

Everywhere Attack: Attacking Locally and Globally to Boost Targeted Transferability

Hui Zeng^{1,2*}, Sanshuai Cui^{3*}, Biwei Chen^{4†}, Anjie Peng¹

¹Southwest university of science and technology, Mianyang, China

²Guangan institute of technology, Guangan, China

³City University of Macau, Macau, China

⁴Beijing normal university, Zhuhai, China

zengh5@mail2.sysu.edu.cn, sanshuaicui@cityu.edu.mo, bchen@bnu.edu.cn, penganjie200012@163.com,

Abstract

Adversarial examples' (AE) transferability refers to the phenomenon that AEs crafted with one surrogate model can also fool other models. Notwithstanding remarkable progress in untargeted transferability, its targeted counterpart remains challenging. This paper proposes an *everywhere* scheme to boost targeted transferability. Our idea is to attack a victim image both globally and locally. We aim to optimize 'an army of targets' in every local image region instead of the previous works that optimize a high-confidence target in the image. Specifically, we split a victim image into non-overlap blocks and jointly mount a targeted attack on each block. Such a strategy mitigates transfer failures caused by attention inconsistency between surrogate and victim models and thus results in stronger transferability. Our approach is method-agnostic, which means it can be easily combined with existing transferable attacks for even higher transferability. Extensive experiments on ImageNet demonstrate that the proposed approach universally improves the state-of-the-art targeted attacks by a clear margin, e.g., the transferability of the widely adopted Logit attack can be improved by 28.8%~300%. We also evaluate the crafted AEs on a real-world platform: Google Cloud Vision. Results further support the superiority of the proposed method.

Code — https://github.com/zengh5/Everywhere_Attack

Introduction

Adversarial example (AE) (Szegedy et al. 2014) is a powerful tool for uncovering potential vulnerability of deep neural networks (DNN) before their deployment in security-sensitive applications (Madry et al. 2018). An exciting property of the AE is that AEs crafted against one model have a non-negligible chance to fool unseen victim models, a.k.a., transferability. Numerous transferable attacks have emerged recently, e.g., stabilizing the optimization direction (Dong et al. 2018; Lin et al. 2020; Wan, Ye, and Huang 2021) or diversifying inputs and surrogates (Xie et al. 2019; Dong et al. 2019; Wang et al. 2021; Li et al. 2020b; Fan et al. 2023).

*These authors contributed equally.

†Corresponding author.



Figure 1: Illustration of the proposed *everywhere* attack. We attempt to synthesize an army of Wukongs (target, the monkey) into every local region of Bajie (victim, the pig).

Despite extensive studies constantly refreshing transferability under the untargeted mode, targeted transferability is much more daunting since it requires unknown models outputting a specific label (Liu et al. 2017). To bridge the gulf, tailored schemes for improving the transferability of targeted attacks have been proposed. For instance, resource-intensive attacks seek extra, target-specific classifiers (Inkawhich et al. 2020) or generators (Naseer et al. 2021; Yang et al. 2022) to optimize adversarial perturbations. Other researchers find that integrating novel loss functions with conventional simple iterative attacks can also enhance targeted transferability (Li et al. 2020a; Zhao, Liu, and Larson 2021; Zeng et al. 2023; Weng et al. 2023).

Despite the recent progress of targeted attacks, the reported transferability is still unsatisfactory. Unlike the attention regions (to the ground truth class) that are critical to untargeted attacks, which tend to overlap among diverse models (Wu et al. 2020), the target class-related attention regions vary significantly across different models (refer to Figure 2), resulting in limited targeted transferability. This paper proposes an *everywhere* scheme to alleviate the attention mismatch dilemma for targeted attacks. Our idea is illustrated in Figure 1: Bajie (the pig) is expected to be attacked as Wukong (the monkey)¹. In contrast to conventional attacks that try to plant a high-confidence Wukong into the victim im-

¹Wukong and Bajie are the two main characters in the Chinese classical novel Journey to the West. Wukong can use his own hair to transform into a large number of clones, thus gaining an advantage in fighting.

age, the proposed *everywhere* attack simultaneously plants an army of Wukong in every local region of the victim image, with the hope that at least one Wukong falls into the attention area of the victim model. Our contributions can be summarized as follows.

- We note that a common cause of targeted transfer failure is the attention mismatch between the surrogate and victim models.
- With this challenge in mind, we propose an *everywhere* attack that tries to cover as much as possible the attention areas of various victim models. To our knowledge, this is the first attempt to enhance transferability by increasing the number of target objects, as opposed to previous works that aim to increase the confidence of the target class object.
- Extensive experiments demonstrate that the proposed method possesses good extensibility and can improve almost all state-of-the-art targeted attacks by a clear margin.

Related Work

An adversarial attack typically has two modes: targeted and untargeted. A targeted attack misguides a classification model to produce an adversary-desired label, whereas an untargeted attack only fools it for misclassification. Targeted attacks are strictly more difficult yet pose a more severe threat to the classification model. In this section, we briefly review conventional tricks to improve untargeted transferability and then discuss tailored schemes for targeted transferability.

Transferable Untargeted Attacks

A plethora of transferable attacks is built up on the well-known iterative fast gradient sign method (IFGSM) (Kurakin, Goodfellow, and Bengio 2016), which can be formulated as:

$$\mathbf{I}'_{n+1} = \text{Clip}_{\mathbf{I}, \epsilon}(\mathbf{I}'_n + \alpha \text{sign}(\nabla_{\mathbf{I}'_n} J(\mathbf{I}'_n, y_o))) \quad (1)$$

where $\mathbf{I}'_0 = \mathbf{I}$, $\nabla_{\mathbf{I}'_n} J(\cdot)$ denotes the gradient of the loss function $J(\cdot)$ with respect to \mathbf{I}'_n , y_o is the original label, and ϵ is the perturbation budget. Researchers have proposed a variety of improved algorithms over IFGSM, e.g., the momentum iterative method (MI) (Dong et al. 2018) integrates a momentum term into the iterative process. Diverse inputs method (DI) (Xie et al. 2019) and translation-invariant method (TI) (Dong et al. 2019) leverage data augmentation to prevent attacks from overfitting a specific source model. Moreover, these enhanced schemes can be integrated for better transferability, e.g., Translation Invariant Momentum Diverse Inputs IFGSM (TMDI).

Transferable Targeted Attacks

In addition to the difficulties untargeted attacks face, targeted attacks have their own challenges, such as gradient vanishing (Li et al. 2020a; Zhao, Liu, and Larson 2021) and the restoring effect (Li et al. 2020a; Zeng et al. 2023). Hence, tailored considerations are necessary for transferable targeted attacks. Existing efforts to boost targeted transferability

can be divided into two families: resource-intensive methods and simple-gradient methods.

Resource-intensive attacks require training auxiliary target-class-specific classifiers or generators on additional data. In the feature distribution attack (Inkawich et al. 2020), a light-weight, one-versus-all classifier is trained for each target class y_t at each specific layer to predict the probability that a feature map is from y_t . Transferable targeted perturbation (TTP) (Naseer et al. 2021) trains an input-adaptive generator to synthesize targeted perturbation and achieves state-of-the-art transferability. However, a dedicated generator must be learned for every (*source model*, *target class*) pair in TTP. Such a limitation is partially addressed by training a conditional generator (Mirza and Osindero 2014) to target multi-class simultaneously (C-GSP, LFAA) (Yang et al. 2022; Wang, Shi, and Wang 2023). However, the number of targeted labels a single generator can cover is limited due to its limited representative capacity. As a consequence, when the number of targeted classes is enormous, e.g., ImageNet, the required training time and storage are still prohibitive.

On the other hand, **simple-gradient methods** only iteratively optimize a victim image and thus have received more attention. Po+Trip attack (Li et al. 2020a) replaces traditional cross-entropy (CE) loss with the Poincare distance loss to address the decreasing gradient problem and introduces a triplet loss to push the attacked image away from y_o . Logit attack (Zhao, Liu, and Larson 2021) uses the Logit loss in the attack and reports better transferability than the CE loss.

$$L_{\text{Logit}} = -l_t(\mathbf{I}') \quad (2)$$

where $l_t(\cdot)$ denotes the logit output with respect to y_t . Moreover, Zhao, Liu and Larson (2021) point out that targeted attacks need significantly more iterations to converge than untargeted ones do. Similarly, Weng et al. (2023) (Margin) point out that the vanishing of the logit margin between the targeted and untargeted classes limits targeted transferability. Thus, they downscale the logits with a temperature factor to address the saturation issue and achieve improved transferability. The object-based diverse input method (ODI) (Byun et al. 2022) proposes diversifying the input image in a 3D object manner to avoid overfitting the source model and achieve improved targeted transferability. The high-confidence label suppressing method (SH) (Zeng et al. 2023) argues that not only the original label y_o , but other high-confidence labels should also be suppressed for better transferability. Such an idea can be realized by updating AEs according to the following direction:

$$\nabla(l_t(\mathbf{I}') - \beta_1 l_o(\mathbf{I}')) - \beta_2 \nabla(\sum_{i=0}^{N_b} l_{\text{high-conf}, i}(\mathbf{I}')) \perp \quad (3)$$

where \perp denotes retaining only the component perpendicular to the first item. Here, the first term is used to enhance the confidence of y_t and suppress y_o simultaneously, the second term suppresses other high-confidence labels. Based on the observation that highly universal adversarial perturbations tend to be more transferable, the self-universality method (SU) (Wei et al. 2023) introduces a feature similarity loss to encourage the adversarial perturbation to be self-universal. The clean feature mixup method (CFM) (Byun et al. 2023)

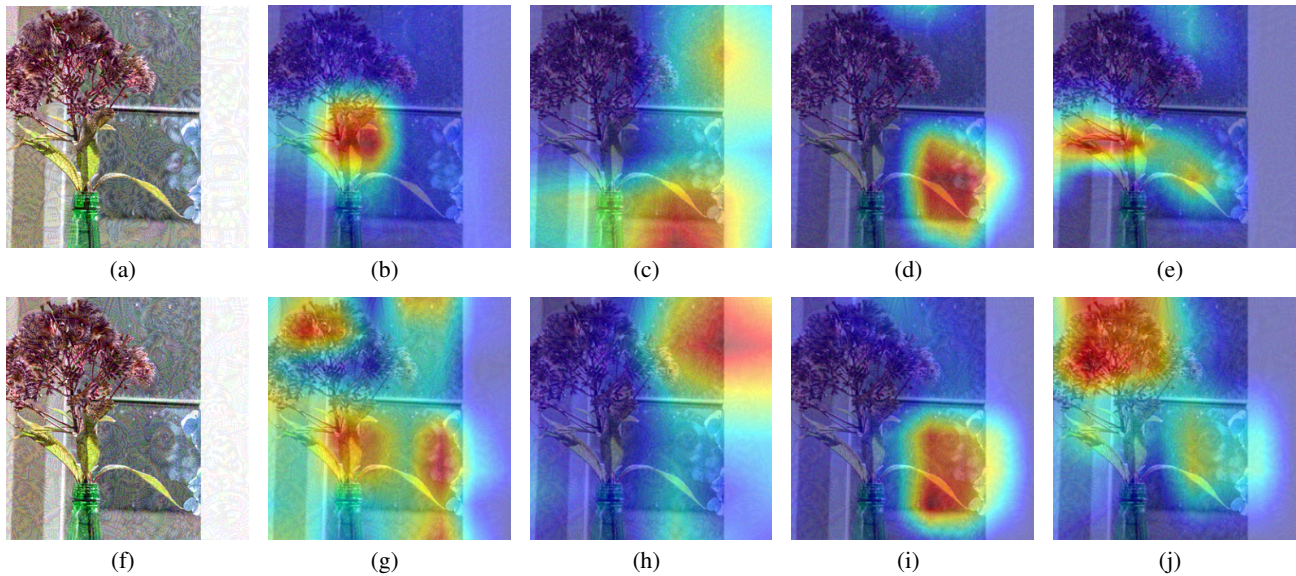


Figure 2: Attentional maps of the target label (*'marmoset'*) on different models. The top row depicts the results of the vanilla CE attack, and the bottom that of the proposed CE+*everywhere* attack. (a, f) Crafted AEs, (b, g) VGG16 (surrogate), (c, h) Inceptionv3 (Inc-v3) (Szegedy et al. 2016), (d, i) Res50, (e, j) Dense121.

borrowed the idea from Admix (Wang et al. 2021) to intentionally introduce competitor noises during optimization, which is achieved by mixing up features from other images in the same batch. Strictly speaking, CFM does not belong to simple-gradient methods since additional images are involved in the optimization. Nevertheless, it exhibits outstanding attack ability according to our experiments.

The Proposed Method

This section revisits a common cause for targeted transfer failure and details the proposed everywhere attack, which can effectively alleviate the attention mismatch issue.

Motivation

In a targeted attack, the adversary attempts to plant a quasi-imperceptible target object (or objects) into a clean image. Due to the attentional mechanism of DNNs, such a planting often focuses on specific image regions. To achieve transferable attacks across victim models, one may expect victim models to center on regions similar to the surrogate model in identifying the target object (or objects). In fact, this assumption is difficult to satisfy in a targeted attack. To illustrate this dilemma, we examine the attentional maps of an AE on different models. The attentional maps are computed with GradCAM (Selvaraju et al. 2017). The AE shown in Figure 2(a) is crafted with the vanilla CE attack, the surrogate model is VGG16bn (VGG16) (Simonyan and Zisserman 2015), and the target label is *'marmoset'*. As can be observed from Figure 2(b), the attack focuses on the lower area of the flower crown. One can imagine that the adversary has planted a *'marmoset'* in this region. However, victim models pay attention to strikingly different regions in recognizing a *'marmoset'*. For example, ResNet50 (Res50) (He et al.

2016) tries to find a *'marmoset'* from the lower right area of the image (Figure 2(d)). As a result, such a transfer attack fails on all three victim models.

One possible way to address the abovementioned challenge is to draw the victim model's attention to the attacked region. However, this is not easy because the adversary in the transfer attack setting cannot access the victim model's attentional region. Since the victim model's attention is unknown in advance, an intuitive strategy is to craft a *bunch of targets* in every region that the victim model may pay attention to. Such a conceptually simple idea motivates the proposed *everywhere* attack.

Everywhere Attack

Figure 3 gives an overview of the proposed *everywhere* attack. To synthesize targets in multiple regions of the image, we split a victim image into $M \times M$ non-overlap blocks. Then, we randomly sample N blocks from the image. For each sampled block, we pad the remaining area with the mean value of the dataset (which will be normalized to zero) and get a 'local' image. Concatenating these 'local' images with the global image delivers $N+1$ images to attack. Finally, we simultaneously mount a targeted attack on these $N+1$ images toward the same target (e.g., *'marmoset'*). In this manner, we expect every block of the obtained AE independently possesses attack capability. The parameter N can be used to balance the attack power and the computational efficiency. Note that the *everywhere* attack degenerates to a baseline attack when $N = 0$. Algorithm 1 summarizes the procedure of integrating the proposed *everywhere* scheme with the CE attack, where DI, TI, and MI are conventional transferability-enhanced methods.

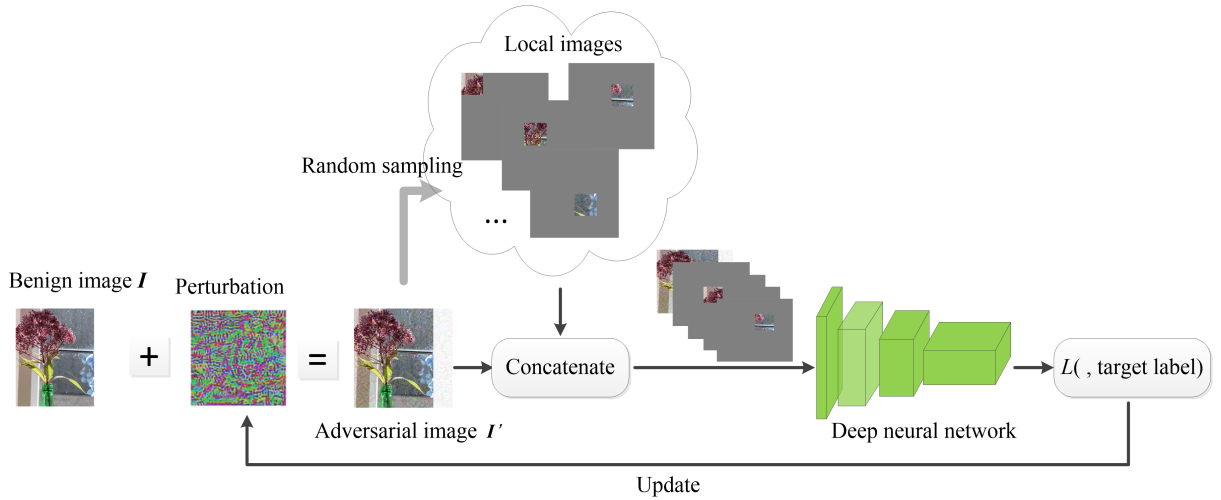


Figure 3: Overview of the proposed *everywhere* attack.

	Res50	Den121	Inc-v3
CE	0.378	0.383	0.251
CE+ <i>everywhere</i>	0.645	0.638	0.504

Table 1: Averaged coverage metric of different victims. Surrogate: VGG16.

The bottom row of Figure 2 shows an AE crafted with the proposed *everywhere* scheme and its attentional maps on different models. The attentional map computed on the surrogate model (Figure 2(g)) presents multiple focal areas. Conceptually, this is similar to the adversary implanting multiple *marmosets* in the image. One of the *marmosets* (the one at the bottom right) is located in the region of interest of Res50 (Figure 2(i)), and another one (the one at the top left) in the region of interest of DenseNet121 (Den121) (Huang et al. 2017) (Figure 2(j)). As a result, our attack successfully transfers to these two victim models.

To conclude this section, we conduct a quantitative experiment on the ImageNet-compatible dataset (ImageNet-compatible 2017). We introduce a coverage metric C to represent the extent of a victim model’s attention (Att_v) being covered by that of the surrogate (Att_s).

$$C = \frac{|Att_v \cap Att_s|}{|Att_v|} \quad (4)$$

where Att is the normalized ($[0, 1]$) and binarized (threshold=2/3) attention map. Table 1 reports the averaged coverage metric over 200 images. Obviously, with the *everywhere* scheme, the victim model’s attention is more likely to overlap with the surrogate’s, i.e., the attention mismatch issue has, in essence, been addressed.

Experimental Results

In this section, we show the efficiency of the proposed *everywhere* attack scheme by integrating it into six iterative attacks: CE, Logit (Zhao, Liu, and Larson 2021), Margin

Algorithm 1: *Everywhere* + CE attack

Input: A benign image I ; target label y_t ; a surrogate model f with loss function J

Parameter: number of partitions for each dimension M , samples N , iterations T

Output: Adversarial Image I'

- 1: Initialize δ_0 and g_0
 - 2: **for** $t=0$ to $T-1$ **do**
 - 3: DI: $I'_t = DI(I + \delta_t)$.
 - 4: Split I'_t into $M \times M$ non-overlap blocks.
 - 5: Randomly sample N blocks and obtain local images $L'_0, L'_1, \dots, L'_{N-1}$ by padding.
 - 6: Concatenate: $I'_t = [I'_t, L'_0, L'_1, \dots, L'_{N-1}]$.
 - 7: Input I'_t to f and obtain gradient $g_{t+1} = \nabla_{\delta} J(I'_t, y_t)$
 - 8: TI and MI: $g_{t+1} = g_t + TI(g_{t+1})$
 - 9: Update and clip δ_{t+1}
 - 10: **end for**
 - 11: **return** $I' = I + \delta_T$
-

(Weng et al. 2023), SH (Zeng et al. 2023), SU (Wei et al. 2023), CFM (Byun et al. 2023) on various transfer scenarios. Since more recent targeted attacks have dominated the Po+Trip attack (Li et al. 2020a), we omit its results for brevity. All the iterative schemes start with the TMDI attack. Then, we contrast *everywhere* attack with two generative attacks: TTP (Naseer et al. 2021) and C-GSP (Yang et al. 2022). Next, the proposed method is used for crafting Data-free Targeted Universal Adversarial Perturbation (DTUAP) (Moosavi-Dezfooli et al. 2017; Zhao, Liu, and Larson 2021), from which our philosophy can be further illustrated. Finally, the crafted AEs are further evaluated using a real-world image recognition system: Google Cloud Vision. The supplementary material provides the ablation study on our key hyper-parameters.

Att.	Source Model: Res50					Source Model: Dense121				
	→Inc-v3	→Den121	→VGG16	→Swin	AVG	→Inc-v3	→Res50	→VGG16	→Swin	AVG
CE	3.9/ 14.1	44.9/ 62.3	30.5/ 52.2	5.2/ 19.0	21.1/ 36.8	2.8/ 10.3	19.0/ 41.7	11.3/ 50.6	1.8/ 19.2	8.7/ 30.5
Logit	9.1/ 22.3	70.0/ 78.5	61.9/ 69.3	13.4/ 28.8	38.6/ 49.7	7.4/ 17.6	42.6/ 58.5	36.3/ 54.2	10.5/ 23.8	24.2/ 38.5
Mar.	10.9/ 21.7	70.8/ 80.8	61.2/ 69.4	16.5/ 33.1	39.9/ 51.3	7.6/ 19.8	44.7/ 58.9	33.4/ 56.4	11.7/ 24.6	24.4/ 39.9
SH	9.9/ 17.8	74.2/ 82.7	62.5/ 78.2	17.1/ 37.3	40.9/ 54.0	8.7/ 12.9	47.4/ 64.3	40.5/ 64.1	9.3/ 23.6	26.6/ 41.2
SU	11.1/ 21.9	72.5/ 79.2	63.9/ 67.4	21.3/ 34.2	42.2/ 50.7	10.0/ 17.2	49.2/ 63.4	42.3/ 55.5	13.5/ 23.1	28.8/ 39.8
CFM	41.4/ 55.3	83.3/ 87.7	77.2/ 81.9	41.5/ 54.2	60.9/ 69.8	35.2/ 43.6	77.3/ 84.8	66.6/ 73.9	27.1/ 43.4	51.6/ 61.4
Att.	Source Model: VGG16					Source Model: Inc-v3				
	→Inc-v3	→Res50	→Den121	→Swin	AVG	→Res50	→Den121	→VGG16	→Swin	AVG
CE	0.0/ 1.8	0.3/ 16.4	0.5/ 15.1	0.1/ 7.6	0.2/ 10.2	1.8/ 6.1	2.5/ 9.6	1.5/ 7.3	0.2/ 0.9	1.5/ 6.0
Logit	0.8/ 3.4	10.6/ 21.8	12.8/ 22.3	6.5/ 13.1	7.7/ 15.2	2.4/ 6.8	3.6/ 14.3	2.2/ 8.9	0.2/ 3.4	2.1/ 8.4
Mar.	0.7/ 3.2	7.9/ 21.1	12.3/ 18.5	6.4/ 10.9	6.8/ 13.4	2.1/ 8.4	3.2/ 14.6	1.9/ 9.3	0.9/ 3.0	2.0/ 8.8
SH	1.1/ 2.6	11.2/ 18.0	13.6/ 22.3	7.0/ 13.7	8.2/ 14.2	2.3/ 7.0	4.5/ 11.5	2.2/ 9.2	0.3/ 2.3	2.3/ 7.5
SU	0.9/ 2.2	13.7/ 25.2	15.7/ 24.6	8.1/ 11.8	9.6/ 16.0	3.0/ 7.4	4.6/ 11.9	3.5/ 8.6	0.9/ 2.8	3.0/ 7.8
CFM	3.8/ 9.3	26.1/ 34.7	28.3/ 39.5	12.4/ 20.8	17.7/ 26.1	12.3/ 29.8	20.9/ 40.3	13.4/ 25.6	4.0/ 11.4	12.7/ 26.8

Table 2: Targeted transfer success rate (%) without/with the proposed *everywhere* scheme, in the random-target scenario. The AVG column is averaged over victims. Best results are in **bold**.

Experimental Settings

Dataset. Following recent work on targeted attacks, our experiments are conducted on the ImageNet-compatible dataset comprised of 1000 images. All these images are with the size of 299×299 pixels and are stored in PNG format.

Networks. Since transferring across different architectures is more demanding, we choose four pretrained models of diverse architectures: Inc-v3, Res50, Den121, and VGG16 as the surrogates. These surrogates and a transformer-based model, Swin (Liu et al. 2021), evaluate AEs’ transferability.

Parameters. For all attacks, the perturbations are restricted by L_∞ norm with $\epsilon = 16$ (The results under lower budgets are provided in the supplementary material), and the step size is set to 2. The total iteration number T is set to 200 to balance speed and convergence. The number of partitions for each dimension M is set to 4, and the number of samples N is set to 9.

Normal Surrogates

Table 2 reports the targeted transferability (random-target) across different models. The proposed *everywhere* scheme boosts all the baseline attacks by a clear margin. Taking the popular Logit attack as a baseline, the average success rate has been improved by 28.8% (49.7% vs. 38.6%) \sim 300% (8.4% vs. 2.1%). Further analysis can provide more insights into the proposed method. First, the weaker the baseline, the more significant the improvement. Hence, the upturn is particularly salient for the CE attack. For example, when VGG16 was the surrogate model, the average success rate of the CE attack increases from 0.2% to 10.2%. Second, the more challenging the transfer scenario is, the more significant the improvement brought by the proposed *everywhere* scheme. For example, the introduced improvement in the ‘Res50→Swin’ scenario is much more salient than that in the ‘Res50→Dense121’, which makes the proposed method even more promising with the popularity of transformer-based networks.

As done in previous works (Zhao, Liu, and Larson 2021; Zeng et al. 2023), we also conduct a worst-case transfer ex-

periment in which the target labels are always the least likely ones. Table 3 compares different attacks: the improvement from the proposed *everywhere* scheme is even more remarkable than the random-target scenario. Taking the Logit attack as the baseline again, the average success rate increases by 39.9% (36.1% vs. 25.8%) when Res50 is the surrogate, and it more than doubles for other surrogates.

Robust Surrogates

Leveraging a slightly robust (adversarially trained) surrogate is accepted as an efficient way to craft transferable targeted AEs (Springer, Mitchell, and Kenyon 2021). We are interested in how the proposed *everywhere* scheme can improve the baselines when robust models are used as surrogates. Specifically, AEs are crafted with adversarially trained models Res18adv and Res50adv and transferred to the same victims used in the last section except Res50. Both models are trained with AEs under $L_2 = 0.01$ budget. Note that there is no architectural overlap between the source and target models. Table 4 presents the targeted transferability in this case. Even though AEs crafted by robust models have shown significantly stronger transferability than those crafted with normal surrogates, the proposed *everywhere* scheme is still helpful, especially when the transformer-based model Swin is the victim. Taking the Logit attack for example, the targeted success rate is doubled in the ‘Res18adv→Swin’ scenario (25.7% vs. 13.1%) and improved by more than a half in the ‘Res50adv→Swin’ scenario (41.6% vs. 23.2%).

Iterative vs. Generative Attacks

Next, we compare the proposed *everywhere* attack with the state-of-the-art generative attacks, TTP and C-GSP. As mentioned before, TTP entails training a generator for each target label and each source model. That means 4×1000 generators are required to perform the random or most difficult-target attack, which is computationally prohibitive. Alternatively, we follow the ‘10-Targets (all source)’ setting of (Naseer et al. 2021) and use ten author-released generators (Res50 being the discriminator during training) to generate

Att.	Source Model: Res50					Source Model: Dense121				
	→Inc-v3	→Den121	→VGG16	→Swin	AVG	→Inc-v3	→Res50	→VGG16	→Swin	AVG
CE	1.3/ 8.8	25.8/ 52.1	15.0/ 42.3	3.2/ 19.9	11.3/ 30.8	1.2/ 6.1	6.5/ 32.7	3.6/ 36.2	0.6/ 12.6	3.0/ 21.9
Logit	3.6/ 9.2	51.6/ 64.7	38.6/ 47.1	9.2/ 23.2	25.8/ 36.1	3.5/ 10.2	22.7/ 46.1	18.3/ 34.9	4.7/ 14.3	12.3/ 26.4
Mar.	4.1/ 12.1	52.3/ 65.8	38.9/ 47.5	10.2/ 22.4	26.5/ 37.0	3.9/ 9.5	24.4/ 44.2	18.2/ 41.3	5.1/ 15.6	12.9/ 27.7
SH	4.0/ 8.8	53.5/ 68.6	41.6/ 60.1	8.1/ 24.8	26.8/ 40.6	3.8/ 7.2	24.5/ 51.1	21.2/ 42.5	5.2/ 15.7	13.7/ 29.1
SU	5.3/ 11.2	54.2/ 66.7	44.1/ 48.6	13.2/ 22.3	29.2/ 37.2	4.4/ 11.3	27.4/ 44.7	24.3/ 39.9	9.0/ 16.8	16.3/ 28.2
CFM	28.2/ 37.3	76.9/ 84.8	61.8/ 70.1	24.5/ 44.2	47.9/ 59.1	27.3/ 36.2	66.1/ 72.6	51.8/ 61.7	21.8/ 36.0	41.8/ 51.6
Att.	Source Model: VGG16					Source Model: Inc-v3				
	→Inc-v3	→Res50	→Den121	→Swin	AVG	→Res50	→Den121	→VGG16	→Swin	AVG
CE	0.0/ 1.1	0.0/ 3.6	0.0/ 6.4	0.0/ 4.8	0.0/ 4.0	2.4/ 7.8	3.8/ 9.1	2.3/ 5.7	0.6/ 2.6	2.3/ 6.3
Logit	0.3/ 0.7	3.3/ 10.9	6.8/ 12.1	5.0/ 11.9	3.9/ 8.9	3.8/ 10.9	4.5/ 10.6	3.2/ 8.3	0.5/ 2.7	3.0/ 8.1
Mar.	0.0/ 0.4	4.4/ 6.8	6.3/ 9.9	6.4/ 8.1	4.3/ 6.3	2.5/ 13.2	4.3/ 13.8	2.0/ 10.9	0.2/ 2.6	2.3/ 10.1
SH	0.1/ 0.3	3.9/ 7.1	6.8/ 8.7	3.1/ 10.2	3.5/ 6.6	3.5/ 10.8	4.9/ 16.7	3.4/ 11.3	0.4/ 3.9	3.1/ 10.7
SU	0.3/ 0.7	5.7/ 9.8	7.4/ 16.8	4.5/ 11.1	4.5/ 9.6	4.3/ 10.2	6.7/ 13.2	3.9/ 8.1	0.8/ 2.1	3.9/ 8.4
CFM	2.7/ 4.2	13.1/ 21.8	17.3/ 28.5	8.6/ 14.9	10.4/ 17.4	16.7/ 35.3	22.2/ 42.1	10.2/ 25.8	4.6/ 17.4	13.4/ 30.2

Table 3: Targeted transfer success rate (%) without/with the proposed *everywhere* scheme, in the most difficult-target scenario.

Att.	Source Model: Res18adv					Source Model: Res50adv				
	→Inc-v3	→Den121	→VGG16	→Swin	AVG	→Inc-v3	→Den121	→VGG16	→Swin	AVG
CE	6.7/ 13.2	29.4/ 44.6	13.2/ 35.9	2.4/ 16.3	12.9/ 27.5	14.4/ 16.9	59.0/ 64.4	24.8/ 53.1	6.6/ 28.8	26.2/ 40.8
Logit	21.8/ 27.0	60.3/ 68.2	46.2/ 50.7	13.1/ 25.7	35.4/ 42.9	26.1/ 30.8	78.6/ 83.9	55.9/ 67.4	23.2/ 41.6	46.0/ 55.9
Mar.	20.4/ 22.4	62.5/ 65.1	43.6/ 51.2	14.2/ 21.1	35.2/ 39.9	26.8/ 29.3	82.3/ 83.6	55.6/ 67.5	25.3/ 38.0	47.5/ 54.6
SH	21.0/ 29.4	68.6/ 75.6	56.1/ 65.4	20.2/ 32.3	41.5/ 50.7	21.4/ 27.5	80.7/ 87.0	67.8/ 76.9	29.1/ 48.2	49.7/ 59.9
SU	23.4/ 30.3	65.8/ 70.3	45.3/ 51.8	15.9/ 25.9	37.4/ 49.6	27.6/ 29.5	79.9/ 81.3	56.8/ 64.2	24.5/ 41.4	47.2/ 54.2
CFM	36.1/ 41.2	75.6/ 80.4	56.7/ 64.6	27.8/ 38.1	49.1/ 56.2	51.8/ 58.3	85.6/ 86.9	74.7/ 79.1	47.5/ 60.2	64.9/ 71.2

Table 4: Targeted transfer success rate (%) without/with the *everywhere* scheme. The AEs are crafted against robust models.

Attack	Inc-v3	Den121	VGG16	Swin	AVG
CE	15.1	63.3	58.8	23.6	40.2
Logit	22.8	83.1	74.0	35.8	53.9
Margin	23.4	83.3	75.4	38.3	55.1
SupHigh	19.0	87.5	85.5	39.4	57.9
SU	24.7	83.1	74.0	36.7	54.6
CFM	59.0	93.7	90.0	59.3	75.5
TTP	39.8	79.5	75.4	44.6	59.8
C-GSP	30.1	67.5	57.0	34.4	47.3

Table 5: Iterative attacks vs. generative attacks, The transfer success rates (%) are averaged over 10 target classes. The upper part of the table presents the results of six iterative attacks, while the lower part shows the results of two generative attacks. Iterative attacks are integrated with the proposed *everywhere* scheme, and the source model is Res50.

AEs. For C-GSP, we train a 10-target conditional generator with Res50 being the discriminator on the ImageNet ‘train’ dataset (Russakovsky et al. 2015).

As shown in Table 5, between two generative methods, the attack ability of the multi-class generator is inevitably inferior to that of the single-class generator. Nevertheless, with the proposed *everywhere* scheme, iterative attacks may yield comparable or even better (CFM + *everywhere*) transferability than generative methods. Such results demonstrate the potential of iterative attacks in the face of generative ones. However, we must admit the intrinsic advantage of the generative attacks: Once the generators are trained, they can craft AEs with much higher computational efficiency than

	Res50	Den121	VGG16	Inc-v3
CE	8.1/ 19.4	8.0/ 28.3	19.2/ 61.5	1.9/ 4.8
Logit	20.7/ 25.1	17.5/ 27.3	64.9/ 70.5	3.6/ 5.0

Table 6: Success rates (%) of the data-free UAPs with $\epsilon = 16$, without/with the proposed *everywhere* scheme.

iterative attacks.

Data-free Targeted UAP

DTUAP is optimized from a random image and can drive multiple clean images into a given class y_t . Due to its data-free nature, DTUAP is a powerful tool for uncovering the intrinsic features of the model of interest. Following (Zhao, Liu, and Larson 2021), we use a mean image (all entrances of which equal 0.5) as the starting point and mount a targeted attack to obtain a DTUAP with CE and Logit attacks ($\epsilon=16$). Then, the obtained DTUAP is applied to all 1000 images in our dataset. Table 6 reports the success rates averaged over 100 classes ($y_t = 0 : 99$). It is observed that the proposed *everywhere* scheme yields more transferable UAPs across input images compared with baselines. For example, with the Logit+*everywhere* scheme, the DTUAPs crafted on the VGG16 model can successfully drive seventy percent of the images into a specified class.

To provide a more intuitive explanation of the proposed *everywhere* scheme, we depict several DTUAP samples in Figure 4. Since features learned by the robust models are more semantically aligned, here we use Res50adv to craft

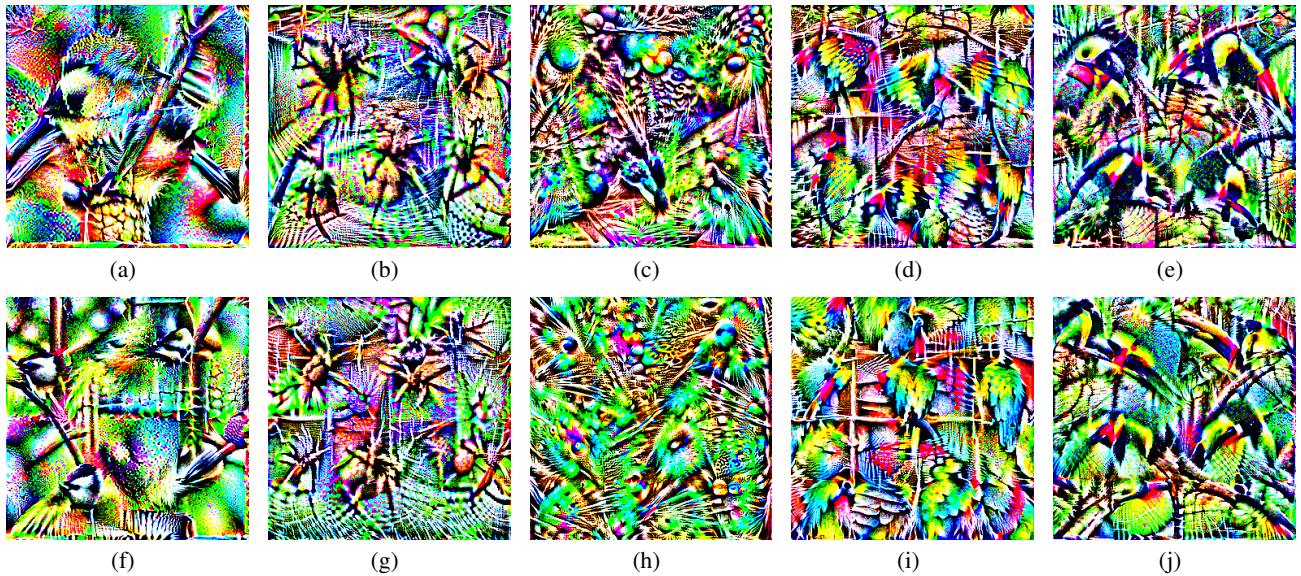


Figure 4: Data-free UAPs of different target classes using Logit (top) and Logit+*everywhere* (bottom). (a, f) ‘chickadee’, (b, g) ‘wolf spider’, (c, h) ‘peacock’, (d, i) ‘macaw’, (e, j) ‘toucan’. The UAPs have been scaled to [0, 1] for better visualization.

Logit	Logit+ <i>everywhere</i>	CFM	CFM+ <i>everywhere</i>
6	11	26	47

Table 7: Success rates (%) of different attacks on Google Cloud Vision. Surrogate: Res50adv.

DTUAPs. Compared to the baseline attack, the *everywhere* attack tends to plant more target objects with smaller sizes into the obtained UAP. Such a distinction is apparent in the case of ‘chickadee’. Only one big *chickadee* can be observed in the DTUAP crafted by the vanilla Logit attack (Figure 4(a)). In contrast, at least four baby *chickadees* can be found in the DTUAP crafted by Logit+*everywhere* attack (Figure 4(f)).

Fooling Google Cloud Vision

Finally, we evaluate the crafted AEs on the Google Cloud Vision API. Specifically, targeted AEs are generated with Res50adv, and the API returns a list of semantic labels for each probe image. As (Zhao, Liu, and Larson 2021), the attack is deemed a success once the target appears in the returned list. Note that we regard semantically similar classes as the same since the semantic label set of the API does not precisely match the ImageNet classes.

Table 7 reports the targeted success rate averaged over 100 images. Google Cloud Vision API is much more difficult to transfer than previously studied models. Nevertheless, the proposed *everywhere* scheme effectively improves the transferability of baselines. Due to page limitations, we provide the sample images and the API outputs in the supplementary material.

Conclusion

The discriminative regions of a target class on victim models are dramatically different from that on the surrogate, which severely constrains the targeted transferability of AEs. To address this challenge, we propose the *everywhere* attack, which optimizes an army of target objects in every local image region that victim models may pay attention to and thus reduces the transfer failures caused by attention mismatch. Extensive experiments demonstrate that the proposed method can universally boost the transferability of existing targeted attacks. It is our hope that the idea of increasing the target quantity opens a new door to boosting targeted transferability for the community.

Acknowledgments

This work was supported by the Opening Project of Guangdong Province Key Laboratory of Information Security Technology (no. 2023B1212060026) and the Start-up Scientific Research Project for Introducing Talents of Beijing Normal University at Zhuhai (no. 312200502504).

Salute to Mr. Cheng’en, the author of the novel *Journey to the West*, for his 520th anniversary of birth.

References

- Byun, J.; Cho, S.; Kwon, M.; and *et al.* 2022. Improving the transferability of targeted adversarial examples through object-based diverse input. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR-2022)*, 15244–15253.
- Byun, J.; Kwon, M.; Cho, S.; and *et al.* 2023. Introducing competition to boost the transferability of targeted adversarial examples through clean feature mixup. In *IEEE/CVF*

- Conf. on Computer Vision and Pattern Recognition*, 24648–24657.
- Dong, Y.; Liao, F.; Pang, T.; and *et al.* 2018. Boosting adversarial attacks with momentum. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 9185–9193.
- Dong, Y.; Pang, T.; Su, H.; and *et al.* 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 4307–4316.
- Fan, M.; Guo, W.; Ying, Z.; and *et al.* 2023. Enhance transferability of adversarial examples with model architecture. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5.
- He, K.; Zhang, X.; Ren, S.; and *et al.* 2016. Deep residual learning for image recognition. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 770–778.
- Huang, G.; Liu, Z.; Laurens, V.; and *et al.* 2017. Densely connected convolutional networks. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2261–2269.
- ImageNet-compatible. 2017. https://github.com/cleverhans-lab/cleverhans/tree/master/cleverhans.v3.1.0/examples/nips17_adversarial_competition. *Online*.
- Inkawhich, N.; Liang, K.; Carin, L.; and *et al.* 2020. Transferable perturbations of deep feature distributions. In *International Conf. on Learning Representations*.
- Kurakin, A.; Goodfellow, I.; and Bengio, S. . 2016. Adversarial examples in the physical world. In *International Conf. on Learning Representations*.
- Li, M.; Deng, C.; Li, T.; and *et al.* 2020a. Towards transferable targeted attack. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 638–646.
- Li, Y.; Bai, S.; Zhou, Y.; and *et al.* 2020b. Learning transferable adversarial examples via ghost networks. In *the 34th AAAI Conf. on Artificial Intelligence*, 11458–11465.
- Liu, Y.; Chen, X.; Liu, C.; and *et al.* 2017. Delving into transferable adversarial examples and black-box attacks. In *International Conf. on Learning Representations*.
- Liu, Z.; Lin, Y.; Gao, Y.; and *et al.* 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF international conf. on computer vision (ICCV-2021)*, 10012–10022.
- Madry, A.; Makelov, A.; Schmidt, L.; and *et al.* 2018. Towards deep learning models resistant to adversarial attacks. In *International Conf. on Learning Representations*.
- Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. arXiv:1411.1784.
- Moosavi-Dezfooli, S. M.; Fawzi, A.; Fawzi, O.; and *et al.* 2017. Universal adversarial perturbations. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.
- Naseer, M.; Khan, S.; Hayat, M.; and *et al.* 2021. On generating transferable targeted perturbations. In *IEEE/CVF international conf. on computer vision*, 7688–7697.
- Russakovsky, O.; Deng, J.; Su, H.; and *et al.* 2015. ImageNet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.
- Selvaraju, R. R.; Cogswell, M.; Das, A.; and *et al.* 2017. Grad-CAM: Visual explanations from deep networks via a gradient-based localization. In *IEEE/CVF international conf. on computer vision*, 618–626.
- Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conf. on Learning Representations*.
- Springer, J.; Mitchell, M.; and Kenyon, G. T. 2021. A little robustness goes a long way: Leveraging robust features for targeted transfer attacks. In *the 35th Conf. on Neural Information Processing Systems*.
- Szegedy, C.; Vanhoucke, V.; Ioffe, S.; and *et al.* 2016. Rethinking the inception architecture for computer vision. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2818–2826.
- Szegedy, C.; Zaremba, W.; Sutskever, I.; and *et al.* 2014. Intriguing properties of neural networks. In *International Conf. on Learning Representations*.
- Wang, K.; Shi, J.; and Wang, W. 2023. LFAA: crafting transferable targeted adversarial examples with low-frequency perturbations. In *the 26th European Conf. on Artificial Intelligence*, 2483–2490.
- Wang, X.; He, X.; Wang, J.; and *et al.* 2021. Admix: Enhancing the transferability of adversarial attacks. In *IEEE/CVF international conf. on computer vision*, 16518–16167.
- Wei, Z.; Chen, J.; Wu, Z.; and *et al.* 2023. Enhancing the self-universality for transferable targeted attacks. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 12281–12290.
- Weng, J.; Luo, Z.; Zhong, Z.; and *et al.* 2023. Logit margin matters: Improving transferable targeted adversarial attack by logit calibration. *IEEE Trans. on Information Forensics and Security*, 18: 3561–3574.
- Wu, W.; Su, Y.; Chen, X.; and *et al.* 2020. Boosting the transferability of adversarial samples via attention. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 1161–1170.
- Xie, C.; Zhang, Z.; Zhou, Y.; and *et al.* 2019. Improving transferability of adversarial examples with input diversity. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2725–2734.
- Yang, X.; Dong, Y.; Pang, T.; and *et al.* 2022. Boosting transferability of targeted adversarial examples via hierarchical generative network. In *European Conf. on Computer Vision*, 725–742.
- Zeng, H.; Zhang, T.; Chen, B.; and *et al.* 2023. Enhancing targeted transferability via suppressing high-confidence labels. In *International Conf. on Image Processing*, 3309–3313.
- Zhao, Z.; Liu, Z.; and Larson, M. 2021. On success and simplicity: a second look at transferable targeted attacks. In *Advances in Neural Information Processing Systems*, 6115–6128.