# Investigating Top-$k$ White-Box and Transferable Black-box Attack

Chaoning Zhang, Philipp Benz, Adil Karjauv, Jae Won Cho, Kang Zhang, In So Kweon
Korea Advanced Institute of Science and Technology (KAIST)
chaoningzhang1990@gmail.com

## Abstract

*Existing works have identified the limitation of top-$1$ attack success rate (ASR) as a metric to evaluate the attack strength but exclusively investigated it in the white-box setting, while our work extends it to a more practical black-box setting: transferable attack. It is widely reported that stronger I-FGSM transfers worse than simple FGSM, leading to a popular belief that transferability is at odds with the white-box attack strength. Our work challenges this belief with empirical finding that stronger attack actually transfers better for the general top-$k$ ASR indicated by the interest class rank (ICR) after attack. For increasing the attack strength, with an intuitive analysis on the logit gradient from the geometric perspective, we identify that the weakness of the commonly used losses lie in prioritizing the speed to fool the network instead of maximizing its strength. To this end, we propose a new normalized CE loss that guides the logit to be updated in the direction of implicitly maximizing its rank distance from the ground-truth class. Extensive results in various settings have verified that our proposed new loss is simple yet effective for top-$k$ attack. Code is available at:* https://bit.ly/3uCiomP

## 1. Introduction

Deep neural networks (DNNs) are widely known to be vulnerable to adversarial examples [15,16,39,43], which are crafted by adding imperceptible or quasi-imperceptible perturbations to natural images. This intriguing phenomenon has inspired a vibrant field of studying the model robustness [30, 32, 36, 40]. One intriguing property of adversarial examples is the widely known transferability from one (surrogate) model to another (target) model [12,25]. This property has been exploited for the transferable black-box attack as well as enhancing query-based black-box attack [19].

It is widely reported that I-FGSM increases the attack strength of FGSM, but at the cost of a lower transfer rate. This leads to a popular belief that the white-box strength of an attack is at odds with its transferability [22]. Lower transfer rates of I-FGSM are often attributed to the conjecture that longer iterations lead to over-fitting to the surrogate model [6, 22]. Partly due to this concern, conventionally, existing works on transferable attack often adopt a limited number of iterations $T$, typically set to $\epsilon/\alpha$ where $\epsilon$ and $\alpha$ are the maximum $L_\infty$ budget and step size, respectively. In contrast, we show that this phenomenon can be at least partially explained by the lower perturbation magnitude of I-FGSM and a larger $T$ improves the transferability, eventually outperforming FGSM given a sufficiently large $T$. We further demonstrate that complementary to existing techniques, increasing $T$ consistently enhances the transferability and then saturates to a plateau.

Conventionally, attack success rate (ASR), also called fooling ratio (FR), is commonly used for evaluating strength in white-box, and transferability in black-box attacks. However, ASR does not provide an in-depth indication of attack strength. In essence, ASR only indicates whether an interest class, ground-truth class in the non-targeted or target class in targeted setting, ranks top-1 in the adversarial example. It would be interesting to know the ASR@$k$, *i.e.* beyond from top-1 to top-$k$, to have a wide-range evaluation of attack strength. To this end, we introduce a new metric termed *interest class rank (ICR)*, which facilitates the ASR@$k$ evaluation and, more importantly constitutes a single unified value indicating the *top-$k$ attack strength*.

Increasing $T$ enhances both top-$k$ adversarial strength and transferability, suggesting top-$k$ attack strength is also transferable. However, simply increasing the $T$ is not enough to lead to a strong top-$k$ attack. We identify that the reason lies in the commonly used cross-entropy (CE) loss or C&W loss which prioritize the speed of fooling the network instead of maximizing its distance from the interest (ground-truth) class. To this end, we propose Relative Cross-Entropy (RCE) loss for boosting stronger top-$k$ attack. Our new loss achieves close-to-optimal top-$k$ strength in white-box attack, outperforming existing losses by a large margin, consequently leading to a stronger top-$k$ transferable attack.

**Contributions.** Our work is the first to attempt the task of *top-$k$ transferable attack*. A major obstacle towards this task is a popular belief on strength and transferability, which we challenge by showing that increasing $T$ enhances both and that top-$k$ attack strength is transferable. We identify the limitation of existing losses and propose a new loss for achieving a strong top-$k$ attack in both white-box and transferable black-box settings. We extensively validate its efficacy for benchmarking top-$k$ strength and transferability of adversarial examples on multiple datasets. Our proposed ICR metric for evaluating top-$k$ attack also provides a unified perspective for non-targeted and targeted setups.

## 2. Related work

**Beyond attack success rate.** Although attack success rate (ASR) is a popular metric for evaluating the attack strength, its limitation comes in that it only shows whether the interest class, ground-truth class in the typical non-targeted setting, ranks top-1 after an attack. This limitation has been first noted in [21]. To this end, ASR@$k$ (with a different term) has been introduced in [29] has been introduced. For a given $k$, an attack is successful if the rank of the ground-truth class is larger than or equal to $k$. When $k$ is larger than 1, the attack is strictly more difficult than adopting the conventional ASR, *i.e*. ASR@1, as the metric. In other words, if an attack is successful under ASR@$k$, it is guaranteed that it is a successful attack with the conventional ASR, but not vice versa. Extended to the targeted setting, an attack is successful if the rank of interest class, *i.e*. target class, is smaller than the given $k$. When $k$ increases, task complexity of non-targeted and targeted settings increases and decreases, respectively. [9] has also proposed alternative metrics, such as old label new ranking (OLNR), new label old ranking (NLOR), cosine similarity (CosSim), normalized rank transformation (NRT). Complementary to their metrics, our work introduces a straightforward metric, interest class rank (ICR), to indicate the rank of the interest class after the attack. A major merit of ICR is that it can be directly transformed to ASR@$k$ for any $k$.

**Transferability and Black-box Attacks.** Various works have attempted to explain transferability from different perspectives. For example, [10] attributes it to the hypothesis on the linear nature of modern DNNs, which has been recently supported by the recent finding that backpropagating more linearly improves transferability [11, 38]. Understanding transferability from the perspective of pixel interaction [37] has also been investigated. Through the lens of non-robust feature [17], a recent work [3] has shown that adversarial tranferability can be improved by removing BN from the surrogate model. Even though the rationality behind transferability is still not fully understood, this intriguing property has been widely exploited for black-box attacks. Early works have shown that adversarial examples naively generated in the direct white-box manner, such as vanilla I-FGSM, have low transferability. An ensemble of multiple surrogate models is found to improve the transferability [26, 35] but at the cost of more computation resources. Some free techniques have been proposed, such as momentum update [6], input diversity [41], and translation-invariant constraint [7]. [14,24] have demonstrated that fine-tuning adversarial examples with the intermediate level attack can further boost the transferability. Backpropagating linearly [11] or smoothly [42] is also found to improve trasnferability. Most investigation on transferable attack centers around non-targeted setting, and recently multiple works [18–20] have attempted the targeted setting via the loss optimization in feature space. This often requires training additional class-wise layer-wise auxiliary classifiers. Directly performing the loss optimization in the output space is more simple but often at the loss of lower targeted transferability. Identifying the gradient vanishing issue of the cross-entropy (CE) loss, [23] has proposed a new loss based the Poincaré ball distance.

**Positioning our work.** Our work is the first to target a strong top-$k$ transferable attack, *i.e*. increasing ASR-$k$ for a wide range of $k$ including $k = 1$. For the top-$k$ attack, prior works exclusively only study in the white-box setting. On the other hand, prior works that study transferability do not take top-$k$ into account. Interestingly, we note that both subproblems boil down to increasing attack strength. Through showing attack strength is transferable, our work aims to realize strong top-$k$ white-box attack and top-$k$ transferable attack, simultaneously. Prior techniques improve the transferability mainly through a regularization effect. Orthogonal and complementary to them, our work focuses on the influence of adversarial strength on transferability and attempt to improve it through increasing white-box attack strength.

## 3. Background and a popular belief

**White-box attacks.** White-box attack methods typically assume that the attacker has full knowledge of a target model, *i.e.* the architecture and parameters [4, 27, 33]. To make the adversarial perturbation imperceptible, the perturbation is often constrained inside a certain allowable perturbation budget or its $L_p$ is smaller than a certain magnitude, *i.e.* $||v||_p \leq \epsilon$ [33]. Under such constraint, the goal of most existing adversarial attacks is to maximize a certain loss $L(x + v, y)$ for which the CE loss is widely used.

**FGSM.** Goodfellow *et al*. proposed FGSM to craft adversarial examples: $X^{adv} = X + \epsilon sign(\nabla_X J(X, y_{true}))$, where $X^{adv}$ is the resulting adversarial example, $X$ is the attacked image, $J$ is the loss, $y_{true}$ is the ground truth label, and $\epsilon$ is the maximum allowable perturbation budget for making the resulting adversarial example look natural to the human eye. Simple FGSM achieves a reasonably high ASR.

**Single-Step Least Likely Class Method (Step-LL).** This attack can be considered as a new variant of FGSM with a loss that targets a non-ground truth class [21]:

$X^{adv} = X + \epsilon \text{sign}(\nabla_X J(X, y_{LL}))$, where $y_{LL} = \arg\min(h(X))$, indicating the least-likely (LL) class based on the model output, *i.e.* logit vector $h(X)$.

**I-FGSM or Iter-LL.** Iterative attack was introduced in [21, 22] to increase ASR by iteratively applying FGSM or Step-LL with the step size $\alpha$: $X_0^{adv} = X, X_{t+1}^{adv} = X_t^{adv} + \alpha sign(\nabla_X J(X_t^{adv}, y))$. The step size $\alpha$ is often set to $\epsilon/T$, where $T$ indicates the number of iterations, for satisfying the $L_\infty$ constraint. It has been widely reported in [6, 21, 22] that iterative attack methods induce a higher ASR than their single-step counterparts, *i.e.* FGSM or step-LL, but *transfer* at lower success rates. For example, it is argued in [22] that *"there might be an inverse relationship between transferability of specific method and ability of the method to fool the network,"* which implies that adversarial strength is at odds with transferability.

**Existing techniques for improving transferability.** Most techniques introduced in popular works for improving transferability play the role of regularization. It has been shown in [48] that adding a regularization term can non-trivially improve the transferability. This is conceptually analogous to the practice of regularizing model training to avoid over-fitting, *i.e.* slightly reducing the training accuracy, for improving the test accuracy. Other works have also introduced other implicit regularization techniques, such as gradient update with momentum [6]:

$$g_{t+1}^{adv} = \mu g_t^{adv} + \frac{\nabla_X J(X_t^{adv}, y)}{||\nabla_X J(X_t^{adv}, y)x||_1}, \qquad (1)$$

$$X_{t+1}^{adv} = X_t^{adv} + \alpha sign(g_{t+1}^{adv}).$$

where $\mu$ indicates the momentum weight, usually set to 1. The above technique is often called MI-FGSM. Another two famous variants of I-FGSM are DI-FGSM introduced in [41] and TI-FGSM in [7]. The DI-FGSM is shown as :

$$X_{t+1}^{adv} = X_t^{adv} + \alpha sign(\nabla_X J(Tr(X_t^{adv}; p), y)) \qquad (2)$$

where $Tr$ indicates transformation with the probability $p$. The TI-FGSM is shown as:

$$X_{t+1}^{adv} = X_t^{adv} + \alpha sign(W * \nabla_X J(X_t^{adv}, y)) \qquad (3)$$

where $W$ is a kernel for smoothing the gradients.

**Experimental Setup.** Following previous works [6, 7, 23], we evaluate our proposed techniques on an ImageNet-compatible dataset composed of 1000 images. This dataset was introduced in the NeurIPS 2017 adversarial challenge[1] and widely used for transferable black-box attack. Consistent with previous methods, we set the maximum perturbation magnitude to $L_\infty = 16/255$.

**Influence of $\alpha$ and $T$ on transfer rate.** The phenomenon that FGSM is more transferable is often attributed to the fact that iterative attack methods tend to over-fit to the surrogate model [6, 22]. However, it remains yet unclear which factor mainly contributes to over-fitting. Technically, the differences between I-FGSM and FGSM consist of two factors: step size $\alpha$ and number of iteration $T$. To demystify this, we analyze the influences of $\alpha$ and $T$ on the transfer rate. The results are shown in Figure 1. We have two major observations: (a) Given a fixed $\alpha$, increasing $T$ enhances the transfer rate; (b) Given a fixed $T$, increasing $\alpha$ significantly boosts the transfer rate, especially when $T$ is not sufficiently large. The results demonstrate that *the factor that contributes to the over-fitting of I-FGSM is $\alpha$ rather than $T$*. We find that given sufficiently large iterations, I-FGSM transfers better than FGSM. We report similar phenomenon in different setups in Figure 2.

**Correlation with the $L_2$ norm.** Why does increasing $\alpha$ and $T$ enhance the transferability? We identify $L_2$ norm of the perturbation as a factor that correlates with the transferability. The results in Figure 1 show that in the setup of I-FGSM, there is a high positive correlation between transfer rate and $L_2$ norm (similar trend is observed for $L_1$ norm). Nonetheless, $L_2$ is not the only influence factor, otherwise, I-FGSM can never outperform FGSM for transferability.

## 4. Top-$k$ attack and ICR metric

**Interest class rank.** With the top-$k$ metric, *i.e.* ASR@$k$, there is no end to what $k$ can be, thus alternatively we also introduce a new metric called Interest Class Rank (ICR) which directly indicates the rank of the interest class after the attack. Note that for any sample, given the ICR value, we can easily tell whether it is a successful attack at any given $k$. For example, in an untargeted setting, if the ICR is 20, the attack is successful when $k$ in ASR@k is set to 10 ($10 < ICR$) and unsuccessful when $k$ is 30 ($30 > ICR$). Thus, without the need to enumerate all $k$, the ICR with a single value indicates the full-spectrum top-$k$ attack strength. Note that ICR can be used for both attack settings, where a larger ICR indicates the attack is *stronger* in the untargeted setting and *weaker* in the targeted setting. We highlight that the ICR is equivalent to ASR@$k$, since ICR can be easily transformed top-$k$ for any $k$.

**Top-$k$ attack strength is transferable.** With the ICR as the metric, we study the new rank between the surrogate model and target model, *i.e.* whether the top-$k$ adversarial strength is transferable. Through analyzing a single sample, we observe that a higher ICR on the surrogate model also leads to a higher ICR on the target model, suggesting top-$k$ adversarial strength is transferable. Averaging on 1000 samples, we show the ICR with different $\alpha$ and $T$ and the results are shown in Figure 3. As a control study, we also report the same results with the metric of ASR-1. The overall trend of the ICR mirrors that of ASR. For example, either increasing $T$ or $\alpha$ significantly boosts

[1] https://github.com/rwightman/pytorch-nips2017-adversarial

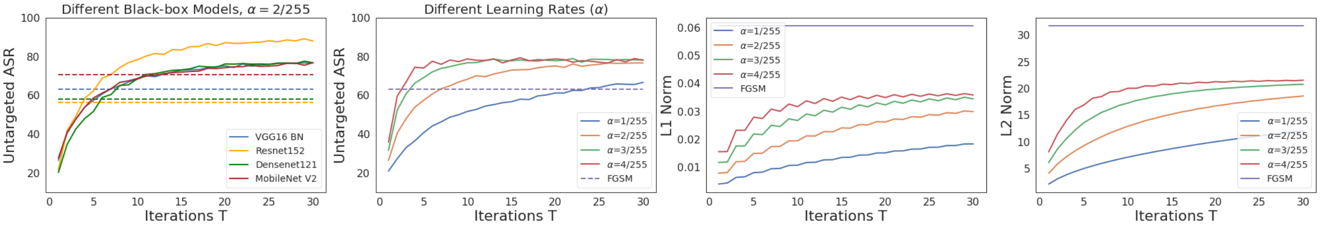

Figure 1. Transferability result for the FGSM (dashed lines) and I-FGSM (solid lines) with source network ResNet50 (RN50) and various black-box models (1st left). Performance for different step sizes ($\alpha$) when black-box model is VGG16 (2nd left). L1 (3rd left) and L2 (4th left) norms of the perturbation over the iterations. L1 norm is calculated on all pixel dimensions as an averarage their absolute values.

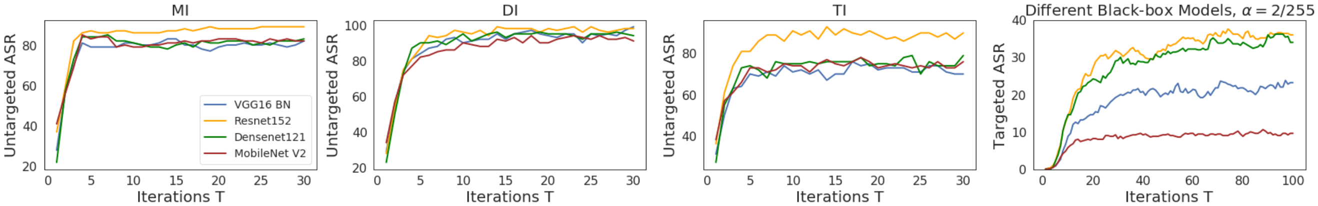

Figure 2. **First three figures from the left:** Non-targeted transferability with MI, DI, and TI. **Rightmost figure:** Targeted transferability with the MI-DI-TI-FGSM. The source network is ResNet50.

the top-$k$ adversarial strength on the target model. However, it is more challenging to get satisfactory performance with the ICR metric. With the $\alpha$ set to 1/255, even after 20 iterations, the black-box average ICR is only around 15/1000 (Maximum $K$ is 1000 for ImageNet). Adopting a higher $\alpha$ boosts convergence, however, the final ICR is still only around 40/1000. An important takeaway from the above results is that *strong top-$k$ black-box attack might be achievable through increasing the white-box top-$k$ adversarial strength*.

## 5. Boosting top-$k$ white-box attack

**One intriguing property of logit vectors.** Let the logit vector be defined as the pre-softmax output of a DNN classifier and be denoted as $\mathbf{Z}$. Here, we report that the sum of all values in the $\mathbf{Z}$ vector is very close to zero in the vast majority of cases. We confirmed this phenomenon over various networks on different datasets for both adversarial examples and natural examples. Refer to the supplementary for detailed results of this intriguing phenomenon as well as a possible explanation. Moreover, the zero-sum phenomenon indicates that the logit values in $\mathbf{Z}$ have to be internally connected to satisfy *zero-sum* constraint. In the following, we present a geometric illustration of the gradient directions of different loss functions, for which the zero-sum property of $\mathbf{Z}$ will constitute an important assumption.

**Gradient directions of common loss functions.** The influence of the loss on the generation of adversarial examples lies in the perturbation gradient update direction. Due to the extremely non-linear behavior of the network, it is intractable to intuitively derive a loss by analyzing the gradient on the network input. To alleviate such concern, we focus on tractable gradients of the logit vector. In other words, we assume that we can directly update the logits. Admittedly, we recognize that directly updating logit is not practical since we can only update the input perturbation. Nevertheless, with the backward-propagation chain rule, an optimal gradient update on the logit will lead to a pseudo-optimal update on the input perturbation. In this part we will first discuss the gradient directions with respect to the logit vector $\mathbf{Z}$ of commonly used loss functions. The detailed derivations can be found in the supplementary and here we present the main results. For the non-targeted setting, the derivative with respect to $\mathbf{Z}$ for CE, CE(LL) and CW losses are $\mathbf{P} - \mathbf{Y}_{gt}$, $\mathbf{Y}_{LL} - \mathbf{P}$ and $\mathbf{Y}_j - \mathbf{Y}_{gt}$, respectively. $\mathbf{P}$ is the post-softmax probability vector and $\mathbf{Y}_{gt}$, $\mathbf{Y}_{LL}$, $\mathbf{Y}_j$ ($j = \arg\max\limits_{i \neq gt} Z(X^{adv})_i$) indicate the ground-truth one-hot label, that of the least likely class, and that of highest class except the ground-truth class, respectively.

**Relative CE Loss.** Next, we present our new loss formulation for boosting the top-$k$ adversarial strength. The loss function, which we term Relative CE loss or $RCE$ in short is formulated as follows:

$$RCE(X_t^{adv}, y_{gt}) = CE(X_t^{adv}, y_{gt}) - \tfrac{1}{K}\textstyle\sum_{k=1}^{K} CE(X_t^{adv}, y_k), \tag{4}$$

which consists of two parts, the commonly used CE, and a normalization part, averaging the CE calculated for each class. Its gradient on the logit vector $\mathbf{Z}$ is derived as follows:

$$\frac{\partial L_{RCE}}{\partial \mathbf{Z}} = \frac{1}{K} - \mathbf{Y}_{gt}. \tag{5}$$

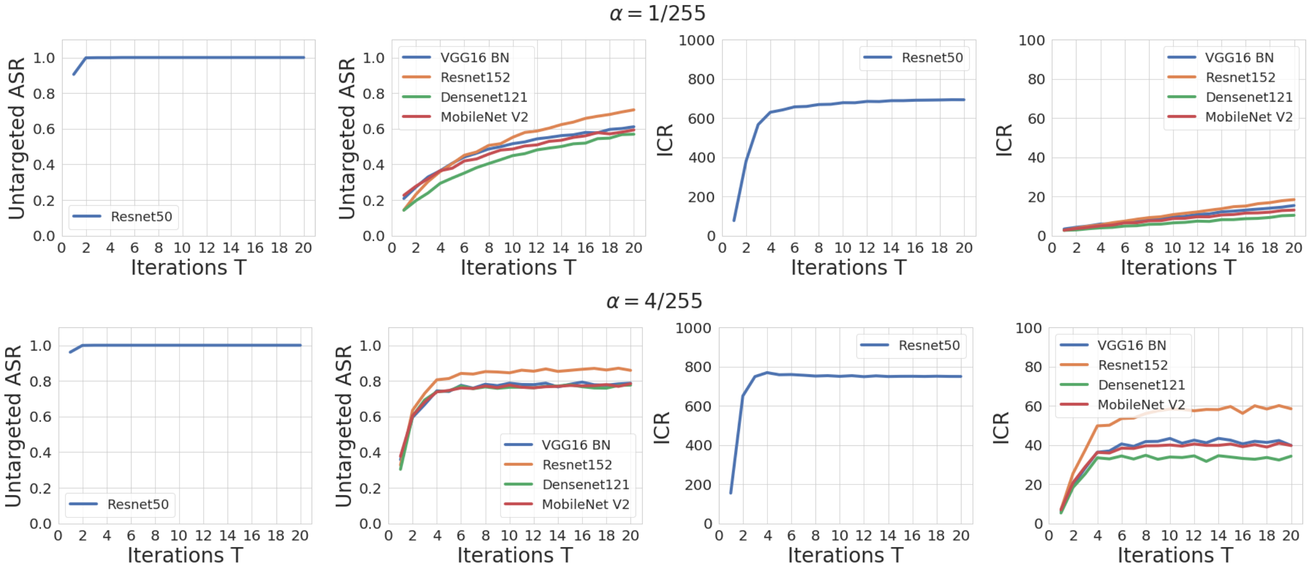

Figure 3. ICR and ASR with $\alpha$ set to 1/255 (top) and 4/255 (bottom) with ResNet50 as the surrogate (white-box) model.

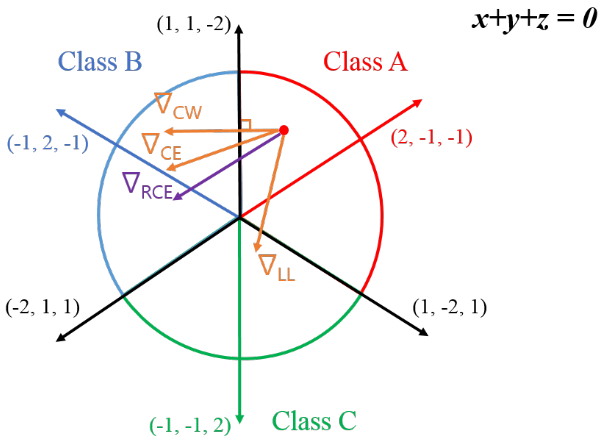

Figure 4. Geometric interpretation of the logit gradient of losses.

After establishing the gradient directions of common loss functions and the introduction of our loss function and its corresponding gradient, we now provide a geometric perspective, to illustrate why the proposed loss increases top-$k$ adversarial strength. In short, we will show that *the gradient direction of the $RCE$ loss pushes a sample most far away from its ground-truth class.*

**Geometric interpretation of the logit gradient.** For illustration purposes, our setup is designed to have only three classes A, B, C. Each class is represented by the corresponding logit value $x$, $y$, and $z$, respectively. First, we assume that there is no constraint on the logits, thus each logit is fully independent. The logit space can be represented in the 3-D space with three orthogonal axes $X$, $Y$, and $Z$. Previously, we described the zero-sum phenomenon of logit vector $\mathbf{Z}$ that the sum of logits is always very close to zero for clean samples and adversarial samples. The logits are constrained to lie on a plane of $x+y+z=0$ (with a normal vector of $(1,1,1)$), which is termed (logit) decision hyperplane. In other words, the *zero sum* constraint decreases the degree of freedom from 3-D space to a 2-D plane. We visualize this 2-D plane in Figure 4. With the symmetric assumption, the direction of the class-wise logit vector for class A, B, C can be set to $(2,-1,-1)$, $(-1,2,-1)$, $(-1,-1,2)$ with a certain scale. We highlight that vector scale is irrelevant and only the direction matters due to the sign function on the input gradient processing, *i.e.* FGSM. It is worth mentioning that the sum of the values in the $\frac{\partial L}{\partial \mathbf{Z}}$ is also always equal to zero for the above discussed three losses. Moreover, all the points on the plane satisfy $x+y+z=0$ given the *zero sum* constraint. Thus, all the discussion here is always on the decision hyperplane $x+y+z=0$. Suppose, at step $t$, the position of the sample on the decision hyperplane is $(x_t, y_t, z_t)$. Without losing generality, we assume the sample is on the region of class $A$ and $y_t > z_t$ indicating the sample is relatively more close to the logit decision boundary with $B$ instead of $C$.

To give a concrete example for facilitating the discussion, we assume $x_t = 1, y_t = 0.2, z_t = -1.2$ and the resulting post-softmax probability vector is $P = (0.64, 0.29, 0.07)$. We assume that the sample is correctly classified, hence its ground truth vector is $Y_{gt} = (1,0,0)$. Following the descriptions above $Y_{LL} = (0,0,1)$ $Y_j = (0,1,0)$, the calculated derivatives for CE, CW and CE(LL) are detailed as:

$$\frac{\partial L_{CE}}{\partial \mathbf{Z}} = \begin{pmatrix}-0.36\\0.29\\0.07\end{pmatrix};\ \frac{\partial L_{CE(LL)}}{\partial \mathbf{Z}} = \begin{pmatrix}-0.64\\-0.29\\0.93\end{pmatrix};\ \frac{\partial L_{CW}}{\partial \mathbf{Z}} = \begin{pmatrix}-1\\1\\0\end{pmatrix};\ \frac{\partial L_{RCE}}{\partial \mathbf{Z}} = \begin{pmatrix}-0.66\\0.33\\0.33\end{pmatrix}$$

With the gradient derivation, we find that CW and CE shift the sample towards class B while the CE(LL) shifts the samples to class C. A detailed comparison shows that the CW gradient direction is orthogonal to the decision boundary between A and B in this 3-class setup. Thus intuitively, CW loss prefers to encourage the sample to find the nearest decision boundary to cross. CE also results in a gradient direction that is close to the decision boundary. Instead, our RCE loss does not explicitly encourage the sample to choose any decision boundary. All CW, CE, and CE(LL)s share one common property: the logit update direction is dependent on the current sample position on the decision hyperplane. Depending on the position of the sample on the decision plane, CE and CW tend to move the sample towards a semantically close class, while CE(LL) loss explicitly moves the sample to a semantically far class. In this example, the interest class is A, conceptually, a strong attack should maximize the semantic distance from class A, *i.e*. updating in the opposite of the interest class logit vector. The gradient of our loss adopts this direction regardless of the sample position on the decision hyperplane to move the sample far from class A. Due to ignorance of the current sample position, one drawback of our approach is that it might lead to relatively slower convergence. Empirically, we confirm that this is a concern in the very early iterations, see the supplementary for relevant discussion.

Table 1. Comparison of RCE loss with other losses in the white box scenario. The discrepancy between ICR and OLNR exists because not all samples in the dataset are correctly classified.

| | non-targeted Acc. | ICR | OLNR | NLOR | NRT | CosSim |
|---|---|---|---|---|---|---|
| CE | 100.00 | 752.90 | 712.35 | 159.52 | 279.53 | 0.25 |
| CW | 100.00 | 391.40 | 349.94 | 21.01 | 257.22 | 0.40 |
| LL | 99.20 | 491.02 | 490.46 | 888.96 | 306.12 | 0.08 |
| FDA | 100.00 | 619.90 | 608.84 | 517.28 | 311.49 | 0.06 |
| RCE(Ours) | 100.00 | **1000.00** | **979.63** | 570.94 | **360.23** | **-0.21** |
| RCE(LL) | 100.00 | 687.36 | 688.72 | **996.32** | 354.58 | -0.17 |

**Strong top-$k$ white-box attack.** Here, we compare our loss with CE, CW, CE(LL), and FDA [9]. The results are shown in Table 1. Additionally to our proposed ICR metric for evaluating top-$k$ adversarial strength, we also report other metrics as in [9] including OLNR, NLOR, cosine similarity (CosSim), normalized rank transformation (NRT), and ASR, for completeness. The $\alpha$ and $T$ are set to $4/255$ and 20 (same for other experiments, unless specified). The results show that our loss achieves the strongest attack among all losses for all metrics except for NLOR with CE(LL). Note that CE(LL) loss explicitly targets the LL class, thus it is expected NLOR would be higher. We further conduct an experiment with RCE(LL) which achieves 996.32 for NLOR, significantly outperforming CE(LL).

**Stronger top-$k$ attack under image transformations.** Following [21], we apply image transformations to the generated adversarial examples to test whether our loss still achieves stronger attack under image transformation (see Table 2). Note that such a setup constitutes testing the robustness of adversarial examples. Please refer to the supplementary for a detailed experimental setup.

Table 2. ICR under image transformations for different loss functions.

| | No transform | Brightness | Contrast | Gaussian Noise |
|---|---|---|---|---|
| CW | 390.00 | 216.27 | 185.01 | 33.18 |
| CE | 752.90 | 488.92 | 460.19 | 71.28 |
| RCE (Ours) | 1000.00 | 897.85 | 876.94 | 201.25 |

**Loss comparison through the lens of temperature.** From Figure 4, we observe that CE gradient direction lies between that of CW and our loss. Table 1 also shows that the performance of CE also lies in between. Here, we show that CW and our loss can be seen as a special case of CE through changing the temperature $T_e$ [13]. $T_e$ is a non-trivial hyperparameter temperature, *i.e*. pre-processing to $\mathbf{Z} = \mathbf{Z}/T_e$ as the softmax input, resulting in $\mathbf{P}_e$. This temperature scaling method has been widely used for knowledge distillation [5, 13] as well as a defense method [31]. With the temperature taken into account, the derivative of the CE is derived as follows:

$$\frac{\partial L}{\partial \mathbf{Z}} = \frac{1}{T_e}(\mathbf{P}_e - \mathbf{Y}_{gt}), \tag{6}$$

Typically, the temperature $T_e$ is set to 1. From our geometric perspective, the $T_e$ balances the preference of the loss to encourage the sample to cross the decision boundary of the semantically closer class, *i.e*. those classes with relatively high logits. A higher $T_e$ indicates decrease of such preference. With the temperature $T_e$ as a control variable, we reveal that the CW loss can be interpreted as a special case of the CE loss by setting $T_e$ to a small value. Our proposed $RCE$ loss can also be seen as a special case of the CE loss by setting $T_e$ to a relatively large value. The proof is given in the supplementary. Empirically, we demonstrate the influence of temperature on the attack strength in Table 3. The results validate that increasing/decreasing the $T_e$ shifts the performance close to RCE/CW loss.

## 6. Top-$k$ transferable attack

Inspired by the finding that top-$k$ adversarial strength is transferable, we believe that our loss might also lead to a stronger top-$k$ transferable attack since it achieves the strongest white-box attack. Unless specified, we always adopt $\alpha = 4/255$. We set $T$ to 20 and 200 for the non-targeted setting and targeted setting, respectively.

**Non-targeted setting.** The results are shown in Table 4, where we compare our loss with CE and CW in two different baselines: vanilla I-FGSM and MI-DI-TI-FGSM. For

Table 3. Influence of different temperature values in the CE loss. Metric adopted is ICR.

| CW | $T_e = 1/100$ | $T_e = 1/8$ | $T_e = 1/4$ | $T_e = 1/2$ | $T_e = 1$ | $T_e = 2$ | $T_e = 4$ | $T_e = 8$ | $RCE$ |
|---|---|---|---|---|---|---|---|---|---|
| 55.53 | 76.89 | 346.74 | 393.98 | 491.71 | 752.90 | 947.48 | 987.93 | 999.60 | 1000.0 |

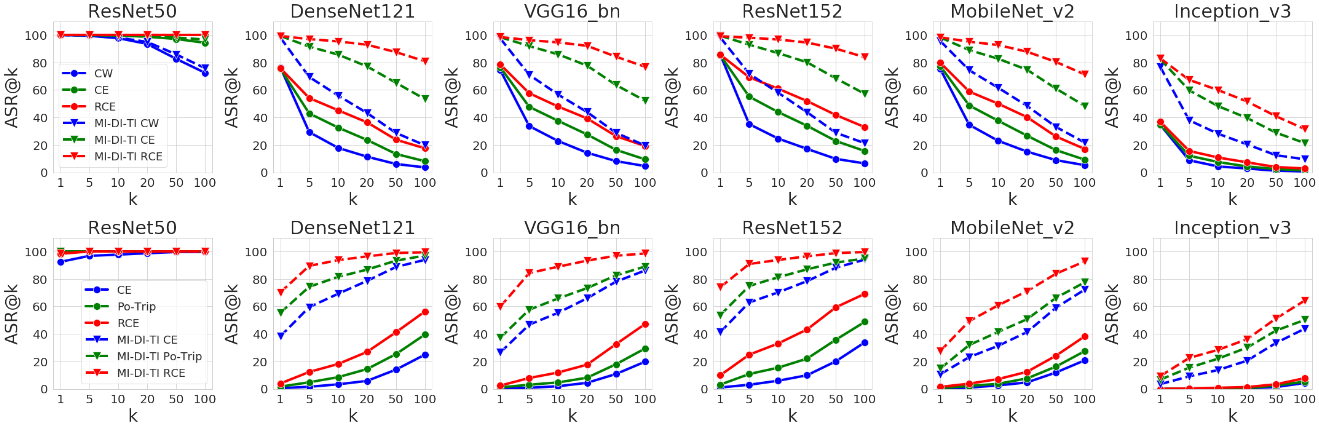

Figure 5. ASR@$k$ transferability with ResNet50 as the surrogate model in untargeted setting (top row) and targeted setting (bottom row)

Table 4. Non-targeted transferability of I-FGSM (top), and MI-DI-TI-FGSM (bottom) attacks for the source network ResNet50. Each entry represents the ICR/non-targeted ASR@1 (%).

| | RN50 | DN121 | VGG16bn | RN152 | MNv2 | IncV3 |
|---|---|---|---|---|---|---|
| CW | 390.00/100.00 | 14.80/76.50 | 18.59/74.30 | 24.15/85.60 | 22.68/75.20 | 5.49/34.60 |
| CE | 752.90/100.00 | 34.16/75.40 | 40.87/76.40 | 61.20/85.20 | 39.21/77.30 | 7.50/34.80 |
| RCE (Ours) | 1000.00/100.00 | 72.11/75.80 | 80.86/78.50 | 144.81/85.60 | 70.39/79.80 | 13.35/36.80 |
| CW | 427.49/100.00 | 77.82/98.10 | 77.13/97.40 | 81.67/98.20 | 84.88/95.60 | 39.03/76.80 |
| CE | 806.85/100.00 | 220.87/99.30 | 213.77/98.40 | 249.02/99.40 | 193.96/98.20 | 89.93/82.40 |
| RCE (Ours) | 999.94/100.00 | 482.58/99.20 | 430.97/98.50 | 517.85/99.00 | 366.30/98.30 | 141.90/83.00 |

both baselines, our RCE loss outperforms CE loss by a large margin. The top row in Figure 5 shows that our loss also achieves higher (untargeted) ASR@$k$, especially when $k$ is set to large for making the task more challenging.

Table 5. Targeted transferability of I-FGSM (Top), and MI-DI-TI-FGSM (bottom) attacks for the source network ResNet50. Each entry represents the ICR/targeted ASR@1 (%).

| | RN50 | DN121 | VGG16bn | RN152 | MNv2 | IncV3 |
|---|---|---|---|---|---|---|
| CE | 2.52/92.40 | 320.73/0.50 | 355.33/0.30 | 264.20/1.00 | 345.40/0.00 | 607.46/0.00 |
| Po-Trip | 1.00/100.00 | 236.37/1.60 | 299.51/1.10 | 192.63/3.10 | 309.81/0.50 | 582.28/0.00 |
| RCE (Ours) | 1.02/98.30 | 161.13/3.90 | 208.61/2.40 | 108.22/9.90 | 244.40/1.40 | 559.95/0.00 |
| CE | 1.00/100.00 | 22.19/38.20 | 45.64/26.50 | 23.61/41.30 | 92.72/10.60 | 245.79/3.40 |
| Po-Trip | 1.00/100.00 | 13.84/55.30 | 40.33/37.20 | 18.46/53.70 | 76.37/14.80 | 215.26/6.70 |
| RCE (Ours) | 1.01/98.90 | 4.51/70.20 | 7.76/59.80 | 3.67/74.00 | 30.90/27.50 | 157.35/9.30 |

Table 6. Targeted transferability of I-FGSM (Top), and MI-DI-TI-FGSM (bottom) attacks for an ensemble of source networks ResNet50 and DenseNet121. Each entry represents the ICR/targeted ASR@1 (%).

| | RN50 | DN121 | VGG16bn | RN152 | MNv2 | IncV3 |
|---|---|---|---|---|---|---|
| CE | 2.07/92.00 | 1.60/96.00 | 242.62/2.20 | 175.88/4.20 | 258.10/1.60 | 521.37/0.00 |
| Po-Trip | 1.00/99.90 | 1.00/100.00 | 203.21/5.30 | 130.12/11.00 | 230.62/2.10 | 492.40/0.40 |
| RCE (Ours) | 1.02/98.30 | 1.02/98.50 | 78.95/16.20 | 44.90/29.00 | 135.20/5.80 | 419.23/0.50 |
| CE | 1.00/100.00 | 1.00/100.00 | 15.74/54.90 | 8.22/66.60 | 43.16/23.50 | 119.14/14.90 |
| Po-Trip | 1.00/100.00 | 1.00/100.00 | 27.86/48.70 | 11.08/65.50 | 53.26/24.20 | 136.38/14.90 |
| RCE (Ours) | 1.01/98.80 | 1.01/98.80 | 2.48/81.70 | 1.86/86.80 | 10.21/50.10 | 59.86/30.70 |

**Targeted setting.** Table 5 shows the results in the targeted setting, where a smaller targeted ICR indicates superior performance. Our approach achieves significantly better performance than CE and Po-Trip loss [23] which constitutes the SOTA approach that generates perturbation in the output space. The bottom row in Figure 5 shows that our loss also results in a higher targeted ASR@$k$, especially when $k$ is set to 1. For example, from ResNet to VGG16, our loss improves the performance of Po-Trip loss from 37.20% to 59.80%. It also outperforms another approach that generates perturbation in the feature space [19] (59.80% vs. 43.5%). Moreover, our loss also achieves comparable targeted transferability as a recent work [47] that adopts logit loss [43] for only maximizing the logit for targeted UAP (UAP is perturbation that fools the model for most images [2, 28, 44, 45]). We further conduct experiments in the ensemble setting by generating adversarial examples on source networks ResNet50 and DenseNet121(see results in Table 6). Following [1], we also report the performance on ViTs [8] and MLP-Mixer [34] (See results in the supplementary.) We observe that our RCE loss consistently outperforms the existing losses by a significant margin.

**CIFAR results.** We further conduct experiment on CIFAR10 (see Table 7) and CIFAR100 (see Table 8). The trend mirrors that on the ImageNet in both non-targeted and targeted settings.

**Images with various types of content.** The core of an

Table 7. Non-targeted (top) and targeted (bottom) ICR/ASR@1 of the MI-FGSM attack for source network ResNet50 trained on CIFAR-10.

| | RN20 | RN56 | VGG19 | DN |
|---|---|---|---|---|
| CW | 6.03/99.70 | 6.07/99.70 | 5.04/98.40 | 6.82/99.60 |
| CE | 6.23/99.50 | 6.28/99.40 | 5.24/98.40 | 6.83/99.20 |
| RCE (Ours) | 8.23/99.10 | 8.00/99.10 | 6.80/96.10 | 8.50/98.70 |
| CE | 1.11/93.40 | 1.12/93.20 | 1.24/87.80 | 1.05/95.50 |
| Po-Trip | 1.64/71.10 | 1.62/68.70 | 1.77/69.00 | 1.43/79.80 |
| RCE (Ours) | 1.08/94.60 | 1.08/94.30 | 1.18/89.80 | 1.03/98.00 |

Table 8. Non-targeted (top) and targeted (bottom) ICR/ASR@1 of the MI-FGSM attack for source network ResNet50 trained on CIFAR-100.

| | RN20 | RN56 | VGG19 | DN |
|---|---|---|---|---|
| CW | 21.57/94.00 | 22.23/95.80 | 20.09/91.60 | 21.62/93.60 |
| CE | 24.31/95.10 | 25.24/96.30 | 24.58/93.70 | 24.44/96.00 |
| RCE (Ours) | 48.32/97.70 | 52.35/97.00 | 44.05/96.40 | 45.82/96.30 |
| CE | 20.74/11.40 | 18.46/16.20 | 31.43/13.60 | 14.52/15.30 |
| Po-Trip | 23.75/10.40 | 21.59/13.60 | 33.24/12.00 | 17.89/14.40 |
| RCE (Ours) | 11.46/22.20 | 10.08/27.50 | 23.08/17.70 | 9.54/20.70 |

adversarial attack against a deep classifier is to add a small perturbation for changing the model output. Here, we perceive this output change as shifting the sample *far from* or *close to* a certain interest class regardless of the original image content. Our ICR can be used to indicate the attack strength in this general setting. Since our loss is sample position-agnostic, the image content itself is irrelevant. Our RCE loss outperforms CE loss in all setups (See results in the supplementary).

## 7. Discussion

**I-FGSM *vs.* PGD.** I-FGSM [22] and PGD [27] are in essence the same except with a technical difference. Specifically, I-FGSM initializes the initial perturbation with zero values while PGD initializes it with random values. Random initialization of PGD allows multiple restarts if the attack fails. However, in the black-box setting, only a single attempt is allowed for the evaluation purpose, thus the community sticks to use I-FGSM based attack for transferable attack. That is why our experiments are also based on initialization-free I-FGSM. In the white-box setting, multiple starts are allowed, however, with a single run (no multiple starts), our loss already achieves 100% ASR@$k$ even when $k$ is set to the maximum $K$.

**Top-$k$ optimization vs. top-$k$ evaluation.** [46] performs the ordered top-k targeted attack. With their definition, their attack considers attacking multiple classes at the same time. With this said, we emphasize that our work does *not* preform a top-k otimization because our goal is to manipulate the rank of a *single class*. Instead, our work only adopts the top-k *metric* as the evaluation. In other words, the $k$ of ASR@$k$ is not utilized in the training stage; in evaluation, it is also recommended to report ASR@$k$ for a wide range of $k$ values. Moreover, our loss is not designed for directly maximizing the ICR (in the untargeted setting) which is a discrete thus non-differential optimization goal. With this said, our loss does not overfit to the ICR metric and improves the performance for all other existing metrics.

**Targeted vs. non-targeted top-$k$ attack.** Top-$k$ targeted attack is *not* very meaningful if we only care about whether the prediction label after the attack is the target class. When the targeted attack goal fails, however, our ICR still conveys non-trivial information, *i.e.* the rank of the target class. It is worth mentioning that non-targeted top-$k$ attack also has an *implicit target* direction (far from the ground-truth class). With this said, we highlight that our ICR provides a unified perspective on targeted and non-targeted attack. For black-box transferability, we highlight that increasing ICR in the non-targeted setting is significantly more challenging than decreasing ICR in the targeted setting. In the scale of 1 to 1000 for ICR on ImageNet, the optimal performance is 1 in targeted setting and 1000 in non-targeted setting. Taking the trasnfer from ResNet50 to DesnseNet121 as example, our results in Table 5 show that the best achieved targeted ICR is 4.51 which is very close to the optimal value of 1. On the other hand, the best achieved non-targeted ICR is 482.58 (see Table 4) which is very far from the optimal value of 1000. Considering the saturated top-1 transferability, we advocate future works to evaluate their attack methods with our ICR, especially for the non-targeted setting. For evaluating the strength of transferable attacks, our ICR can be a more reliable and informative metric than top-1 ASR.

## 8. Conclusion

Our work is the first to investigate the top-$k$ transferable attack. Motivated from the finding that top-$k$ attack strength is transferable, we explore how to achieve strong top-$k$ white-box attack. With the limitation of existing losses identified as mainly prioritizing the speed to fool the network based on an intuitive geometric perspective, we propose a new RCE loss that conceptually maximizes its semantic distance from the ground-truth class. The proposed RCE loss achieves a significantly stronger top-$k$ white-box attack for a wide range of metrics, including our proposed ICR. Due to the transferable property of top-$k$ attack strength, our loss also achieves a top-$k$ transferable attack that is significantly stronger than the SOTA approaches. We further extend our loss to the targeted setting, where a significant performance boost is also observed.

## References

[1] Philipp Benz, Chaoning Zhang, Soomin Ham, Adil Karjauv, and In So Kweon. Robustness comparison of vision transformer and mlp-mixer to cnns. In *CVPR 2021 Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges (AML-CV)*, 2021. 7

[2] Philipp Benz, Chaoning Zhang, Tooba Imtiaz, and In So Kweon. Double targeted universal adversarial perturbations. In *ACCV*, 2020. 7

[3] Philipp Benz, Chaoning Zhang, and In So Kweon. Batch normalization increases adversarial vulnerability and decreases adversarial transferability: A non-robust feature perspective. In *ICCV*, 2021. 2

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *SP*, 2017. 2

[5] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *ICCV*, 2019. 6

[6] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018. 1, 2, 3

[7] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*, 2019. 2, 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 7

[9] Aditya Ganeshan and R Venkatesh Babu. Fda: Feature disruptive attack. In *ICCV*, 2019. 2, 6

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 2

[11] Yiwen Guo, Qizhang Li, and Hao Chen. Backpropagating linearly improves transferability of adversarial examples. *arXiv preprint arXiv:2012.03528*, 2020. 2

[12] Atiye Sadat Hashemi, Andreas Bär, Saeed Mozaffari, and Tim Fingscheidt. Transferable universal adversarial perturbations using generative models. *arXiv preprint arXiv:2010.14919*, 2020. 1

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 6

[14] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *ICCV*, 2019. 2

[15] Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. *ICLR*, 2020. 1

[16] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, 2018. 1

[17] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, 2019. 2

[18] Nathan Inkawhich, Kevin J Liang, Lawrence Carin, and Yiran Chen. Transferable perturbations of deep feature distributions. *ICLR*, 2020. 2

[19] Nathan Inkawhich, Kevin J Liang, Binghui Wang, Matthew Inkawhich, Lawrence Carin, and Yiran Chen. Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability. *NeurIPS*, 2020. 1, 2, 7

[20] Nathan Inkawhich, Wei Wen, Hai Helen Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *CVPR*, 2019. 2

[21] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR2017 workshop*, 2016. 2, 3, 6

[22] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *ICLR*, 2017. 1, 3, 8

[23] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. In *CVPR*, 2020. 2, 3, 7

[24] Qizhang Li, Yiwen Guo, and Hao Chen. Yet another intermediate-level attack. In *ECCV*, 2020. 2

[25] Yingwei Li, Song Bai, Cihang Xie, Zhenyu Liao, Xiaohui Shen, and Alan L Yuille. Regional homogeneity: Towards learning transferable universal adversarial perturbations against defenses. *arXiv preprint arXiv:1904.00979*, 2019. 1

[26] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and blackbox attacks. *ICLR*, 2017. 2

[27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 2, 8

[28] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, 2017. 7

[29] Konda Reddy Mopuri, Vaisakh Shaj, and R Venkatesh Babu. Adversarial fooling beyond" flipping the label". In *CVPRW*, 2020. 2

[30] Chaithanya Kumar Mummadi, Thomas Brox, and Jan Hendrik Metzen. Defending against universal perturbations with shared adversarial training. In *ICCV*, 2019. 1

[31] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *SP*, 2016. 6

[32] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *NeurIPS*, 2019. 1

[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[34] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey

Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021. 7

[35] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *ICLR*, 2018. 2

[36] Jianyu Wang and Haichao Zhang. Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks. In *ICCV*, 2019. 1

[37] Xin Wang, Jie Ren, Shuyun Lin, Xiangming Zhu, Yisen Wang, and Quanshi Zhang. A unified approach to interpreting and boosting adversarial transferability. *ICLR*, 2021. 2

[38] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. *arXiv preprint arXiv:2002.05990*, 2020. 2

[39] Jing Wu, Mingyi Zhou, Shuaicheng Liu, Yipeng Liu, and Ce Zhu. Decision-based universal adversarial attack. *arXiv preprint arXiv:2009.07024*, 2020. 1

[40] Cihang Xie and Alan Yuille. Intriguing properties of adversarial training at scale. *ICLR*, 2020. 1

[41] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, 2019. 2, 3

[42] Chaoning Zhang, Philipp Benz, Gyusang Cho, Adil Karjauv, Soomin Ham, Chan-Hyun Youn, and In So Kweon. Backpropagating smoothly improves transferability of adversarial examples. In *CVPR 2021 Workshop Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges (AML-CV)*, 2021. 2

[43] Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In-So Kweon. Understanding adversarial examples from the mutual influence of images and perturbations. In *CVPR*, 2020. 1, 7

[44] Chaoning Zhang, Philipp Benz, Adil Karjauv, and In So Kweon. Universal adversarial perturbations through the lens of deep steganography: Towards a fourier perspective. *AAAI*, 2021. 7

[45] Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon. A survey on universal adversarial attack. *IJCAI*, 2021. 7

[46] Zekun Zhang and Tianfu Wu. Learning ordered top-k adversarial attacks via adversarial distillation. In *CVPR Workshops*, 2020. 8

[47] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. On success and simplicity: A second look at transferable targeted attacks. *NeurIPS*, 2021. 7

[48] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *ECCV*, 2018. 3