Logit Margin Matters: Improving Transferable Targeted Adversarial Attack by Logit Calibration

Juanjuan Weng, Zhiming Luo, Shaozi Li, Nicu Sebe, Zhun Zhong

Abstract—**Previous works have extensively studied the transferability of adversarial samples in untargeted black-box scenarios. However, it still remains challenging to craft targeted adversarial examples with higher transferability than non-targeted ones. Recent studies reveal that the traditional Cross-Entropy (CE) loss function is insufficient to learn transferable targeted adversarial examples due to the issue of vanishing gradient. In this work, we provide a comprehensive investigation of the CE loss function and find that the logit margin between the targeted and untargeted classes will quickly obtain saturation in CE, which largely limits the transferability. Therefore, in this paper, we devote to the goal of continually increasing the logit margin along the optimization to deal with the saturation issue and propose two simple and effective logit calibration methods, which are achieved by downscaling the logits with a temperature factor and an adaptive margin, respectively. Both of them can effectively encourage optimization to produce a larger logit margin and lead to higher transferability. Besides, we show that minimizing the cosine distance between the adversarial examples and the classifier weights of the target class can further improve the transferability, which is benefited from downscaling logits via L2-normalization. Experiments conducted on the ImageNet dataset validate the effectiveness of the proposed methods, which outperform the state-of-the-art methods in black-box targeted attacks. The source code is available at Link.**

Index Terms—**Adversarial machine learning, convolutional neural networks, Targeted adversarial examples, logit calibration**

I. Introduction

In the past decade, deep neural networks (DNNs) have achieved remarkable success in various fields, *e.g.*, image classification [1], image segmentation [2], and object detection [3]. However, Goodfellow et al. [4] revealed that the DNNs are vulnerable to adversarial attacks, in which adding imperceptible disturbances into the input can lead the DNNs to make an incorrect prediction. The adversarial perturbation will raise a vital threat for the real-world applications of the CNNs, especially in scenarios of speech recognition [5], [6], facial verification systems [7], [8] and person re-identification systems [9], [10]. Many following approaches [11], [12], [13], [14], [15], [16], [17] have been proposed to construct more destructive adversarial samples for investigating the vulnerability of the DNNs. Besides, some studies [4], [18] also showed that the adversarial samples are transferable across different networks, raising a more critical robustness threat under the black-box scenarios. Therefore, it is vital to explore the vulnerability of the DNNs, which is extremely useful for designing robust DNNs.

Currently, most of the works [11], [15], [19], [20], [21], [22], [23] have been devoted to the untargeted black-box attacks, in which adversarial examples are crafted to fool unknown CNN models predicting unspecified incorrect labels. For example, [11], [15] leveraged input-level transformation or augmentation to improve the non-targeted transferability. [20] proposed a powerful intermediate feature-level attack. [21], [22] demonstrated that backpropagating more gradients through the skip-connections can increase the transferability. [24] proposed the stochastic variance reduced ensemble (SVRE) attack to reduce the gradient variance of the ensemble models for improving transferability. Despite the success in non-targeted cases, targeted transferability remains challenging, which requires eliciting the black-box models into a pre-defined target category.

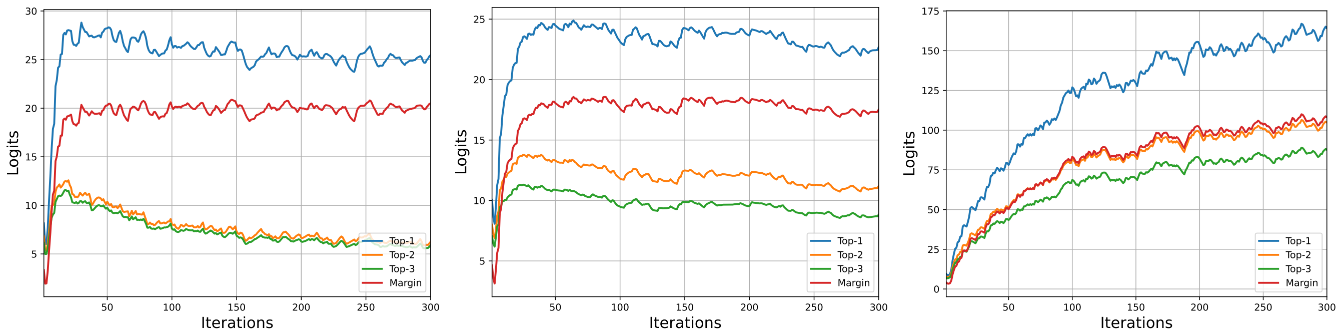
For learning the transferable adversarial samples in untargeted cases, most methods have leveraged the Cross-Entropy (CE) as the loss function. However, recent studies [25], [26] showed that the CE loss is insufficient for learning the adversarial perturbation in the targeted case due to the issue of vanishing gradient. To deal with this issue, Li et al. [25] adopted the Poincaré distance to increase the gradient magnitude adaptively during the optimization. Zhao et al. [26] demonstrated that an effortless logit loss equal to the negative value of the targeted logits could alleviate the gradient vanishing issue and achieve surprisingly strong targeted transferability. Besides, Zhao et al. [26] also showed that optimizing with more iterations can significantly increase the targeted transferability. Although it demonstrated that continually enlarging the logit of the targeted class along the whole training iteration (as shown in Figure 1c) can improve the transferability of adversarial samples, it still does not thoroughly analyze the insufficient issue in the CE loss function

Different from [25], [26], in this study, we take a closer look at the vanishing gradient issue in the CE loss and find the logit margin between the targeted and non-targeted classes will quickly get saturated during the optimization (as shown in Figure 1a). Moreover, this issue will influence the attack performance of the adversarial examples and thus essentially limits the transferability. Specifically, along with the training iterations in CE, we observe the logits of the targeted and non-targeted classes increase rapidly in the first few iterations. However, after reaching the peak, the logit margin between the targeted and non-targeted classes will get saturated, and further training will decrease the logits

This work is supported by the National Natural Science Foundation of China (No. 62276221); the Natural Science Foundation of Fujian Province of China (No. 2022J01002); EU H2020 project AI4Media (No. 951911).

J. Weng, Z. Luo (Corresponding author), and S. Li are with the Department of Artificial Intelligence, Xiamen University, Xiamen 361005, China.

N. Sebe and Z. Zhong are with the Department of Information Engineering and Computer Science, University of Trento, Trento 38100, Italy.



(a) Cross-Entropy (b) Po+Trip (c) Logit

Fig. 1: The average Top-3 logits and logit margin of 50 adversarial samples generated from the ResNet-50 by the Cross-Entropy, Po+Trip [25] and Logit [26] loss functions. (* Training and computation details of this figure are in Section III-B)

simultaneously to maintain this margin. This phenomenon is mainly due to the fact that the softmax function in CE will approximately output the probability of the target class to 1 when reaching the saturated margin (*e.g.*, 10). Thus, it raises the problem that the transferability will not be further increased even with more optimization iterations. A similar saturated phenomenon also can be observed in the Po+Trip loss [25] as shown in Figure 1b. While in practice, we are encouraged to improve the transferability by keeping increasing the logit for the targeted class and its margin along the optimization to avoid saturation against other non-targeted classes to cross the decision boundaries of other black-box models.

In this paper, we devote to keeping enlarging logit margins along the optimization to alleviate the above saturation issue in CE. Inspired by the temperature scaling technique employed in knowledge distillation where logits are divided by a temperature parameter T [27], a higher temperature T produces a softer probability distribution over different classes after applying the softmax function. We first utilize the scaling technique for targeted attacks to calibrate the logits. Then the logit margin between the targeted and non-targeted classes will not be saturated after only a few iterations and will keep improving the transferability. Besides, instead of using a constant T, we further explore an adaptive margin-based calibration by scaling the logits based on the logit margin of the target class and the highest non-target class. In addition, we also investigate the effectiveness of calibrating the targeted logit into the unit length feature space by L2-normalization, which is equivalent to minimizing the angle between the adversarial examples and the classifier weights of the targeted class.

Finally, we conduct experiments on the ImageNet dataset to validate the effectiveness of the logits calibration for crafting transferable targeted adversarial examples. Experimental results demonstrate that the calibration of the logits helps achieve a higher attack success rate than other state-of-the-art methods. Besides, the combination of different calibrations can further provide mutual benefits.

II. Related Works

In this section, we give a brief introduction of the related works from the following two aspects: *untargeted black-box attacks* and *targeted attacks*.

A. Untargetd Black-box Attacks

After Szegedy et al. [28] exposed the vulnerability of deep neural networks, many attack methods [15], [12] have been proposed to craft highly transferable adversaries in the non-targeted scenario. We first review several gradient-based attack methods that focus on enhancing the transferability against black-box models.

Iterative-Fast Gradient Sign Method (I-FGSM) [29] is an iterative version of FGSM [4], which iteratively adds the perturbation with a small step size α in the gradient direction:

$$\hat{x}_0 = x, \quad \hat{x}_{i+1} = \hat{x}'_i + \alpha \cdot \mathrm{sign}(\nabla_{\hat{x}} J(\hat{x}'_i, y)), \qquad (1)$$

where $\hat{x}'_i$ denotes the adversarial image in the i_{th} iteration, $\alpha = \epsilon / T$ ensures the adversary is constrained within an upper-bound perturbation ϵ through the l_p-norm when optimized by T iterations.

Following the seminal I-FGSM [29], a series of methods have been proposed to improve the transferability of attacking black-box models from different aspects, *e.g.*, gradient-based, input augmentation-based. For example, the **Momentum Iterative-FGSM (MI-FGSM)** [11] introduces a momentum term to compute the gradient of the I-FGSM, encouraging the perturbation is updated in a stable direction. The **Translation Invariant-FGSM (TI-FGSM)** [12] adopts a predefined kernel W to convolve the gradient $\nabla_{\hat{x}} J(\hat{x}'_i, y)$ at each iteration t, which can approximate the average gradient over multiple randomly translated images of the input $\hat{x}_t$. On the other aspects, the **Diverse Input-FGSM (DI-FGSM)** [15] leverages the random resizing and padding to augment the input $\hat{x}_t$ at each iteration. Currently, most targeted attack methods [25], [26], [30] simultaneously use the MI [11], TI [12] and DI [15] to form a strong baseline with better transferability.

B. Targeted Attacks

Targeted attacks are different from non-targeted attacks, which need to change the decision to a specific target class. [31] integrates the above non-targeted attack methods into targeted attacks to craft targeted adversarial examples. However, the performance is limited because it is insufficient to fool the black-box model only by maximizing the probability of the target class in the CE loss.

Po+Trip [25] finds the insufficiency of CE is mainly due to the vanishing gradient issue. Then, [25] leverages the Poincaré space as the metric space and further utilizes Triplet loss to improve targeted transferability by forcing adversarial examples toward the target label and away from the ground-truth labels. To further address this gradient issue, **Logits** [26] adopts a simple and straightforward idea by directly maximizing the target logit to pull the adversarial examples close to the target class, which can be expressed as:

$$L_{Logit} = -z_t(\boldsymbol{x}'), \tag{2}$$

where $z_t(\cdot)$ is the output logits of the target class. Based on the prior **Logits** [26], the ODI [16] proposes the object-based diverse input (ODI) method to diversify the input for further improving targeted transferability.

On the other hand, many studies employ resource-intensive approaches to achieve targeted attacks, which train target class-specific models (auxiliary classifiers or generative models) on additional large-scale data. For example, the FDA [32], [33] uses the intermediate feature distributions of CNNs to boost the targeted transferability by training class-specific auxiliary classifiers to model layer-wise feature distributions. The GAP [34] trains a generative model for crafting targeted adversarial examples. Subsequently, [35] adopts a relativistic training objective to train the generative model for improving attack performance and cross-domain transferability. Recently, the TTP [30] utilizes the global and local distribution matching for training target class-specific generators for obtaining high targeted transferability. However, the TTP requires actual samples from the target class and brings expensive training costs. Different from the above methods, we introduce three simple and effective logit calibrations into the CE loss, which can achieve competitive performance without additional data and training.

III. Method

A. Problem Definition

Given a white-box model $\mathbb{F}_s$ and an input image x not from the targeted class t, our primary goal is to learn an imperceptible perturbation δ that can fool the $\mathbb{F}_s$ to output the target t for $\hat{x} = x + \delta$. Besides, the prediction of $\hat{x}$ will also be t when feeding into other unknown black-box models. The l_∞-norm is usually used to constrain the perturbation δ within an upper-bound ϵ, denoted as $||\delta||_\infty \leq \epsilon$.

For the surrogate model $\mathbb{F}_s$, we denote the feature of the final classification layer of the input x as $\phi(x)$. The logit z_i of the category i is computed by $z_i = W_i^T \phi(x) + b_i$, where W_i and b_i are the classifier weights and bias for category i. The corresponding probability p_i after the softmax function is calculated by $p_i = \frac{e^{z_i}}{\sum e^{z_j}}$.

B. Logit Margin

When successfully attacking the $\mathbb{F}_s$, the logit z_t of the target class t will be higher than the logits z_{nt} of any other non-target class in the classification task. Their logit margins can be computed by,

$$G(\phi(\hat{x})) = z_t - z_{nt} = W_t^T \phi(\hat{x}) + b_t - W_{nt}^T \phi(\hat{x}) - b_{nt}. \tag{3}$$

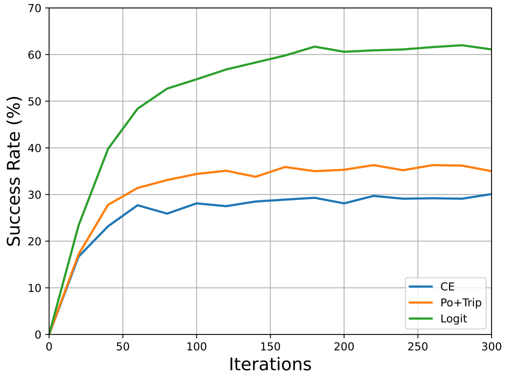


Fig. 2: The targeted attack success rate (%) on VGG-16 by using the ResNet-50 as the surrogate model.

[25], [26] showed that it is insufficient to obtain transferable targeted adversarial samples that are only close to the target class while not far away enough from the true class and other non-targeted classes. Based on this property, it encourages us to continually enlarge this logit margin along the optimization to increase the separation between the targeted and other non-targeted classes, thereby improving transferability.

To have a better understanding of the relationship between the logit margins and the targeted transferability, we visualize the average Top-3 logits (1 targeted class and other two non-targeted classes) of 50 random adversarial samples trained on ResNet50 by the CE, Po+Trip [25], and the Logit [26] loss functions. We also compute the average logit margin of the targeted class against the Top-20 non-targeted classes. The logit and the average logit margin are shown in Figure 1, and the transfer targeted attack success rate of these three loss functions from ResNet50 to VGG16 is plotted in Figure 2.

From Figure 1, we can observe that the logits of the targeted class and the Top-2 non-targeted classes increase rapidly in the first few iterations for the CE and Po+Trip loss, as well as their logit margins. When reaching the peak, the margin is saturated, and the logits start to decrease simultaneously to maintain the saturated margin. By comparing the CE and Po+Trip, the Po+Trip needs slightly more iterations to reach the saturated status and thus shows a marginal better transferability than CE, as shown in Figure 2. In comparison, the Logit loss function will keep increasing the logits of the targeted category and the logit margin. Hence, the Logit loss function shows a much better transfer targeted-attack success rate than CE and Po+Trip. On the other hand, the Logit loss also significantly increases the logits from other non-targeted classes when training with more iterations.

To further analyze why the logit margin will quickly reach saturation in the CE loss function and explore the effectiveness of increasing the margin during training. In the following sections, we will revisit the cross-entropy loss function and introduce the logit calibration to achieve this goal.

C. Revisiting the Cross-Entropy Loss

Firstly, our objective is to maximize the logit margin in Eq. 3. After computing the gradient w.r.t. to $\phi(\hat{x})$, we can get

$$\frac{\partial G}{\partial\phi(\hat{x})} = W_t - W_{nt}. \tag{4}$$

This gradient indicates that the adversarial feature $\phi(\hat{x})$ needs to move towards the target class while apart from those non-target classes. Next, we compute the gradient w.r.t. to $\phi(\hat{x})$ in the Cross-Entropy loss function

$$L_{ce} = -\log(p_t) = -z_t + \log(\sum e^{z_j}), \tag{5}$$

where the p_t is probability about target class t and the z_t is the logit of the category t. And we can get the gradient

$$\begin{aligned}\frac{\partial L_{ce}}{\partial\phi(\hat{x})} &= -\frac{\partial z_t}{\partial\phi(\hat{x})} + \frac{1}{\sum e^{z_j}}\cdot\frac{\partial\sum e^{z_j}}{\partial\phi(\hat{x})} \\ &= -\frac{\sum e^{z_i}}{\sum e^{z_j}}\cdot\frac{\partial z_t}{\partial\phi(\hat{x})} + \frac{1}{\sum e^{z_j}}\sum e^{z_i}\frac{\partial z_i}{\partial\phi(\hat{x})} \\ &= \sum\frac{e^{z_i}}{\sum e^{z_j}}\cdot(\frac{\partial z_i}{\partial\phi(\hat{x})} - \frac{\partial z_t}{\partial\phi(\hat{x})}) \\ &= \sum -p_i(W_t - W_i),\end{aligned} \tag{6}$$

where W_t is the classifier weights and bias for category t. From Eq. 6, we actually find that the CE loss is designed to adaptively optimize the $\phi(\hat{x})$ towards W_t and away from other W_i. However, after optimization with several iterations, the p_i of the non-targeted class will soon approach to 0 and then the influence of $W_t - W_i$ significantly vanishes.

Let's consider the case only with 2 classes (t and nt), we have the probabilities p_t and p_{nt} as:

$$p_t = \frac{e^{z_t}}{e^{z_t} + e^{z_{nt}}} = \frac{1}{1 + e^{-(z_t - z_{nt})}}, \tag{7}$$

$$p_{nt} = \frac{e^{z_{nt}}}{e^{z_t} + e^{z_{nt}}} = \frac{1}{1 + e^{(z_t - z_{nt})}}. \tag{8}$$

As shown in Figure 3, the p_t will get close to 1 when $z_t - z_{nt} > 6$ (*e.g.*, $p_{nt} \approx 2e^{-9}$ when $z_t - z_{nt} = 20$). In such a context, the gradient will significantly vanish. Recall that, in the CE loss function (Figure 1a), the logit margin increases rapidly but will reach saturation when approaching a certain value. This further indicates that the optimization of the CE loss function is largely restrained when the logit margin reaches a certain value. To this end, we raise the question *if we explicitly enforce the optimization to increase the logit margin ($z_t - z_{nt}$) continually during the training process, could we get better transferable targeted adversarial samples?*

To answer this, we propose to downscale the $z_t - z_{nt}$ by a factor s in the CE and extend the informative optimization for more iterations. Since in such circumstances, $z_t - z_{nt}$ will be enlarged by the factor s. Specifically, suppose that the optimization will be saturated when $z_t - z_{nt}$ reaches a certain value v. Using $z_t - z_{nt}$ and $\frac{z_t - z_{nt}}{s}$ in the CE will both approach the saturated value v. Then, it is easy to infer that, for the latter case, $z_t - z_{nt}$ will be $v \times s$.

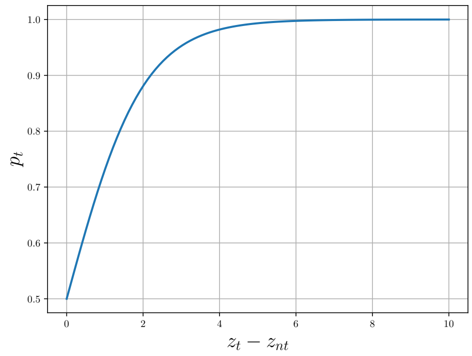


Fig. 3: The probability of p_t under different $z_t - z_{nt}$.

D. Calibrating the Logits

To downscale the $z_t - z_{nt}$ during the optimization, we investigate three different types of logit calibration methods in this study, *i.e.*, Temperature-based, Margin-based, and Angle-based.

1) Temperature-based: Inspired by the Temperature-scaling used in the knowledge distillation [27], our first logit calibration directly downscales the logits by a constant temperature factor T,

$$\tilde{z}_i = \frac{z_i}{T}. \tag{9}$$

After introducing the T, the probability distribution $\boldsymbol{p}$ will be softer over different classes. The corresponding gradient can be computed by:

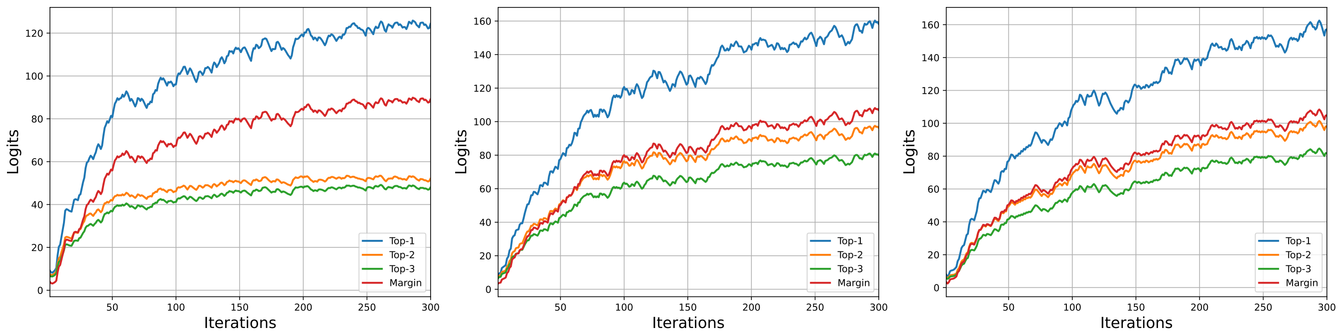
$$\begin{aligned}\frac{\partial L_{ce}^T}{\partial\phi(\hat{x})} &= \frac{e^{z_j/T}}{\sum e^{z_j/T}}\cdot\frac{1}{T}(\frac{\partial z_j}{\partial\phi(\hat{x})} - \frac{\partial z_t}{\partial\phi(\hat{x})}) \\ &= \sum -\hat{p}_i\frac{(W_t - W_i)}{T}.\end{aligned} \tag{10}$$

After downscaling by the factor T, the new $\hat{p}_i$ after the softmax will not quickly approach 0 when only trained with a few iterations. In Figure 4a and Figure 4b, we visualize the trend of logits and margins using $T = 5$ and $T = 20$. We can find that targeted logits and the logit margin will keep increasing as the same as the Logit loss [26] in Figure 1. Meanwhile, the trend of $T = 20$ is very similar to the Logit [26].

2) Margin-based: The previous Temperature-based logit calibration contains a hype-parameter T, which could be different for different surrogate models $\mathbb{F}_s$. To migrate this issue, we further introduce an adaptive margin-based logit calibration without extra hype-parameters. Specifically, we calibrate the logits by using the margin between the Top-2 logits in each iteration, denoted as:

$$\tilde{z}_i = \frac{z_i}{\hat{z}_1 - \hat{z}_2}, \tag{11}$$

where $\hat{z}_1$ and $\hat{z}_2$ are the Top-1 and the Top-2 logit, respectively. Suppose the $\tilde{z}$ is sorted, the Top-1 logit $\tilde{z}_1$ will be the target



(a) T=5

(b) T=20

(c) Margin

Fig. 4: After the logit calibration, the average Top-3 logits and logit margin of 50 adversarial samples crafted from ResNet-50.

class $\tilde{z}_t$ after a few iterations. Therefore, the corresponding calibrated probability of the target class will be:

$$\begin{aligned} p_t &= \frac{1}{1+\sum_{i\neq t} e^{-(\tilde{z}_t-\tilde{z}_i)}} \\ &= \frac{1}{1+e^{-(\tilde{z}_t-\tilde{z}_i)}+\sum_{i=2} e^{-(\tilde{z}_t-\tilde{z}_i)}} \\ &= \frac{1}{1+e^{-\frac{\hat{z}_t-\hat{z}_2}{\hat{z}_t-\hat{z}_2}}+\sum_{i=2} e^{-(\tilde{z}_t-\tilde{z}_i)}} \\ &< \frac{1}{1+e^{-1}}. \end{aligned} \tag{12}$$

Correspondingly, we can have the probability of $1-p_t > 1-\frac{1}{1+e^{-1}}$. Therefore, the probability $p_{\hat{1}}$ of the Top-1 non-target class will be larger than the average probability of all non-target classes, denoted as:

$$p_{\hat{1}} = \frac{1}{e^{\tilde{z}_{\hat{1}}-\tilde{z}_t}+\sum_{i\neq t} e^{\tilde{z}_i-\tilde{z}_{\hat{1}}}} > \frac{1}{N-1}(1-\frac{1}{1+e^{-1}}). \tag{13}$$

Then, it can adaptively deal with the vanishing gradient issue in the original CE loss function. The logits and the margin is shown in Figure 4c.

3) Angle-based: On the other aspect, we also can adaptively calibrate the logit $W_i^T\phi(\hat{x})+b_i$ of each category based on the weights W_i. Besides, the weights W_i for different category i usually has a different norm. To further alleviate the influence of various norms, we calibrate the logit into the feature space with unit length by L2-normalization, denoted as $\frac{W_i^T\phi(\hat{x})+b_i}{||W_i||||\phi(\hat{x})||}$.

If omit the b_i, this calibration is actually computed the $\cos(\theta)$ between $\phi(\hat{x})$ and W_i, and we term it as angle-based calibration. *Since this angle-based calibration will bound each logit smaller than one.* Instead of using the CE loss function, we directly minimize the angle between the $\phi(\hat{x})$ and the targeted weights W_t. The corresponding optimization loss function is:

$$L_{cosine} = -\frac{W_t^T\phi(\hat{x})}{||W_t||||\phi(\hat{x})||}. \tag{14}$$

Notice that, the angle-based classifiers have been widely used in Face-Recognition task [36], [37].

Relationship of these three calibrations: The temperature-based is the simplest one which only calibrates the logits by a constant value of T. However, the optimal T is different for different models, as shown in Tables I and II. Therefore, we investigate the margin-based and angle-based calibrations to deal with this hyper-parameter issue. The margin-based method calculates the temperature parameter T adaptively by considering the Top-2 logits at each iteration, instead of using a constant value. On the aspect, since $z_i = W_i^T\phi(\hat{x})+b_i$ and the L_2 norm of W_i is different for each class i, we could perform the calibration by normalizing the classifier weight W_i of each class i and normalizing the feature to the unit length by L_2 normalization. This L_2 normalization-based calibration is actually computing the cosine between the W_i and $\hat{x}$ without considering their norms. Therefore, we refer to it as "angle-based calibration". Additionally, the three different calibration methods used in this study are complementary to other, and we will evaluate their performance and their mutual benefits in the following experimental section.

Difference between our Margin-based calibration and the DLR loss [38]: The targeted-DRL loss is $\mathrm{DRL}(x,y) = \frac{z_y-z_t}{\hat{z}_1-(\hat{z}_3+\hat{z}_4)/2}$, which is a margin loss function and requires the original label y of each sample for optimization. The targeted DRL is very similar to the Logit loss $\mathrm{Logit}(x) = -z_t$. However, in our margin-based calibration, we only adaptively calibrate the logit $\tilde{z}_i = \frac{z_i}{\hat{z}_1-\hat{z}_2}$ and still optimize with the CE loss. Besides, our margin-based calibration doesn't need the original label for computing the loss function.

IV. The connection with Logit Loss

In this section, we theoretically analyze the relationship between the Logit loss [26] and Temperature-based calibration with a large T. When using a large T, the distribution $\hat{p}_i$ will be extremely smooth over different classes. And we can get the $\hat{p}_i \approx \frac{1}{N}$ for each class, where N is the number of classes. Besides, the $\sum_i W_i$ are with very small values (i.e., $\sum_i W_i \approx 0$). In this study, we conduct experiments on the ImageNet dataset ($N = 1000$). Then the gradient of Cross-Entropy with a large T in Eq. 10 will become:

$$\begin{aligned} \frac{\partial L_{ce}^T}{\partial \phi(\hat{x})} &\approx \sum_i -\frac{(W_t - W_i)}{NT} \\ &\approx -\frac{W_t}{T} + \frac{1}{NT}\sum_i W_i \\ &\approx -\frac{W_t}{T}. \end{aligned} \tag{15}$$

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2023.3284649

TABLE I: The targeted transfer success rates (%) in the single-model transfer scenario. (Results with 20/100/300 iterations are reported, and the highest one at 300 iterations is shown **bold**.)

Attack	Surrogate Model: ResNet50			Surrogate Model: Dense121		
	→Dense121	→VGG16	→Inc-v3	→Res50	→VGG16	→Inc-v3
CE	27.0/40.2/42.7	17.4/27.6/29.1	2.3/4.1/4.6	12.3/17.2/18.4	8.6/10.5/10.9	1.6/2.3/2.8
Po+Trip	27.9/51.2/54.8	17.9/35.5/34.7	3.2/6.8/7.8	11.0/14.8/15.0	7.3/9.2/8.6	1.6/2.8/2.8
Logit	31.4/64.0/71.8	23.8/55.0/62.4	3.1/8.6/10.9	17.4/38.6/43.5	13.7/33.8/37.8	2.3/6.6/7.5
T=5	33.3/69.9/**77.8**	24.8/59.9/66.1	3.1/9.4/**12.2**	19.3/43.4/47.5	14.6/36.6/39.4	2.3/7.3/8.8
T= 10	31.6/68.5/77.0	23.6/58.5/**66.4**	2.8/9.4/11.6	17.9/43.2/**49.3**	13.4/36.8/**41.5**	2.2/7.7/8.8
Margin	33.3/65.8/76.5	23.1/58.6/65.7	3.0/9.5/12.2	18.8/42.8/47.2	14.5/36.5/41.4	2.5/7.7/**9.4**
Angle	38.9/72.5/77.2	29.2/60.7/65.2	4.4/10.7/11.1	20.6/43.2/47.8	16.5/35.7/39.3	3.0/7.7/8.9

Attack	Surrogate Model: VGG16			Surrogate Model: Inc-v3		
	→Res50	→Dense121	→Inc-v3	→Res50	→Dense121	→VGG16
CE	0.5/0.3/0.6	0.6/0.3/0.3	0/0/0.1	0.7/1.2/1.8	0.6/1.3/1.9	0.4/0.8/1.3
Po+Trip	0.7/0.6/0.7	0.7/0.6/0.5	0.1/0.1/0.1	1.0/1.6/1.7	0.6/1.7/2.5	0.7/1.2/1.8
Logit	3.4/9.9/11.6	3.5/12.0/13.9	0.3/1.0/**1.3**	0.6/1.1/2.0	0.6/1.9/3.0	0.6/1.5/**2.8**
T=5	3.1/7.0/6.9	3.3/7.6/7.8	0.2/0.9/0.8	0.7/1.7/2.1	0.5/1.9/**3.3**	0.4/1.6/2.6
T= 10	3.6/9.0/9.7	3.4/10.5/11.7	3.2/1.1/**1.3**	0.5/1.3/1.9	0.6/2.0/2.7	0.4/1.5/**2.8**
Margin	3.3/10.3/**12.0**	3.5/12.5/**14.5**	0.3/1.1/**1.3**	0.5/1.4/1.7	0.7/2.1/3.1	0.5/1.7/2.7
Angle	0.4/0.7/0.5	0.6/0.4/0.5	0/0/0.1	0.8/1.8/**2.6**	0.8/2.2/3.0	0.9/1.7/2.4

For the Logit loss [26], we can obtain the gradient w.r.t. to input feature $\phi(\hat{x})$ in Eq. 2 as:

$$\frac{\partial L_{Logit}}{\partial \phi(\hat{x})} = -W_t. \tag{16}$$

From the Eq. 15 and Eq. 16, we can observe the gradient of Cross-Entropy with a large T is approximate $\frac{1}{T}$ of the gradient of the Logit loss [26]. On the other aspect, the I-FGSM is used for optimization,

$$\hat{x}_{i+1} = \hat{x}'_i + \alpha \cdot \text{sign}(\nabla_{\hat{x}} L(\hat{x}'_i, y)), \tag{17}$$

which only considers the sign of the gradient. Therefore, Eq. 16 and Eq. 15 will update the perturbation in a similar direction. Based on the above analysis, we can consider the Logit loss function as a special case of Temperature-based calibration with a large T. Besides, we compare the performance between the Logit and CE ($T = 50$ & $T = 100$) in the following experimental section.

V. Experiments

A. Experimental Setup.

In this section, we evaluate the effectiveness of logit calibration for improving transferable targeted adversarial attacks. Following the recent study [26], we conduct the experiments on the difficult ImageNet-Compatible Dataset[1]. This dataset contains 1,000 images with 1,000 unique class labels corresponding to the ImageNet dataset [39]. We implement our methods based on the source code[2] provided by the Logit [26]. The same four diverse CNN models are used for evaluation, *i.e.*, ResNet-50 [40], DenseNet-121 [41], VGG-16 with Batch Normalization [1] and Inception-v3 [42]. The perturbation is bounded by $L_\infty \leq 16$. The TI [12], MI [11] and DI [15] are used for all attacks, and $||W||_1 = 5$ is set for TI. The I-FSGM is adopted for optimization with the $\alpha = 2$. The attacks are trained for 300 iterations on an NVIDIA-2080 Ti GPU. We run all the experiments 5 times and report the average targeted transfer success rates (%). Besides, we also conduct more experimental results about the ensemble attack and the real-world attack.

B. Comparison with Others in Single-Model Transfer

We first compare the proposed (temperature-based, margin-based and angle-based) logit calibrations with the original CE, Po+Trip [25], and Logit [26] in the single-model transfer task. In this task, we take one surrogate model for training and test the targeted transferability in attacking the other 3 models. As shown in Table I, the original CE loss function produces the worst performance than the Po+Trip and Logit. But after performing the logit calibration in the CE loss function, we can find a significant performance is boosted compared with the original CE. Our calibration methods can outperform the Logit, especially when using the ResNet50 and Dense121 as the surrogate model. These results indicate that continuously increasing the logit margin can significantly influence the performance of the targeted transferability. On the other aspect, we find that $T = 10$ has better performance than $T = 5$ on the VGG-16, suggesting that different models may need different T. Instead of finding the best T for a different model, the Margin-based calibration can solve the issue and reach the overall best transferability in all four models. However, we find that the Angle-based calibration is not working on the VGG16, which needs further investigation.

C. Influence of Different T in CE

In this section, we evaluate the influence of using different T in the CE loss function. The results are reported in Table II. From the Table, we can have the following observations. **1)** The scaling factor T has a significant influence on the targeted transferability. Specifically, the performance drops significantly when using a small $T = 0.5$. After increasing the T, we can observe the number of successfully attacked samples will increase. **2)** The optimal T for different model is different. For example, $T = 5$ can produce the overall best performance for ResNet50, Dense121, and Inception v3, while the VGG16 with fewer convolutional layers requires a large T to obtain better

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

[2] https://github.com/ZhengyuZhao/Targeted-Tansfer

TABLE II: The targeted transfer success rates (%) by using different T in CE loss function. (Results with 20/100/300 iterations are reported.)

Attack	Surrogate Model: ResNet50			Surrogate Model: Dense121		
	→Dense121	→VGG16	→Inc-v3	→Res50	→VGG16	→Inc-v3
Logit	31.4/64.0/71.8	23.8/55.0/62.4	3.1/8.6/10.9	17.4/38.6/43.5	13.7/33.8/37.8	2.3/6.6/7.5
T=0.5	13.2/16.0/19.5	7.1/9.5/11.0	1.2/1.8/2.4	4.2/5.0/6.2	2.5/3.5/3.2	0.6/0.9/1.1
T=1	27.0/40.2/42.7	17.4/27.6/29.1	2.3/4.1/4.6	12.3/17.2/18.4	8.6/10.5/10.9	1.6/2.3/2.8
T=2	34.2/62.8/67.7	24.4/52.3/53.9	3.3/7.2/8.5	18.7/35.0/36.1	13.2/27.3/27.0	2.2/5.5/6.1
T=5	33.3/69.9/**77.8**	24.8/59.9/66.1	3.1/9.4/**12.2**	19.3/43.4/47.5	14.6/36.6/39.4	2.3/7.3/**8.8**
T=10	31.6/68.5/77.0	23.6/58.5/**66.4**	2.8/9.4/11.6	17.9/43.2/**49.3**	13.4/36.8/**41.5**	2.2/7.7/**8.8**
T=20	30.4/65.6/74.3	22.9/55.4/63.6	3.2/9.0/11.6	17.6/40.3/46.2	13.4/35.4/40.1	2.3/6.7/8.7
T=50	30.2/64.7/72.7	23.3/55.1/62.9	2.9/8.8/11.4	17.3/39.6/44.8	12.7/34.3/38.3	2.4/6.7/8.3
T=100	30.0/64.7/72.3	22.8/54.4/61.9	3.1/8.7/10.7	17.0/39.7/44.7	13.0/33.7/39.1	2.2/6.5/8.1

Attack	Surrogate Model: VGG16			Surrogate Model: Inc-v3		
	→Res50	→Dense121	→Inc-v3	→Res50	→Dense121	→VGG16
Logit	3.4/9.9/**11.6**	3.5/12.0/13.9	0.3/1.0/1.3	0.6/1.1/2.0	0.6/1.9/3.0	0.6/1.5/**2.8**
T=0.5	0.2/0.1/0.2	0.1/0.1/0.1	0/0/0	0.3/0.9/0.9	0.3/0.8/1.4	0.3/0.6/1.3
T=1	0.5/0.3/0.6	0.6/0.3/0.3	0/0/0.1	0.7/1.2/1.8	0.6/1.3/1.9	0.4/0.8/1.3
T=2	1.6/1.8/1.8	1.8/1.9/1.6	0.2/0.2/0.2	0.6/1.5/2.0	0.4/1.7/2.2	0.5/1.2/2.0
T=5	3.1/7.0/6.9	3.3/7.6/7.8	0.2/0.9/0.8	0.7/1.7/2.1	0.5/1.9/**3.3**	0.4/1.6/2.6
T=10	3.6/9.0/9.7	3.4/10.5/11.7	0.3/1.1/1.3	0.5/1.3/1.9	0.6/2.0/2.7	0.4/1.5/**2.8**
T=20	3.4/9.7/11.1	3.6/12.7/13.8	0.3/1.2/1.3	0.5/1.4/**2.3**	0.6/1.8/3.1	0.5/1.6/2.4
T=50	3.1/10.2/11.4	3.9/12.0/**14.5**	0.1/1.1/1.3	0.6/1.8/2.1	0.6/2.0/3.0	0.3/1.7/2.7
T=100	3.6/9.8/11.3	3.4/11.8/13.9	0.4/1.2/**1.4**	0.6/1.6/2.0	0.4/2.1/3.0	0.4/1.7/**2.8**

TABLE III: The targeted transfer success rate (%) with the ResNet-18 and VGG-19 as the surrogate models by using different T. (Results with 20/100/300 iterations are reported.)

Attack	Surrogate Model: ResNet-18				Surrogate Model: VGG-19			
	Inc-v3	ResNet-50	Dense-121	VGG-16	Inc-v3	ResNet-50	Dense-121	VGG-16
Logit	3.7/10.0/12.2	24.8/55.6/60.7	24.3/53.6/58.5	21.2/49.4/54.9	0.3/0.9/1.0	4.1/10.8/12.3	5.1/13.3/15.8	65.6/84.0/84.2
T=1	2.1/3.0/3.0	19.2/24.0/26.0	18.6/24.0/24.6	15.9/19.3/19.0	0.2/0.1/0.2	1.0/1.8/1.3	1.4/0.8/0.9	40.7/41.1/43.1
T=5	3.9/10.8/11.9	27.8/60.7/63.6	27.2/57.5/61.6	23.7/53.0/56.6	0.5/1.3/0.9	4.6/8.7/8.7	5.1/11.7/10.6	71.4/90.0/90.5
T=10	3.6/11.2/**13.2**	25.9/59.7/**66.9**	25.9/57.2/**64.2**	22.2/53.0/**59.7**	0.6/1.4/1.7	4.3/10.5/11.4	5.1/13.9/15.9	67.9/91.7/**93.1**
T=20	3.9/11.4/13.0	25.2/57.8/64.2	24.8/54.3/60.7	21.1/49.7/57.1	0.5/2.0/**1.9**	4.4/11.3/**13.0**	5.1/14.5/**17.1**	65.3/87.8/90.9
T=50	3.0/10.4/12.9	27.1/57.0/61.7	25.7/54.5/59.3	22.5/49.0/54.6	0.4/1.7/1.7	4.2/10.6/12.8	4.9/13.8/16.2	64.4/84.1/85.3

transferability. **3)** The performances are comparable when using $T = 5$ and $T = 10$ for ResNet50, Dense121, and Inception v3. This is because we use I-FSGM for optimization, which only considers the sign of the gradients. **4)** When using a large T, the performance will be similar to the Logit loss. For example, we compare the Logit with CE ($T = 50$ & $T = 100$) in the single-model transfer scenario, and we can see the performances are approximate for the targeted transfer success. Besides, in Figure 5, we visualize the trend of logits and margins using the Logit loss [26] and the CE loss ($T = 50$ & $T = 100$). We can find the trends are very similar when the adversarial samples are crafted from the ResNet-50 by the Logit and CE ($T = 50$ & $T = 100$) loss functions.

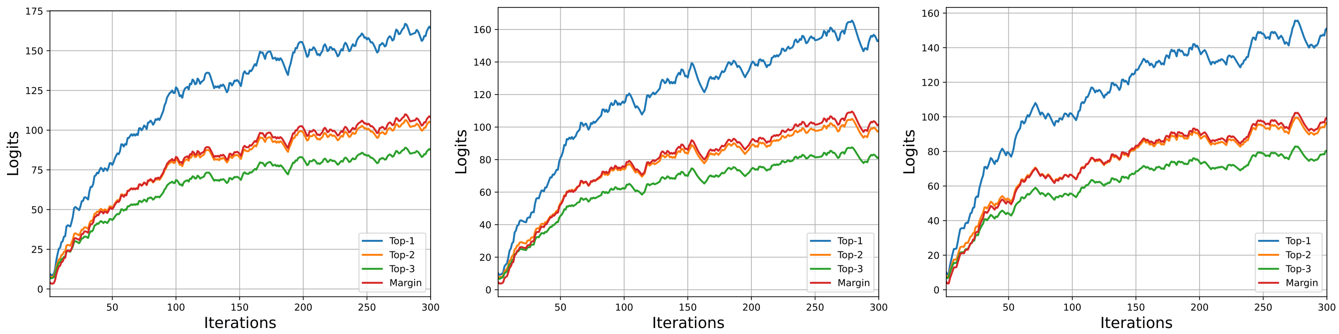
To further analyze the influence of the number of convolutional layers, we leverage the ResNet-18 and VGG-19 as the surrogate models and report the results in Table III. We can find that the overall best performance is achieved at $T = 10$ for ResNet-18 and $T = 20$ for VGG-19. Compared with $T = 5$ for ResNet-50 and $T = 50$ for VGG-16, these results suggest that a relatively large T is preferred for CNNs with few layers to achieve better performance under the same CNN architecture.

D. Influence of Different T on datasets with different size

To further investigate whether the dataset size will influence the performance, we conduct the following experiments by using the ResNet-50 as the surrogate model and varying the size of the dataset. Specifically, we randomly selected 1/5/10 images for each class in the validation set of the ImageNet dataset, and then obtained a dataset of 1K/5K/10K images covering all 1000 classes. Since there is no given targeted label for each image, we follow the worst-case transfer scenario in Logits [26] by setting the 2nd predicted class of each image as the targeted class. The results are reported in Table VI and we can find that the optimal values are consistent at T=5 for the ResNet-50 with the varying dataset size of 1K/5K/10K. These results can suggest that the size of the dataset does not affect the overall final performance. And the optimal T is mainly related to the model architecture of the training surrogate models.

E. Mutual Benefits of Different Logit Calibrations

In this part, we evaluate the mutual benefits of combining different calibrations. As shown in Table IV, we can have the following findings. **1)** Combining the $T = 5/10/20$ and Margin, there is no increase in performance compared with using one of them. This is because the gradient directions of these two methods are very similar. **2)** Combining the $T = 5$ and Angle, we can observe a further improvement when using ResNet50 and Dense121 as the surrogate model, *e.g.*, the transferable rate of "ResNet50 → Dense121" is increased to 82.4% with 300 iterations. Since the Angle obtains poor performance on VGG16, the transferable rates of corresponding combinations are also low in $T = 5/10$+Angle, but $T = 20$+Angle can deal



(a) Logit (b) CE ($T = 50$) (c) CE ($T = 100$)

Fig. 5: The average Top-3 logits and logit margin of 50 adversarial samples trained by the Logit, CE ($T = 50$) and CE ($T = 100$) loss functions for crafting the ResNet-50.

TABLE IV: The comparison of combining logit calibrations. (The targeted transfer success rates (%) with 20/100/300 iterations are reported.)

Attack	Surrogate Model: ResNet50			Surrogate Model: Dense121		
	→Dense121	→VGG16	→Inc-v3	→Res50	→VGG16	→Inc-v3
T=5 + Margin	33.8/69.8/77.2	24.0/59.0/65.5	3.3/9.6/11.1	19.3/44.3/47.8	14.1/37.7/40.8	2.5/7.5/9.4
T=5 + Angle	34.5/74.3/**82.4**	25.6/66.5/**72.2**	3.6/10.5/**13.1**	20.3/52.7/**61.9**	15.8/45.0/**53.6**	2.3/9.2/**12.7**
T=10 + Margin	32.7/69.5/77.3	22.8/59.4/66.3	12.9/9.7/11.5	18.3/44.1/49.1	13.7/36.9/41.6	2.4/8.3/9.2
T=10 + Angle	33.0/69.8/79.1	24.4/59.0/68.9	3.4/10.0/12.9	19.4/47.2/56.1	14.8/40.1/47.0	2.5/8.3/11.0
T=20 + Margin	33.0/69.2/76.2	23.1/58.4/65.8	3.2/9.5/11.8	19.1/43.4/48.5	13.9/36.7/41.4	2.4/7.8/9.5
T=20 + Angle	34.2/68.6/76.5	24.7/58.7/66.6	3.4/9.7/12.7	20.0/44.4/50.9	15.5/38.4/43.7	2.5/8.2/9.5
Margin + Angle	34.4/70.8/78.1	24.3/60.2/67.4	3.5/10.4/12.6	19.9/46.6/52.7	15.2/39.3/44.5	2.7/8.2/9.9

Attack	Surrogate Model: VGG16			Surrogate Model: Inc-v3		
	→Res50	→Dense121	→Inc-v3	→Res50	→Dense121	→VGG16
T=5 + Margin	3.5/10.2/11.4	3.7/12.4/14.6	0.3/1.1/1.3	0.5/1.4/1.6	0.6/2.1/2.9	0.5/1.7/2.8
T=5 + Angle	2.2/2.5/2.3	2.4/2.6/2.3	0.2/0.1/0.2	0.5/1.6/**2.4**	0.6/2.0/3.1	0.5/1.7/2.5
T=10 + Margin	3.2/10.7/11.7	3.4/12.9/**15.0**	0.2/1.0/**1.4**	0.5/1.4/1.9	0.5/1.9/3.0	0.3/1.5/2.3
T=10 + Angle	3.4/6.2/5.1	3.5/7.5/7.0	0.2/0.6/0.6	0.6/1.3/1.9	0.6/2.0/3.2	0.5/1.6/2.6
T=20 + Margin	3.5/10.1/**11.8**	3.4/12.0/14.9	0.3/1.2/**1.4**	0.6/1.2/1.9	0.5/1.9/2.9	0.5/1.6/2.7
T=20 + Angle	3.2/9.7/10.1	3.9/11.9/13.3	0.3/1.0/1.2	0.6/1.6/2.0	0.6/2.0/**3.5**	0.5/1.7/**2.9**
Margin + Angle	3.3/9.8/11.1	3.5/12.6/14.6	0.3/1.2/**1.4**	0.6/1.4/2.0	0.6/1.7/3.1	0.5/1.5/2.6

TABLE V: The targeted transfer success rate (%) with the ResNet-152 and VGG-19 as the surrogate models. (Results with 20/100/300 iterations are reported.)

Attack	Surrogate Model: ResNet-152				Surrogate Model: VGG-19			
	Inc-v3	ResNet-50	Dense-121	VGG-16	Inc-v3	ResNet-50	Dense-121	VGG-16
T=5+Angle	4.3/11.6/**14.2**	43.6/81.0/**86.7**	29.6/65.7/**72.6**	16.3/42.5/**49.4**	0.6/0.6/0.3	3.4/4.3/3.6	4.0/3.9/3.7	71.5/84.6/82.7
T=10+Angle	4.4/9.5/10.6	42.2/74.1/82.9	29.4/58.8/67.9	17.3/38.6/46.0	0.7/1.1/1.1	4.3/7.7/7.2	5.0/11.1/10.0	69.5/92.4/92.6
T=20+Angle	3.7/8.2/9.8	38.8/71.7/78.2	26.4/56.1/62.4	15.3/37.1/40.8	0.9/1.6/1.8	4.7/10.1/12.0	4.6/14.6/16.1	68.9/89.4/91.5
T=5+Margin	3.9/9.9/10.2	38.2/74.1/78.7	27.1/51.7/61.8	13.4/35.4/40.7	0.4/1.8/1.9	4.6/9.5/13.0	5.5/14.6/16.8	68.9/90.9/92.6
T=10+Margin	3.9/9.1/10.8	37.7/72.7/80.1	25.8/53.3/60.5	12.1/35.4/40.4	0.5/1.5/1.6	4.4/11.7/13.0	5.2/15.4/17.4	68.5/91.2/**92.9**
T=20+Margin	4.3/9.7/9.6	37.5/71.2/77.4	25.6/53.1/61.6	15.1/34.1/41.3	0.8/1.8/1.8	4.2/11.6/**13.1**	4.9/15.2/**17.5**	68.0/90.4/92.6
Angle+Margin	4.0/9.3/10.6	39.1/74.1/79.5	27.1/55.6/62.2	12.6/36.4/40.8	0.6/1.8/**2.1**	4.4/11.8/12.9	4.9/14.9/16.8	68.4/90.9/92.6

TABLE VI: Targeted transfer success rate (%) when varying the size of datasets from small to large. The targeted class for each image is its 2nd predicted class.

Attack	→DenseNet-121			→VGG-16		
	1K	5K	10K	1K	5K	10K
T=1	73.4/79.2/81.2	75.3/80.4/82.4	75.8/80.6/82.3	59.6/65.6/68.0	62.3/68.2/70.3	61.9/67.8/70.1
T=5	73.1/88.8/**92.0**	72.6/89.3/**92.6**	72.6/89.7/**92.6**	68.7/85.6/**88.8**	68.8/86.6/**89.9**	68.2/86.0/**89.6**
T=10	65.4/82.9/89.8	67.7/84.3/90.2	68.4/84.7/90.4	64.5/80.6/86.8	64.9/81.5/86.7	65.1/81.4/87.2
T=20	66.1/79.1/83.0	67.0/80.1/84.2	67.4/80.0/84.8	65.6/77.9/81.6	64.2/76.4/81.3	64.4/77.9/81.7

with this issue. **3)** Combining the Margin and Angle, there are only slight improvements on ResNet50 and Dense121, while it can alleviate the negative effects caused by the angle-based calibration.

To further investigate the best combination of different models, we use the ResNet-152 and VGG-19 models as surrogate models to generate the adversarial examples. From Table V, we can observe that "T=5+Angle" is most effective in the ResNet-152 model, which is consistent with the ResNet-50. For the VGG-19 model, combining the other two logit calibration methods (Angle and T=5/10/20) and Margin can achieve similar performance, which is the same as the observations of the VGG-16.

Finally, by jointly considering all the experimental results of single calibration and combined calibration in Table I, II, III, IV, and V, we suggest that: **1)** for the single calibration, it prefers to use the Margin-based calibration. **2)** For different CNNs with the same architecture, the one with few convolutional layers usually requires a relatively large T for achieving better-targeted transferability. **3)** For the ResNet and DenseNet architectures, the combination of T=5 + Angle is preferred for further improving the performance. **4)** For the VGG and Inception architectures, combining Margin with the other two logit calibration methods (Angle and T=10/20) can achieve similar performance, which is slightly higher than only using single Margin-based calibration.

F. Targeted Transfer of Ensemble Attacks

Since ensemble attacks can significantly improve transferability, we also use an ensemble of white-box source models to optimize the perturbations for further evaluation. Following the realistic ensemble setting in the Logit [26], we also select the four models—ResNet50, DenseNet121, VGG16, and Inc-v3, where each hold-out target model shares no similar architecture with the other three source models used for the ensemble. Besides, we simply assign equal weights to the source models during the training.

The results are shown in Figure 6, and we can obtain the following observations. **1)** The Po+Trip produces the worst performance than the original CE and the Logit in the realistic ensemble setting. Besides, the original CE is also weak for ensemble transfer tasks. **2)** After performing the logit calibration (CE (T = 5) and Margin), we can find a significant performance is boosted compared with the original CE. **3)** Meanwhile, our methods can outperform the Logit [26], comparing all results. These results verify the effectiveness of logit calibration in the ensemble targeted attacks.

G. Transfer Success Rates with Varied Targets

Following the settings in [26], we further evaluate the performance of different calibration methods in the worse-case transfer scenario by gradually varying the target class from the highest-ranked to the lowest one. The ResNet-50 and the DesNet-121 models are used for training and evaluation, respectively. From Table VII, we can have the following findings: **1)** The three types of logit calibration methods can improve the targeted transfer success rate over the original CE. The angle-based calibration can achieve the best performance. **2)** The Temperature-based ($T = 5/10$) and the Angle-based calibrations can outperform the Logit loss by a large margin, especially the Angle-based calibration. **3)** The margin-based calibration doesn't work well in this setting. When the target class is after the 200th, the performance of margin-based will be significantly lower than the Temperature-based and Angle-based calibrations.

TABLE VII: Targeted transfer success rate (%) when varying the target from the high-ranked class to low.

	2nd	10th	200th	500th	800th	1000th
Logit	83.7	83.2	74.5	71.5	64.9	52.4
CE	77.4	58.6	26.9	23.7	16.7	7.0
CE/5	91.3	88.7	77.1	75.8	70.1	58.8
CE/10	89.0	87.8	81.0	79.2	73.5	62.5
Margin	87.4	81.7	61.3	51.6	43.1	23.0
Angle	92.4	89.1	80.3	79.2	76.1	66.3

H. Transfer-based Attacks on Google Cloud Vision

Following the evaluation protocol in [26], we randomly select 100 images to conduct a real-world adversarial attack on the Google Cloud Vision API. The attacking performance is computed based on transfer-based attacks of the ensemble of four CNNs (*i.e.*, Inc-V3, ResNet-50, Dense-121 and VGG-16). The results are shown in Table VIII. We can find that the results of the Logit and CE (T = 5) are very similar. But the Margin-based calibration performs worse than Logit and CE (T = 5). These results reveal that our logit calibration-based targeted transfer attacks can cause a potential threat to the real-world Google Cloud Vision API. Besides, we present several successful targeted adversarial images with CE (T = 5) in Figure 7. We can find that the adversarial images can achieve well-targeted transferability with quasi-imperceptible perturbations.

TABLE VIII: Non-targeted and targeted transfer success rates (%) on Google Cloud Vision API.

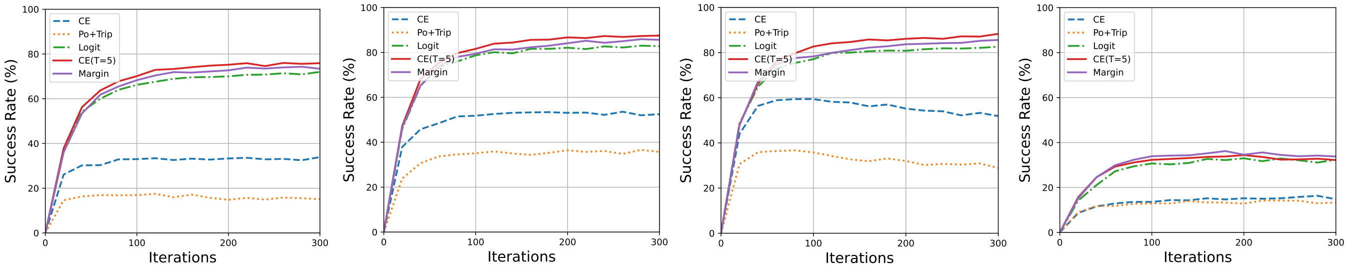
	Logit	CE (T=5)	Margin
Targeted	16	15	12
Non-targeted	51	53	42

I. Evaluating Logits Calibration in TTP and ODI

In this section, we further evaluate the proposed temperature-based logit calibration in the GAN-based targeted attacks (TTP [30]) and the most recently published object-based diverse input (ODI) method [16].

TTP: Following the setting in TTP [30], we sample 50K images from the ImageNet training set and 50K images from the Painting dataset[3], which are used to train the targeted generators from different source domains. Instead of using the distribution matching and neighborhood similarity matching loss [30], we only use the cross-entropy function for training the targeted generators while keeping other settings identical. More training and evaluation details used by TTP can be referred to [30]. We use the ResNet50 as the surrogate model and report the results in Table IX. We have the following findings from Table IX,. **1)** By using ImageNet as the training dataset, the TTP shows better transferability than the CE in attacking other black-box models. The average targeted accuracy of TTP is around 3% higher than that of CE. **2)** After downscaling the logit by 5 in the CE loss function (T = 5), we can observe a significant boost of the Top-1 targeted accuracy for all models, reaching an average value of 91.98% (ImageNet). For both ImageNet

[3] https://www.kaggle.com/c/painter-by-numbers

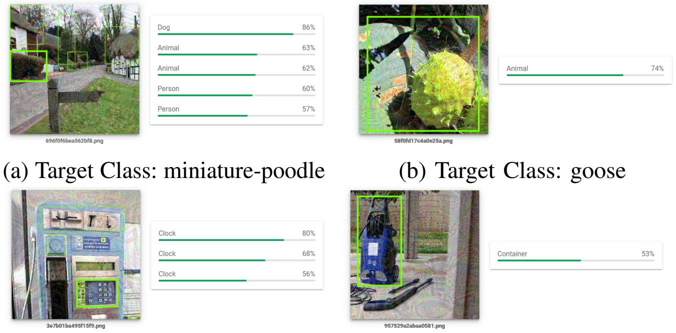


(a) ResNet-50 (b) DenseNet-121 (c) VGG-16 (d) Inc-v3

Fig. 6: Targeted transfer success rates (%) in a realistic ensemble setting, where each hold-out target model shares no similar architecture with the source models used for the ensemble. We select four models—ResNet50, DenseNet121, VGG16 and Inc-v3. In each sub-figure, the caption indicates the name of the hold-out model, and the adversarial examples are generated for the ensemble of the other three models.

TABLE IX: **Comparison with TTP [30] on Target Transferablity.** The averaged Top-1 targeted accuracy (%) across 10 targets is computed with 49.95K ImageNet validation samples. Perturbation budget: $l_\infty \leq 16$. * indicates the training surrogate model.

Dataset	Loss	ResNet50*	VGG19$_{BN}$	Dense121	ResNet152	WRN-50-2	Average
ImageNet	TTP	97.02*	78.15	81.64	80.56	78.25	83.12
	CE	97.15*	70.44	78.96	76.22	78.24	80.20
	CE (T=5)	**99.18***	86.65	90.55	90.30	93.22	91.98
	Margin	98.85*	80.76	87.32	86.79	89.86	88.72
	Angle	99.17*	**93.09**	**93.62**	**93.19**	**94.37**	**94.69**
Painting	TTP	96.63*	73.09	84.76	76.27	75.92	81.33
	CE (T=5)	**98.95***	**82.97**	**87.07**	**87.81**	**91.70**	**89.70**



(a) Target Class: miniature-poodle (b) Target Class: goose

(c) Target Class: analog clock (d) Target Class: measuring-cup

Fig. 7: Successful sample targeted adversarial images on Google Cloud Vision API (object detection) generated by the CE ($T = 5$) with ensemble transfer.

and Painting as the training source, the CE ($T = 5$) can surpass the TTP by a large margin (91.98% vs. 83.12% & 89.70% vs. 81.33%). **3)** Margin-based calibration also can outperform the orginal CE and TTP, but the target transferatblity is lower than the CE ($T = 5$) and Angle-based. **4)** Angle-based calibration can achieve the best target transferability. Specifically, the top-1 targeted accuracies are higher than 93% for four different black-box CNN models. Note that, compared to TTP, our logit calibration has the benefit without any data from the target class. These experimental results demonstrate that the proposed temperate-based logit calibration is also effective in training generator-based targeted attackers.

ODI: For the ODI [16], we follow the setting in ODI, which improves the targeted transferability by projecting the adversarial examples on 3D objects and uses the Logit loss [26] for optimizing the targeted adversarial examples. To compare with the original ODI, we use the "Package, Pillow, Book" as the 3D objects and simply replace the Logit loss [26] with the CE ($T = 5$) and Margin-based calibration for training with 300 iterations. Besides, we also select the same six additional models for more comprehensive comparisons, including ResNet18 (RN-18), a lightweight model MobileNet-v2 (Mob-v2) [43], Inception ResNet-v2 (IR-v2) [44], Inception-v4 (Inc-v4) [44] and two adversarial trained models (adversarially trained Inc-v3 (Adv-Inc-v3) [29], and ensemble-adversarially trained IR-v2 (Ens-adv-IR-v2) [29]). The results are reported in Table X. We have the following observations from Table X: **1)** For the source model ResNet-50, VGG-16, and Dense-121, the CE ($T = 5$) and Margin-based calibration significantly outperform the Logit loss when using the ODI [16] for data augmentation. Besides, the Margin-based calibration can perform slightly better than the CE ($T = 5$). **2)** For the Inception-V3, there is no significant difference between the Logit, CE ($T = 5$) and Margin-based calibration.

J. Evaluating Logits Calibration in Non-Target Attack

To evaluate the performance in Non-target attacks, we conduct experiments on the CIFAR-10 dataset and the ImageNet-Compatible dataset.

For the CIFAR-10 dataset, we implement the untargeted attack based on the code provided by [45]. The ResNet-18 model serves as the white-box surrogate model for crafting perturbations using the MI-FGSM algorithm with 20 iterations. The black-box models include DenseNet, GoogLeNet, and SENet18. In Table XI, we report the untargeted success rate

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2023.3284649

TABLE X: **Comparison with the ODI-MI-TI [16] on Target Transferablity**. Targeted attack success rates (%) against nine black-box target models with the four source models. For each attack, we also report the average attack success rate (Avg.) against all black-box models.

Source:RN-50	Target model									
Attack	VGG-16	RN-18	DN-121	Inc-v3	Inc-v4	Mob-v2	IR-v2	Adv-Inc-v3	Ens-adv-IR-v2	Avg.
ODI-MI-TI	76.8	77.0	86.8	67.4	55.4	66.8	48.0	0.7	1.7	53.4
CE (T=5)	80.7	80.7	88.9	68.1	56.7	71.0	51.8	0.6	1.2	55.5
Margin	83.7	84.4	93.3	69.4	58.7	71.3	51.7	0.7	1.2	57.2
Source:VGG-16	Target model									
Attack	VGG-16	RN-18	DN-121	Inc-v3	Inc-v4	Mob-v2	IR-v2	Adv-Inc-v3	Ens-adv-IR-v2	Avg.
ODI-MI-TI	60.8	64.3	71.1	37.0	38.0	47.0	21.1	0.0	0.0	37.7
CE (T=5)	69.4	70.0	79.5	39.6	42.6	52.6	23.3	0.0	0.0	41.9
Margin	67.6	69.8	78.3	41.3	42.2	52.8	23.2	0.0	0.0	41.7
Source:DN-121	Target model									
Attack	VGG-16	RN-18	DN-121	Inc-v3	Inc-v4	Mob-v2	IR-v2	Adv-Inc-v3	Ens-adv-IR-v2	Avg.
ODI-MI-TI	64.5	63.4	71.6	53.5	46.4	44.2	38.3	0.4	0.7	42.6
CE(T=5)	66.8	67.7	74.7	56.6	47.8	47.7	42.1	0.5	1.3	45.0
Margin	69.7	68.5	77.1	56.6	47.8	49.2	40.5	0.3	0.7	45.6
Source:Inc-v3	Target model									
Attack	VGG-16	RN-18	DN-121	Inc-v3	Inc-v4	Mob-v2	IR-v2	Adv-Inc-v3	Ens-adv-IR-v2	Avg.
ODI-MI-TI	15.7	14.7	17.4	30.4	32.1	14.1	26.9	0.3	0.6	16.9
CE(T=5)	15.0	16.2	18.7	34.5	31.5	15.3	27.8	0.2	0.5	17.7
Margin	13.5	15.3	16.6	31.4	31.3	13.2	24.9	0.3	0.3	16.3

when attacking the CIFAR-10 testing set, which consists of 10,000 images. The following observations can be made: **1)** After using T, the untargeted success rates of attacking the white-box ResNet-18 and other black-box CNNs are increased. For the white-box ResNet-18, the attack success rate continually increases along with the T. For the black-box CNNs, the best performance is achieved at T=5. **2)** The Angle-based calibration shows the overall best transfer untargeted success rate and the Margin-based calibration also leads to a slight increase. However, we also notice that Margin-based and Angle-based will decrease the performance of the white-box ResNet-18. In summary, these findings provide supporting evidence for the effectiveness of logit calibration in non-targeted attacks on a small-scale dataset.

For the ImageNet-Compatible dataset, we follow the setting [11] to implement the untargeted attack. The ResNet-50 is used as the white-box model for crafting the adversarial examples using the MI-FGSM algorithm with 20 iterations. The black-box CNNs include DenseNet-121, VGG-16, and Inc-V3. From the results in Table XII, we can find that: **1)** When T=5, the transfer untargeted success rates are increased on three black box models. However, when T is further increased to 10 or 20, the transfer untargeted success rates will decrease. **2)** The Margin-based calibration improves the untargeted success rate in all black-box CNNs. The Angle-based calibration is effective for transfer attacks on Inc-v3. In summary, these results suggest that logit calibration can improve the performance of non-targeted attacks to a certain degree.

TABLE XI: The transfer untargeted success rate of training with ResNet-18 and testing by the DenseNet-121, GoogleNet and SENet-18 on CIFAR-10.

Attack	ResNet-18*	DenseNet-121	GoogleNet	SENet-18
CE	89.89*	61.92	50.96	62.21
T=5	89.14*	63.98	54.59	63.97
T=10	90.39*	63.55	54.46	63.75
T=20	**90.67***	63.08	54.26	63.35
Margin	88.38*	62.77	54.18	62.57
Angle	87.70*	**64.45**	**56.12**	**64.46**

TABLE XII: The transfer untargeted success rate of training with ResNet-50 and testing by the DenseNet-121, VGG16 and Inc-v3 on Imagenet.

Attack	ResNet-50*	DenseNet-121	VGG-16	Inc-v3
CE	**100.0***	85.2	82.7	51.3
T=5	**100.0***	86.8	83.6	52.5
T=10	**100.0***	85.2	82.6	52.4
T=20	**100.0***	86.0	82.7	51.5
Margin	**100.0***	**87.4**	**84.6**	52.8
Angle	**100.0***	85.6	82.3	**53.5**

K. Evaluating Logits Calibration in ViT and MLP architectures

We evaluate the targeted transferability of the proposed method in the setting where both the surrogate and victim models are either ViTs or MLPs. For the ViT, we select the ViT-B/16 [46] as the white-box surrogate model and the DeiT-B [47] as the black-box victim model. For the MLP, we use the MLP-Mixer with SiGLU (MLP-Glu/24) [48] as the white-box model, and the ResMLP/24 [49] as the black-box model. Note that the adversarial examples will overfit the white-box model when the training iterations are more than 20 in the ViT and MLP. Therefore, we report the results with 5/10/20 iterations.

From Table XIII, we can find the following conclusions: **1)** Whether it is ViT or MLP, all methods can achieve superior performance in the white-box models, but the transferability performance degrades significantly due to the architectural differences between the traditional CNNs and the ViT and MLP. **2)** Compared with the CE and Logit, our CE (T=5) and

TABLE XIII: The targeted transfer success rates (%) in the single-model transfer scenario. (Results with 5/10/20 iterations are reported and the highest one at 20 iterations is shown **bold**).

Attacks	ViT		MLP	
	ViT-B/16 *	DeiT-B	MLP-Glu/24 *	ResMLP/24
CE	79.9/99.7/**100.0**	0.5/1.1/1.1	96.8/100.0/**100.0**	1.6/3.0/2.7
Logit	80.7/99.7/**100.0**	0.6/1.6/1.1	95.1/100.0/**100.0**	2.0/2.7/3.0
CE/5	82.0/99.8/**100.0**	0.7/2.2/**1.7**	95.8/100.0/**100.0**	1.6/2.7/3.5
Margin	83.1/99.9/**100.0**	0.7/2.0/1.5	96.4/100.0/**100.0**	1.8/3.5/**3.8**

Margin method can slightly enhance the targeted transferability, which further verifies our method's effectiveness.

L. Evaluating Logits Calibration on Defense Models

We evaluate the targeted transferability of our approach against the ResNet-50 trained from different defense strategies including Augmix [50], Stylized [51], Adversarial [52] with l_2 norm. Besides, we also use the Neural Representation Purifier (NRP) [53] to cure the infection of adversarial perturbations before feeding the images into the normal trained ResNet-50 model. We report the untargeted and targeted success rates (UnTar and Tar) in Table XIV.

From Table XIV, we can have the following observations. **1)** Compared to the original CE, the Logit loss leads to a significant improvement in the untargeted and targeted success rates against all defense strategies. **2)** The T=5/10 and Margin-based calibrations can achieve similar performance, which outperforms the CE, Po+Trip, and Logit on the targeted attack. **3)** The Angle-based calibrations exhibit lower performance than other logit calibrations. However, its results are still higher than the original CE loss. Finally, these results can support the effectiveness of logit calibration in improving the untargeted and targeted success rates of attacks against defense strategies.

TABLE XIV: The performance (Untargeted and Targeted success rates) against the ResNet-50 trained from different defense strategies including Stylized (SIN-IN) [51], Augmix [50], NRP [53], and Adversarial [52].

Attack	SIN-IN		Augmix		NRP		Adversarial	
	UnTar	Tar	UnTar	Tar	UnTar	Tar	UnTar	Tar
CE	90.9	65.9	66.8	15.5	57.0	7.8	53.2	3.5
Po+Trip	86.5	69.9	61.7	19.6	49.0	5.8	47.3	5.6
Logit	97.6	85.6	**81.2**	37.9	63.5	11.8	63.9	11.0
T=5	97.2	**91.2**	79.6	39.7	63.0	12.6	63.9	**13.4**
T=10	97.4	90.6	80.9	**42.5**	64.7	**13.8**	64.2	12.8
Margin	**97.7**	90.3	80.7	40.8	**65.7**	13.5	**65.3**	13.2
Angle	94.4	81.1	73.2	31.0	54.2	6.2	57.6	9.5

VI. Conclusion

In this study, we analyze the logit margin in different loss functions for the transferable targeted attack and find that the margin will quickly get saturated in the CE loss and thus limits the transferability. To deal with this issue, we propose to use logit calibrations in the CE loss function, including Temperature-based, Margin-based, and Angle-based. Experimental results verify the effectiveness of using the logit calibration in the CE loss function for crafting transferable targeted adversarial samples. The proposed logit calibrations are simple and easy to implement, which can achieve state-of-the-art performance in transferable targeted attacks. Besides, the logit calibrations also can improve the performance of non-targeted attack to a certain degree.

Potential Social Impact

Our findings in targeted transfer attacks can potentially motivate the AI community to design more robust defenses against transferable attacks. In the long run, it may be directly used for suitable social applications, such as protecting privacy. Contrariwise, some applications may use targeted transferable attacks in a harmful manner to damage the outcome of AI systems, especially in speech recognition and facial verification. Finally, we firmly believe that our investigation in this study can provide valuable insight for future researchers of using logit calibration for both adversarial attack and defense.

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015. 1, 6

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015. 1

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NeurIPS*, 2015. 1

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015. 1, 2

[5] Q. Wang, B. Zheng, Q. Li, C. Shen, and Z. Ba, "Towards query-efficient adversarial attacks against automatic speech recognition systems," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 896–908, 2020. 1

[6] J. Wu, B. Chen, W. Luo, and Y. Fang, "Audio steganography based on iterative adversarial attacks against convolutional neural networks," *IEEE transactions on information forensics and security*, vol. 15, pp. 2282–2294, 2020. 1

[7] Y. Zhong and W. Deng, "Towards transferable adversarial attack against deep face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1452–1466, 2020. 1

[8] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu, "Efficient decision-based black-box adversarial attacks on face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7714–7722. 1

[9] W. Ding, X. Wei, R. Ji, X. Hong, Q. Tian, and Y. Gong, "Beyond universal person re-identification attack," *IEEE transactions on information forensics and security*, vol. 16, pp. 3442–3455, 2021. 1

[10] F. Yang, J. Weng, Z. Zhong, H. Liu, Z. Wang, Z. Luo, D. Cao, S. Li, S. Satoh, and N. Sebe, "Towards robust person re-identification by defending against universal attackers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[11] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *CVPR*, 2018. 1, 2, 6, 11

[12] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *CVPR*, 2019. 1, 2, 6

[13] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *ICML*, 2019. 1

[14] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *ICLR*, 2018. 1

[15] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *CVPR*, 2019. 1, 2, 6

[16] J. Byun, S. Cho, M.-J. Kwon, H.-S. Kim, and C. Kim, "Improving the transferability of targeted adversarial examples through object-based diverse input," in *CVPR*, 2022. 1, 3, 9, 10, 11

[17] X.-C. Li, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Decision-based adversarial attack with frequency mixup," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1038–1052, 2022. 1

[18] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *ICLR*, 2016. 1

[19] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, “Nesterov accelerated gradient and scale invariance for adversarial attacks,” in *ICLR*, 2020. 1

[20] Q. Huang, I. Katsman, H. He, Z. Gu, S. Belongie, and S.-N. Lim, “Enhancing adversarial example transferability with an intermediate level attack,” in *CVPR*, 2019. 1

[21] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, and X. Ma, “Skip connections matter: On the transferability of adversarial examples generated with resnets,” in *ICLR*, 2020. 1

[22] Y. Guo, Q. Li, and H. Chen, “Backpropagating linearly improves transferability of adversarial examples,” in *NeurIPS*, 2020. 1

[23] Q. Xu, G. Tao, and X. Zhang, “Bounded adversarial attack on deep content features,” in *CVPR*, 2022. 1

[24] Y. Xiong, J. Lin, M. Zhang, J. E. Hopcroft, and K. He, “Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability,” in *CVPR*, 2022. 1

[25] M. Li, C. Deng, T. Li, J. Yan, X. Gao, and H. Huang, “Towards transferable targeted attack,” in *CVPR*, 2020. 1, 2, 3, 6

[26] Z. Zhao, Z. Liu, and M. Larson, “On success and simplicity: A second look at transferable targeted attacks,” *NeurIPS*, vol. 34, 2021. 1, 2, 3, 4, 5, 6, 7, 9, 10

[27] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network (2015),” *arXiv preprint arXiv:1503.02531*, 2015. 2, 4

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *CVPR*, 2015. 2

[29] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” in *Artificial intelligence safety and security*, 2018, pp. 99–112. 2, 10

[30] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, “On generating transferable targeted perturbations,” in *ICCV*, 2021. 2, 3, 9, 10

[31] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” *arXiv preprint arXiv:1611.01236*, 2016. 2

[32] N. Inkawhich, K. J. Liang, L. Carin, and Y. Chen, “Transferable perturbations of deep feature distributions,” *arXiv preprint arXiv:2004.12519*, 2020. 3

[33] N. Inkawhich, K. Liang, B. Wang, M. Inkawhich, L. Carin, and Y. Chen, “Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability,” *NeruIPS*, 2020. 3

[34] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, “Generative adversarial perturbations,” in *CVPR*, 2018. 3

[35] M. M. Naseer, S. H. Khan, M. H. Khan, F. Shahbaz Khan, and F. Porikli, “Cross-domain transferability of adversarial perturbations,” *NeurIPS*, 2019. 3

[36] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *CVPR*, 2017. 5

[37] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *CVPR*, 2019. 5

[38] F. Croce and M. Hein, “Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks,” in *ICML*. PMLR, 2020, pp. 2206–2216. 5

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR*. Ieee, 2009, pp. 248–255. 6

[40] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016. 6

[41] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *CVPR*, 2017. 6

[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *CVPR*, 2016. 6

[43] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520. 10

[44] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017. 10

[45] Q. Huang, I. Katsman, H. He, Z. Gu, S. Belongie, and S.-N. Lim, “Enhancing adversarial example transferability with an intermediate level attack,” in *ICCV*, 2019. 10

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020. 11

[47] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357. 11

[48] R. Wightman, “Pytorch image models,” https://github.com/huggingface/pytorch-image-models, 2019. 11

[49] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, G. Izacard, A. Joulin, G. Synnaeve, J. Verbeek *et al.*, “Resmlp: Feedforward networks for image classification with data-efficient training,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 11

[50] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, “Augmix: A simple data processing method to improve robustness and uncertainty,” *arXiv preprint arXiv:1912.02781*, 2019. 12

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

[52] H. Salman, A. Ilyas, L. Engstrom, A. Kapoor, and A. Madry, “Do adversarially robust imagenet models transfer better?” *Advances in Neural Information Processing Systems*, vol. 33, pp. 3533–3545, 2020. 12

[53] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, “A self-supervised approach for adversarial robustness,” in *CVPR*, 2020, pp. 262–271. 12

Juanjuan Weng is currently a Ph.D. candidate at Xiamen University, China. She received her bachelor’s degree in Computer Science and Technology from Henan University of Science and Technology, China, in 2019. Her research interests include person re-identification and deep adversarial learning.

Zhiming Luo received the B.S. degree from the Cognitive Science Department, Xiamen University, Xiamen, China, in 2011; the Ph.D. degree in computer science with Xiamen University and University of Sherbrooke, Sherbrooke, QC, Canada, in 2017. He was a postdoc and is an Associate Professor at Xiamen University, China. His research interests include surveillance video analytic, computer vision, and machine learning.

Shaozi Li received the B.S. degree from Hunan University, and the M.S. degree from Xi’an Jiaotong University, and the Ph.D. degree from National University of Defense Technology. He is currently a full professor in the Department of Artificial Intelligence, Xiamen University, the Vice Director of Technical Committee on Collaborative Computing of CCF, the Vice Director of the Fujian Association of Artificial Intelligence. He is also the senior Member of IEEE, ACM and China Computer Federation (CCF). His research interests cover Artificial Intelligence and Its Applications, Moving Objects Detection and Recognition, Machine Learning, Computer Vision, Multimedia Information Retrieval, etc.

Nicu Sebe is Professor in the University of Trento, Italy, where he is leading the research in the areas of multimedia analysis and human behavior understanding. He was the General Co-Chair of the IEEE FG 2008 and ACM Multimedia 2013. He was a program chair of ACM Multimedia 2011 and 2007, ECCV 2016, ICCV 2017 and ICPR 2020. He is a general chair of ACM Multimedia 2022 and a program chair of ECCV 2024. He is a fellow of IAPR.

Zhun Zhong received his Ph.D. degree in the Department of Artificial Intelligence of Xiamen University, China, in 2019. He was also a joint Ph.D. student at University of Technology Sydney, Australia. He was a postdoc and is an Assistant Professor at University of Trento, Italy. His research interests include person re-identification, novel class discovery, data augmentation and domain adaptation.