more significant. It demonstrates the effectiveness of each component in the proposed SU attack.

Effect of hyper-parameters. The evaluations are conducted by using densenet121 as the white-box model and averaging TASR among black-box models. The scale parameter s determines the area of cropped images. When the area is large, SU overlooks the local structure of inputs. Thus, it is vital to study optimal values of s. We fix the weighted parameter $\lambda = 1.0$, and the feature extraction layer as layer 3 due to the high transferability of middle layers [30]. Figure 4a shows the results with $s_l \in [0.1, 0.3, 0.5, 0.7]$ and $s_{int} \in [0.0, 0.1, 0.2, 0.3]$. We observe that the smaller the area of cropped images, the higher the performance. It suggests that local image patches that differ significantly from the original images provide more diverse input patterns for perturbation optimization and contribute to targeted transferability. Besides, when $s_l = 0.1$ and $s_{int} = 0.1$, the optimal result is reached. However, the area relative to the original image may fluctuate within $[0.1, 0.2]$. To keep it stable, we use the parameters $s = (0.1, 0.0)$ to conduct subsequent experiments. Figure 4b shows the results of performing attacks on different layers and λ. Extracting features from layer 3 is better than other layers. This is because the features captured by shallow layers represent the low-level patterns (such as edges),

Region	Res50	VGG16	Inc-v3
Center	33.0	29.2	9.9
Corner	33.6	30.9	10.8
Whole	39.4	32.4	10.8
Uniform	3.1	1.7	0.4

Table 6. TASR (%) of DTMI-CE-SU with different distributions of local images. The top three rows show different regions of random cropping. The last row denotes the uniform distribution $\mathcal{U}(0, 1)$ of local images.

while the deepest layer contains more semantic information hence is more related to the classification task of the white-box model [30]. For the weighted parameter λ, we set $\lambda \in [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}]$. When $\lambda = 10^{-3}$, SU achieves the best result.

Effect of different regions. The regions of random cropping determine the distribution of local images. The center may be more related to the object of the image than the corner. Therefore, we conduct experiments on the center, the corner and the whole regions. We also compare the uniform distribution $\mathcal{U}(0, 1)$ of local images. As shown in Table 6, there is no significant performance difference between the center and the corner. The reason is that SU concentrates on the target class regardless of the object related to the original class. Therefore, directly providing more local input patterns within one image leads to better performance in SU. The whole region indeed achieves the best performance. Moreover, the uniform distribution makes it difficult to converge, resulting in the worst performance.

5. Conclusion

In this paper, we provide new insight into transfer-based targeted attacks: more universal perturbations yield better transferability. Based on this observation, we propose the Self-University (SU) attack, which optimizes perturbations on the global image and more diverse local images, and aligns intermediate features between them. In this way, SU can make perturbations to be agnostic to different image regions, resulting in high self-transferability. We conduct extensive experiments to show that SU can improve targeted transferability regardless of the single-model attack or ensemble-based attack.

6. Acknowledgments

This project was supported by National Key R&D Program of China (No. 2020AAA0140001), NSFC (No. 62072116), and Science and Technology Commission of Shanghai Municipality (No. 21JC1400600).

References

[1] Junyoung Byun, Seungju Cho, Myung-Joon Kwon, Hee-Seon Kim, and Changick 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*, pages 15244–15253, 2022. 7

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[3] 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*, pages 9185–9193, 2018. 1, 3, 4

[4] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4307–4316, 2019. 1, 3, 4

[5] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[6] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. 6

[7] 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, 2020. 1, 3

[8] Nathan Inkawhich, Kevin J Liang, Lawrence Carin, and Yiran Chen. Transferable perturbations of deep feature distributions. *arXiv preprint arXiv:2004.12519*, 2020. 1, 3, 6

[9] 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*, pages 7066–7074, 2019. 3, 6

[10] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *ArXiv*, abs/1611.01236, 2017. 2

[11] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 2

[12] 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)*, pages 638–646, 2020. 3, 6

[13] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv: Learning*, 2020. 2, 3, 7

[14] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In *Proceedings of the AAAI conference on artificial intelligence*, 2019. 1

[15] Aishan Liu, Jiakai Wang, Xianglong Liu, Bowen Cao, Chongzhi Zhang, and Hang Yu. Bias-based universal adversarial patch attack for automatic check-out. In *ECCV*, 2020. 1

[16] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017. 1

[17] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 86–94, 2017. 1, 4

[18] 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*, pages 7708–7717, 2021. 1, 3

[19] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359, 2017. 3

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015. 6

[21] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 6

[22] Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan Yuille, et al. Robustart: Benchmarking robustness on architecture design and training techniques. *arXiv preprint arXiv:2109.05211*, 2021. 1

[23] 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)*, pages 1924–1933, 2021. 3

[24] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16158–16167, 2021. 2, 7

[25] Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. Boosting adversarial transferability through enhanced momentum. *arXiv preprint arXiv:2103.10609*, 2021. 7

[26] 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)*, pages 7619–7628, 2021. 3

[27] Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards transferable

adversarial attacks on vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2668–2676, 2022. 1

[28] Zhipeng Wei, Jingjing Chen, Xingxing Wei, Linxi Jiang, Tat-Seng Chua, Fengfeng Zhou, and Yu-Gang Jiang. Heuristic black-box adversarial attacks on video recognition models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12338–12345, 2020. 1

[29] Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Boosting the transferability of video adversarial examples via temporal translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2659–2667, 2022. 1

[30] Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Cross-modal transferable adversarial attacks from images to videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15064–15073, 2022. 1, 6, 8

[31] 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, 2020. 3, 6

[32] 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)*, pages 1158–1167, 2020. 3

[33] 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*, pages 2730–2739, 2019. 1, 2, 4

[34] Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In So Kweon. Understanding adversarial examples from the mutual influence of images and perturbations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14509–14518, 2020. 4

[35] Chongzhi Zhang, Aishan Liu, Xianglong Liu, Yitao Xu, Hang Yu, Yuqing Ma, and Tianlin Li. Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. *IEEE Transactions on Image Processing*, 2021. 1

[36] 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. 1, 3, 4, 5, 6