ON STRENGTH AND TRANSFERABILITY OF ADVERSARIAL EXAMPLES: STRONGER ATTACK TRANSFERS BETTER

Chaoning Zhang*, **Philipp Benz***, **Adil Karjauv*** **& In So Kweon**
Korea Advanced Institute of Science and Technology (KAIST)
{chaoningzhang1990, mikolez}@gmail.com, {pbenz, iskweon77}@kaist.ac.kr

ABSTRACT

Our work revisits adversarial attack by perceiving it as shifting the sample semantically close to or far from a certain class, *i.e*. interest class. With this perspective, we introduce a new metric called interest class rank (ICR), *i.e*. the rank of interest class in the adversarial example, to evaluate adversarial strength. The widely used attack success rate (ASR) only taking the top-1 prediction into account can be seen as a special case of ICR. Considering top-k prediction, our ICR constitutes a fine-grained evaluation metric and it can also be readily extended to transfer-based black-box attack. With the widely observed phenomenon that I-FGSM transfers worse than FGSM, adversarial transferability, *i.e*. attack strength on the black-box target model, is widely reported to be at odds with white-box attack strength. Our work challenges this widely held belief with the finding that increasing the number of iterations boosts both white-box strength and black-box transferability. This finding provides a non-trivial insight that adversarial transferability can be enhanced through improving the white-box adversarial strength. To this end, we provide a geometric perspective on the logit gradient and propose a new loss that achieves SOTA white-box attack strength, consequently, also leading to SOTA attack strength in the black-box setting.

1 INTRODUCTION

Deep neural networks (DNNs) (He et al., 2016; Huang et al., 2017) are widely known to be vulnerable to adversarial examples, which are crafted by adding imperceptible image-specific perturbations (Szegedy et al., 2013; Goodfellow et al., 2015; Xie & Yuille, 2020; Zhang et al., 2020b) or universal ones (Moosavi-Dezfooli et al., 2017; Poursaeed et al., 2018; Zhang et al., 2020b;a; Benz et al., 2020; Zhang et al., 2021a; Benz et al., 2021; Zhang et al., 2021b). quasi-imperceptible perturbations to natural images. One intriguing property of adversarial examples is the widely known transferability from one (substitute) model to another (target) model (Kurakin et al., 2016; Dong et al., 2018; Hashemi et al., 2020; Li et al., 2019). This property has been exploited for the transfer-based black-box attack as well as enhancing query-based black-box attack (Inkawhich et al., 2020b). It is widely reported in the literature that I-FGSM increases the attack strength of FGSM, but at the cost of a lower transfer rate. This suggests that the strength of an adversarial attack is at odds with its transferability (Kurakin et al., 2017). Lower transfer rates of I-FGSM are often attributed to the conjecture that iterative attack methods tend to be over-fitting to the substitute model (Kurakin et al., 2017; Dong et al., 2018). We study the influence of step size α and number of iterations T on the transfer rate. Our results demonstrate that a small step size α is a cause of "over-fitting" and that adversarial strength is not (necessarily) at odds with transferability. Given enough iterations, we find that the vanilla I-FGSM can even transfer better than FGSM. Recognizing that many factors might influence transferability, our work is positioned to improve transferability through increasing the adversarial strength, which is orthogonal to most existing techniques.

We introduce a new metric called *interest class rank (ICR)* to, intuitively speaking, indicate the semantic distance of the sample from the interest class, *i.e*. the ground-truth class for a normal

*Equal Contribution

non-targeted attack. With the introduced new metric, we find that semantic adversarial strength and transferability are somewhat positively correlated. We propose a new loss for increasing the semantic adversarial strength. Our new loss is mainly inspired by a geometric illustration of the logit gradient analysis. Intuitively, the proposed Relative Cross-Entropy (RCE) loss maximizes the distance from the interest class, independent of the current sample position on the logit decision hyperplane. The proposed loss achieves the strongest white-box attack, and consequently, also leads to a stronger black-box attack for both non-targeted and targeted scenarios. Even though our focus is mainly to achieve semantically stronger white-box and transfer-based black attack, the results show that our new loss also achieves a higher transfer rate with the attack success rate as the metric, especially for the challenging targeted attack.

2 Background and Motivation

Experimental Setup. Following previous works (Dong et al., 2018; 2019; Li et al., 2020a), we evaluate our proposed techniques on an ImageNet-compatible dataset composed of 1000 images. This dataset was originally introduced in the NeurIPS 2017 adversarial challenge. Consistent with previous methods, we set the maximum perturbation magnitude to $L_\infty = 16$. For a detailed description of the experimental setup please refer to the appendix (A.2).

Influence of α and T on Transfer Rate. The phenomenon that FGSM is more transferable is often attributed to the statement that iterative attack methods tend to over-fit to the substitute model (Kurakin et al., 2017; Dong et al., 2018). However, it remains yet unclear which factor mainly contributes to the over-fitting. Technically, the differences between I-FGSM and FGSM consist of two factors: step size α and number of iteration T. To demystify this, we analyze the influences of α and T on the transfer rate. Please refer to the appendix for the analysis (A.2) and results (Figure 1).

Is I-FGSM always less transferable than FGSM? Interestingly, we find that given sufficiently large iterations, I-FGSM transfers better than FGSM. To our knowledge, we are the first to report this phenomenon. Based on this finding, we get the following non-trivial implication: for understanding the relationship between strength and transferability, adversarial strength is not (necessarily) at odds with transferability; on the contrary, with T as the control variable, increasing strength with larger T is even beneficial for boosting transfer rate. Given similar setups, to improve adversarial transferability, we can simply increase the adversarial strength by increasing the number of iterations T.

3 A Closer Look at Adversarial Strength and Transferability

The attack success rate is the most widely adopted metric for comparing the performance of various attacks. It measures the percentage of misclassified images after the attack, which does not unveil the full picture of the attack. It has been highlighted in (Kurakin et al., 2016) that fooling an image from one type of shred dog to another type of shred dog is less interesting than to an airplane, inspired from which the authors (Kurakin et al., 2016) proposed the iterative least-likely class attack (iter-LL) with the CE(LL) loss. The iter-LL can indeed flip the label into a semantically far class while the original class retains a (comparatively) higher class probability. These semantic strength concerns are not reflected in the metric of attack success rate. For evaluating the sample-wise adversarial semantic strength, old label new ranking (OLNR) and new label old ranking (NLOR) have been proposed in the feature disruptive attack (FDA) work (Ganeshan & Babu, 2019).

Interest class rank. In this work, we propose a new metric called interest class rank (ICR). For a given sample, the model outputs a vector of logit values, termed Z corresponding to each of the K classes. We can sort the logit values in Z in descending order, *i.e.*the highest logit is ranked 1 and the lowest is ranked K. Without losing generality, we perceive the attack as a semantic manipulation on an image that can have any content. Conceptually, if we aim to manipulate the sample regarding a certain interest class as the benchmark, we can manipulate the sample far from or close to the interest class. The attack strength can thus be measured based on the semantic distance to the interest class. There is no established metric to measure such semantic distance. In this work, we propose a straightforward metric to measure the rank of the interest class. Intuitively, if the interest class is ranked 1 for a given (adversarial) example, it is considered as semantically close to the interest class; if the interest class is ranked K, the adversarial example is perceived to be semantically most far from the interest class. Even though ICR is proposed as the metric for sample-wise adversarial

strength, we report the average ICR on all samples. The relationship with other metrics, such as Cosine Similarity, OLNR, and NLOR are introduced in the appendix (A.2).

Adversarial strength is transferable. Regarding the transferability, prior works with the attack success rate as the metric mainly focused on whether the original predicted label is flipped after the attack without analyzing its new rank. In other words, on the target model, the new rank is irrelevant when it ranges from 2 to K. With the ICR as the metric, we study the new rank between the substitute model and target model, *i.e.* whether the semantic adversarial strength is transferable. Specifically, we show the ICR with different α and T, and the results are shown in Figure 2 in the appendix. As a control study, we also report the same results with the metric of attack success rate. The overall trend of the ICR mirrors that of the attack success. For example, either increasing T or α boosts the semantic adversarial strength on the target model significantly. However, it seems to be much more challenging to get satisfactory performance with the ICR metric. With the α set to 1/255, even after 20 iterations, the black-box average ICR is only around 15 (note that the optimal value is 1000 for ImageNet with 1000 classes). Adopting a higher α makes it converge much faster; however, the final average ICR is still only around 40. Overall, the results support that semantic adversarial strength is transferable. Thus, it is possible to realize a semantically strong black-box attack by increasing the semantic adversarial strength on the source model.

4 New Loss for Boosting Semantic Adversarial Strength

The Relative Cross-Entropy Loss. Next, we present our new loss formulation for boosting the semantic adversarial strength. The loss function, which we term Relative Cross-Entropy loss or RCE in short is formulated as follows:

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

The loss function consists of two parts, the commonly used cross-entropy (CE), and a normalization part, averaging the cross-entropy calculated for each class. Our new loss is inspired by the geometric illustration of the logit vector gradient. Refer to the discussion in A.6 and Figure 4 in the appendix for more details. Here, we summarize the key findings that Carlini & Wagner (2017) (CW) loss and CE loss often take a relatively short path to cross the decision boundary, and their gradient directions are dependent on the sample position on the decision boundary. Our loss leads to a gradient direction that maximizes the distance from the GT class regardless of the sample position on the decision boundary.

Semantically Strong White-box Attack. Here, we compare our loss with CE, CW, CE(LL), and FDA (Ganeshan & Babu, 2019). The results are shown in Table 1. Additionally to our proposed ICR metric for evaluating semantic adversarial strength, we also report other metrics as in (Ganeshan & Babu, 2019) including OLNR, NLOR, cosine similarity (CosSim), normalized rank transformation (NRT), and attack success rate, for completeness. The α 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). We also find that our loss leads to superior performance under image transformations. For more details, please refer to A.9 and Table 14 in the appendix.

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

5 Transfer-based Semantically Stronger Attack

In Sec. 3, we show that semantic adversarial strength is transferable, inspired by which we believe that our loss might also lead to transfer-based semantically stronger adversarial attack since it achieves

Table 2: Non-targeted transferability of I-FGSM (top), and MI-DI-TI-FGSM (bottom) attacks for source network ResNet50. Each entry represents the ICR/non-targeted success rate (%).

	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

the strongest white-box attack. In this work, unless specified, we always adopt the $\alpha = 4/255$. We set T to 20 and 200 for the non-targeted attack and targeted attack, respectively. For the targeted attack, we find a small T, such as 20, is not sufficient for achieving satisfactory performance. We first perform a non-targeted attack and the results are shown in Table 2, where we compare our loss with CE and CW in two different baselines: vanilla I-FGSM and MI-DI-TI-FGSM, which is the combination of MI, DI, and TI-FGSM. For both baselines, our RCE loss outperforms CE loss by a large margin in terms of semantic adversarial strength, which is somewhat expected. We also report the non-targeted attack success rate, for which our loss also achieves visibly better performance with the vanilla I-FGSM baseline and equivalent performance with the strong baseline MI-DI-TI FGSM. The transfer-based targeted attack results are shown in Table 3. For the semantic adversarial strength, our approach achieves significantly better performance than CE and Po-Trip loss (Li et al., 2020a) with both baselines. Surprisingly, our loss also achieves significantly better performance in terms of targeted attack success rate. Po-Trip constitutes the SOTA approach that generates perturbation in the output space, somewhat surprisingly our approach also outperforms it by a large margin. For example, from ResNet to VGG16, our loss improves the performance from 37.20% to 59.80%. Moreover, it also outperforms the SOTA approach that generates perturbation in the feature space. For example, after a greedy search with multiple layers, their reported best performance from ResNet50 to VGG16 is 43.5% Inkawhich et al. (2020b). Additional transfer-based black-box results with more source models as well as an ensemble of source models are reported in the appendix (refer to section A.7). Our RCE loss consistently outperforms the existing losses by a large margin.

Table 3: Targeted transferability of I-FGSM (Top), and MI-DI-TI-FGSM (bottom) attacks for source network ResNet50. Each entry represents the ICR/targeted success rate (%).

	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

6 CONCLUSION

Generalizing beyond the top-1 to top-k prediction, we propose a unified metric ICR for evaluating the sample-wise adversarial strength, based on which we revisit adversarial attack for both white-box and transfer-based black-box scenarios. We are the first to report that vanilla I-FGSM can transfer better than FGSM and show that adversarial strength is not necessarily at odds with transferability. On the contrary, we reveal that stronger attack often transfers better. Moreover, our work introduces a geometric perspective on the gradient of the network output logits, motivated from which we propose a new loss that achieves a semantically stronger white-box attack, consequently also resulting in a stronger transfer-based attack. Even though our loss is mainly motivated to improve the semantic adversarial strength indicated by ICR, as a by-product, we also achieve a competitive targeted attack success rate in transfer-based attack, outperforming existing approaches by a large margin. Overall, our work opens a new avenue for benchmarking adversarial strength in transfer-based attack based on our ICR metric and our proposed new loss constitutes a strong baseline in this avenue.

REFERENCES

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

Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Universal adversarial training with class-wise perturbations. *ICME*, 2021.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*, 2017.

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

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

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

Aditya Ganeshan and R Venkatesh Babu. Fda: Feature disruptive attack. In *ICCV*, 2019.

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

Sven Gowal, Jonathan Uesato, Chongli Qin, Po-Sen Huang, Timothy Mann, and Pushmeet Kohli. An alternative surrogate loss for pgd-based adversarial testing. *arXiv preprint arXiv:1910.09338*, 2019.

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

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

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

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

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.

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

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

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*, 2020b.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR2017 workshop*, 2016.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *ICLR*, 2017.

Saehyung Lee, Hyungyu Lee, and Sungroh Yoon. Adversarial vertex mixup: Toward better adversarially robust generalization. In *CVPR*, 2020.

Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. In *CVPR*, 2020a.

Qizhang Li, Yiwen Guo, and Hao Chen. Yet another intermediate-level attack. In *ECCV*, 2020b.

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.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *ICLR*, 2017.

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

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.

Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *CVPR*, 2018.

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.

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

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.

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

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.

Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In-So Kweon. Cd-uap: Class discriminative universal adversarial perturbation. In *AAAI*, 2020a.

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

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

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

A Appendix

A.1 Related Work

Our work aims to achieve a strong white-box attack, consequently boosting the transferability for a stronger black-box attack. For the white-box scenario, the feature disruptive attack (FDA) (Ganeshan & Babu, 2019) is most similar to ours for proposing several metrics, such as OLNR and NLOR, for evaluating the semantic adversarial strength but only for the non-targeted attack. Our work also proposes a unified metric termed ICR that can be used for both non-targeted attacks and targeted attacks. The FDA attack is the SOTA white-box semantically strong attack, and our proposed loss outperforms it by a significant margin for all metrics. For the transfer-based attacks, early works have shown that adversarial examples naively generated on the substitute model in the direct white-box manner, such as vanilla I-FGSM, have low transferability. An ensemble of multiple substitute models is found to improve the transferability (Liu et al., 2017; Tramèr et al., 2018) but at the cost of more

computation resources. Some free techniques have been proposed, MI-FGSM uses a momentum term (Dong et al., 2018), DI-FGSM introduces input diversity (Xie et al., 2019), and TI-FGSM leverages a translation-invariant constraint (Dong et al., 2019). Huang et al. and Li et al. have demonstrated that fine-tuning adversarial examples with the intermediate level attack can further boost the transferability. Enhancing the transferability from the linear perspective (Goodfellow et al., 2015) has been investigated in (Wu et al., 2020; Guo et al., 2020). Most of the above methods are developed for a transfer-based non-targeted attack. The investigation for the more challenging targeted attack is still in its infancy. Most of them are generated in feature space (Inkawhich et al., 2019; 2020a;b), which requires training additional class-wise layer-wise auxiliary classifiers. Similar to our work, (Li et al., 2020a) proposes a new loss that results in a higher targeted transfer rate. Our proposed new loss is originally motivated for achieving a semantically stronger attack. As a byproduct, it also results in a higher transfer-based targeted attack rate.

A.2 Experimental Setup

General setup. Following previous works (Dong et al., 2018; 2019; Li et al., 2020a), we evaluate our proposed approach on an ImageNet-compatible dataset composed of 1000 images. This dataset was originally introduced in the NeurIPS 2017 adversarial challenge. Consistent with prior works, we set the maximum perturbation magnitude to $L_\infty = 16/255$ and set the input size to $3 \times 299 \times 299$. The step size α is set to 4/255 and unless specified, we set the number of iterations T to 20 and 200 for the non-targeted and targeted attack, respectively.

Setup Regarding Image Transformations. To test the robustness of the generated adversarial examples to image transformations, we apply brightness, contrast, and Gaussian noise transformations. For the brightness and contrast transformations, we increase the brightness and contrast by a factor of 2. For the Gaussian noise augmentation, we apply Gaussian noise centered around zero mean, with a standard deviation of 0.1.

Influence of α and T on transfer rate. The influence of α and T on the transfer rate is shown in Figure 1. We have two major observations: (a) Given a fixed α, increasing T enhances the transfer rate; (b) Given a fixed T, increasing α significantly boosts the transfer rate, especially when T is not sufficiently large. The results demonstrate that *the factor that contributes to over-fitting of I-FGSM is α rather than T*.

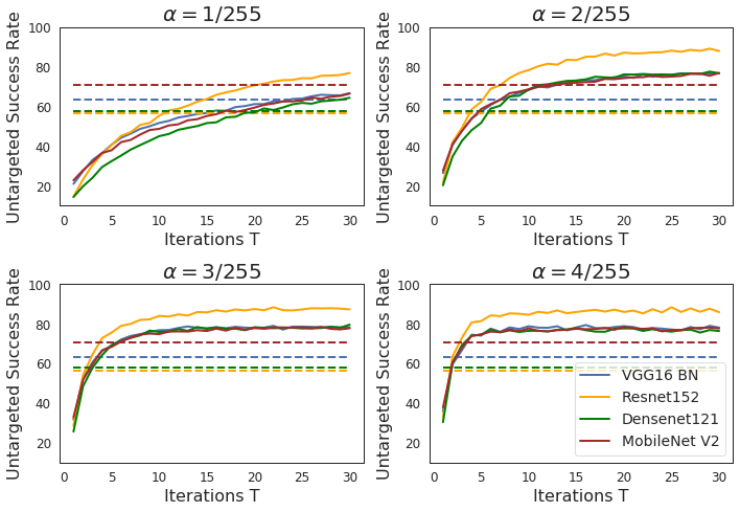

Figure 1: Transferability result for the FGSM (dashed lines) and I-FGSM (solid lines) attack trained on source network ResNet50 (RN50) and transferred to various black-box models. The performance is presented for different step sizes (α) and over the number of iterations. The adopted metric is the non-targeted attack success rate ($\%$).

Relationship with Other Metrics. The attack success rate only measures whether the ground-truth class is still the highest, while our ICR metric captures more fine-grained details of the new rank of

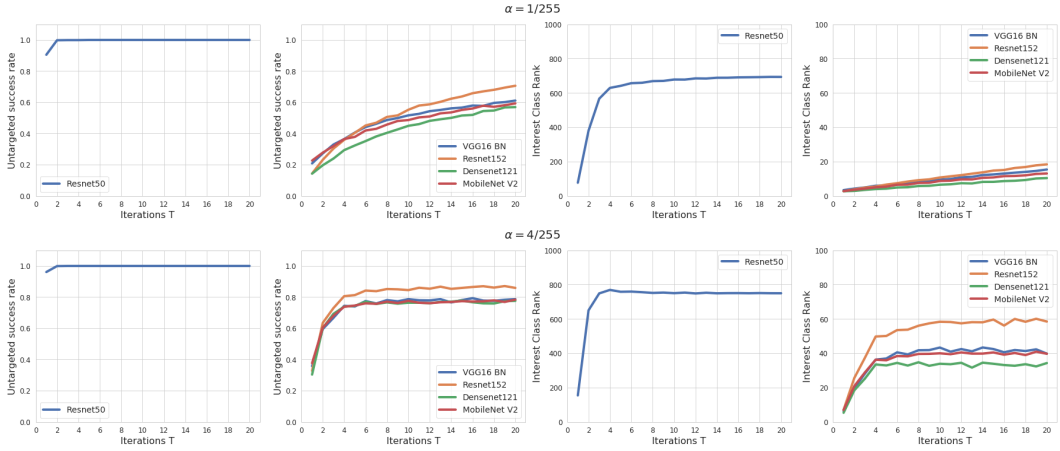

Figure 2: ICR and attack success rate with α set to 1/255 (top) and 4/255 (bottom) for white-box and black-box scenarios.

the ground-truth class. Note that top-k accuracy, such as top-1 or top-5, is often adopted for evaluating the model accuracy. With the ICR results, we can easily calculate the top-k accuracy for ranging k from 1 to k. One alternative metric for reflecting the semantic distance to the interest class can be the predicted class probability, *i.e*. higher probability indicates being semantically closer to the interest class. However, when the image is misclassified after the attack, the probability is very close to zero in most cases, thus it is not a good metric for indicating the attack strength. In a normal non-targeted attack setup, our metric is very similar to the OLNR introduced in (Ganeshan & Babu, 2019) since the so-called old label, *i.e*. the predicted label before the attack, is the same as the ground-truth class, *i.e*. interest class, for most images. Note that if the old label is different from the ground-truth class, it is not practically meaningful to discuss attack since the sample is misclassified even without an attack. OLNR can be seen as a special case of the ICR in the non-targeted attack when the image has the interest class content. Our metric is not limited to this specific scenario. For example, when the image has random content, we can still evaluate the semantic intervention strength by setting the attack goal to shift the sample semantically far from the interest class. Moreover, our ICR can also easily be extended to targeted attack for indicating how close the attack intervention shifts the sample towards the interest class. Note that neither OLNR nor NLOR can be directly adopted in the targeted attack scenario.

A.3 Zero-sum Constraint

Table 4: Zero-sum experiment results on the ImageNet dataset with various network architectures. The metrics reported are the average (with standard deviation) over the validation samples.

	Sum	Abs. Sum	Std	Min	Max
RN50	0.01±0.00	1830.44±284.83	2.45±0.36	-5.80±0.93	16.87±4.70
DN121	0.02±0.00	1876.99±291.23	2.50±0.36	-6.23±1.01	15.96±4.14
VGG16bn	0.01±0.00	2324.19±410.37	3.09±0.56	-6.76±1.26	19.12±6.55
RN152	0.00±0.00	1825.42±302.96	2.45±0.38	-5.84±0.97	17.67±4.58
MNv2	0.08±0.06	2304.12±298.43	3.03±0.40	-8.12±1.28	17.22±4.90

Phenomenon. The *zero-sum* phenomenon of logit vector $\mathbf{Z}$ shows that the sum of the logits mostly results in a value close to zero for both clean and adversarial samples. Here, we empirically demonstrate the phenomenon of the "zero-sum" constraint by evaluating the sum of the logit value $\mathbf{Z}$ on the ImageNet-compatible dataset introduced in the NeurIPS 2017 (See the Experimental general setup) and CIFAR10/CIFAR100 for different network architectures. From the results in Table 4 and Table 5, the first observation is that the sum of all logit values in $\mathbf{Z}$, *i.e*.$\sum_{i=1}^{i=K} z_i$ is indeed close to

Table 5: Zero-sum experiment results on the CIFAR10 (top) and CIFAR100 (bottom) datasets with various network architectures. The metrics reported are the average (with standard deviation) over the validation samples.

	Sum	Abs. Sum	Std	Min	Max
RN20	0.02±0.02	56.21±12.76	7.73±1.90	-8.29±2.01	19.15±6.48
RN32	0.02±0.01	45.75±8.84	6.48±1.27	-6.46±1.57	16.60±4.60
RN44	0.05±0.02	46.41±9.12	6.59±1.39	-6.45±1.44	16.99±4.95
RN56	0.03±0.02	39.99±7.88	5.78±1.10	-5.49±1.43	15.15±3.86
VGG19bn	0.00±0.00	22.55±2.45	3.49±0.38	-3.67±1.00	9.42±1.38
DenseNet-BC-190-40	0.01±0.00	27.82±4.04	4.28±0.49	-3.54±0.87	11.89±1.79
RN20	0.32±0.15	483.70±98.10	6.21±1.29	-13.60±2.97	20.33±6.77
RN32	0.20±0.13	511.02±93.67	6.58±1.24	-14.07±2.87	22.71±7.19
RN44	0.20±0.14	498.21±87.48	6.45±1.17	-13.81±2.78	22.95±7.24
RN56	0.29±0.17	474.85±79.69	6.15±1.08	-13.00±2.60	22.39±6.95
VGG19bn	0.00±0.00	223.78±22.21	2.89±0.27	-4.34±0.51	12.70±2.12
DenseNet-BC-190-40	0.03±0.00	189.25±32.73	2.74±0.49	-4.69±0.96	15.02±5.01

Table 6: Zero-sum experiment results for the adversarial images crafted with different losses: CW (top), CE (middle), RCE (bottom) on ImageNet dataset with ResNet50 (white-box model) and DenseNet121.

	Sum	Abs. Sum	Std	Min	Max
RN50	0.01±0.00	1917.60±232.74	2.49±0.32	-6.30±0.87	14.61±4.55
DN121	0.02±0.00	2224.91±371.48	2.95±0.52	-7.33±1.31	21.04±6.42
RN50	0.01±0.00	2001.94±251.94	2.63±0.36	-6.49±0.93	16.91±5.61
DN121	0.02±0.00	2373.04±419.24	3.19±0.61	-7.70±1.51	24.05±7.69
RN50	0.01±0.00	1720.35±178.80	2.20±0.23	-5.92±0.78	10.29±2.26
DN121	0.02±0.00	1869.99±262.75	2.38±0.35	-7.78±1.97	9.52±2.40

zero with a very small variance among the validation samples, indicating $\sum_{i=1}^{i=K} z_i$ is very close to zero for all validation samples. To further demonstrate that this phenomenon is not just occurring, due to very small values in $\mathbf{Z}$, we further present the absolute sum, *i.e.* $\sum_{i=1}^{i=K} |z_i|$. Additionally, the relatively large values for the standard deviation, the minimum, and maximum value for the Z statistics demonstrate that there exists a balance between the negative and positive logit values, which results in their sum being zero. This phenomenon is also observed for adversarial samples. We report the results for adversarial examples with different losses (CW, CE, RCE) on ImageNet dataset with ResNet50 transferring to DenseNet121 in Table 6. The results show that the zero-sum constraint also holds for adversarial examples.

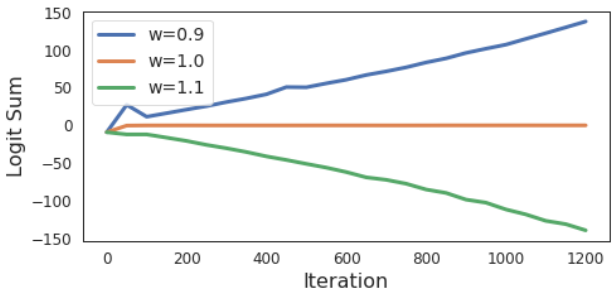

Figure 3: Influence of w on the $\sum_{i=1}^{i=K} z_i$ in the training stage.

Possible Explanation. First of all, we admit that we do not have a clear explanation for this phenomenon. Here, we only attempt to provide a possible explanation. Note that the DNN is often trained with the CE loss. Taking a closer look at the derivation of the CE loss with respect to the logit vector, *i.e.* $\frac{\partial L_{CE}}{\partial \mathbf{Z}} = \mathbf{P} - \mathbf{Y}_{gt}$, it can be observed that sum of all values in both $\mathbf{P}$ and Y_{gt} is 1. We believe that this constitutes a *necessary* condition for making the $\sum_{i=1}^{i=K} z_i$ close to zero. To verify this claim, we experiment with a new loss that results in $\frac{\partial L}{\partial \mathbf{Z}} = w\mathbf{P} - \mathbf{Y}_{gt}$. When w is set to a value larger than 1, such as 1.1, the loss makes the $\sum_{i=1}^{i=K} z_i$ smaller and smaller as the network training goes on (See Figure 3). Similarly, when w is set to a value smaller than 1, such as 0.9, the loss tends to increase the $\sum_{i=1}^{i=K} z_i$. In the above two cases, we observe that $\sum_{i=1}^{i=K} z_i$ eventually becomes infinitely negative/positive given enough iterations and consequently the network training does not converge. When w is set to 1, which is identical to the original CE loss, we observe that $\sum_{i=1}^{i=K} z_i$ converges to zero. Overall, we find that the sum of all values in $\frac{\partial L_{CE}}{\partial \mathbf{Z}}$ is a *necessary*, but probably not *sufficient*, condition for making $\sum_{i=1}^{i=K} z_i$ approach zero. We leave a more elaborate explanation for this phenomenon to future work.

A.4 Logit Vector Gradient Derivations

Here, we provide a detailed derivation of the partial derivative of various loss functions with respect to the logit vector Z.

CE Loss. Before demonstrating the derivative for the CE loss, we will first calculate the derivatives of the softmax output ($\mathbf{P}$) with respect to its input (the logit vector $\mathbf{Z}$). Each entry of the logit vector $\mathbf{Z}$ is indicated with index i, while each entry of the $\mathbf{P}$ is indicated with index j. For simplicity, we divide into two scenarios, $j = i$ and $j \neq i$, and conduct the derivation respectively. First, let's consider *i.e.*$j = i$, and the derivative $\frac{\partial p_j}{\partial z_i}$ can be calculated as follows:

$$\begin{aligned}
\frac{\partial p_i}{\partial z_i} &= \frac{\partial (\frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}})}{\partial z_i} \\
&= \frac{e^{z_i} \sum_{k=1}^{K} e^{z_k} - e^{2z_i}}{(\sum_{k=1}^{K} e^{z_k})^2} \\
&= p_i - p_i^2 \\
&= p_i(1 - p_i),
\end{aligned} \tag{2}$$

with $p_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}$. Eq. 2 enables us to obtain the derivative of the CE loss, *i.e.* $L_{CE} = -\log p_{gt}$, with respect to the softmax input which has the ground-truth index, *i.e.*$i = j = gt$:

$$\begin{aligned}
\frac{\partial L_{CE}}{\partial z_{gt}} &= \frac{\partial (-\log p_{gt})}{\partial z_{gt}} \\
&= -\frac{1}{p_{gt}} \frac{\partial p_{gt}}{\partial z_{gt}} \\
&= -\frac{1}{p_{gt}} (p_{gt}(1 - p_{gt}) \\
&= p_{gt} - 1.
\end{aligned} \tag{3}$$

On the other hand, for the case when $j \neq i$, the derivative $\frac{\partial p_j}{\partial z_i}$ can be calculated as follows:

$$\begin{aligned}
\frac{\partial p_j}{\partial z_i} &= \frac{\partial (\frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}})}{\partial z_i} \\
&= -\frac{e^{z_j} e^{z_i}}{(\sum_{k=1}^{K} e^{z_k})^2} \\
&= -p_j p_i.
\end{aligned} \tag{4}$$

With Eq. (4), we further calculate the derivative of the CE loss with respect to the softmax inputs which are different from the ground-truth index, *i.e.*$i \neq gt$:

$$\begin{aligned}\frac{\partial L_{CE}}{\partial z_i} &= \frac{\partial(-\log p_{gt})}{\partial z_i} \\ &= -\frac{1}{p_{gt}}\frac{\partial p_{gt}}{\partial z_i} \\ &= -\frac{1}{p_{gt}}(-p_{gt}p_i) \\ &= p_i.\end{aligned} \tag{5}$$

From Eq. 3 and Eq. 5, we arrive at the final formulation:

$$\frac{\partial L_{CE}}{\partial \mathbf{Z}} = \mathbf{P} - \mathbf{Y}_{gt}, \tag{6}$$

with $\mathbf{Y}_i$ indicating a one-hot encoded vector with the position at index i being one. Thus, the derivative of the CE(LL) loss, *i.e.* $L_{CE} = -\log P_{LL}$, to the logit vector can be derived similarly with the final formulation shown as follows:

$$\frac{\partial L_{CE(LL)}}{\partial \mathbf{Z}} = \mathbf{P} - \mathbf{Y}_{LL}. \tag{7}$$

CW Loss. CE and CW loss are the two most widely used losses for the white-box attack (Gowal et al., 2019; Lee et al., 2020). In the above, we derive the gradient for CE loss and we further conduct a similar derivation for CW loss which is denoted as $L_{CW} = z_j - z_{gt}$ (Gowal et al., 2019; Lee et al., 2020) with $j = \arg\max_{i \neq gt} z_i$ indicating the highest class except for the gt class. The derivative of the L_{CW} to the Z is denoted as $\frac{\partial L_{CW}}{\partial \mathbf{Z}}$. $\frac{\partial L_{CW}}{\partial z_i} = 0$ when $i \neq j$ and $i \neq gt$. $\frac{\partial L_{CW}}{\partial z_i}$ is 1 and -1 when $i = j$ and $i = gt$, respectively. Therefore, we arrive at:

$$\frac{\partial L_{CW}}{\partial \mathbf{Z}} = \mathbf{Y}_j - \mathbf{Y}_{gt}. \tag{8}$$

Relative Cross-Entropy (RCE) Loss. With Eq. 6, we can calculate the derivative of the proposed RCE loss:

$$\begin{aligned}\frac{\partial L_{RCE}}{\partial \mathbf{Z}} &= \frac{\partial(L_{CE_{gt}} - \frac{1}{K}\sum_{k=1}^{K} L_{CE_k})}{\partial \mathbf{Z}} \\ &= \frac{\partial L_{CE_{gt}}}{\partial \mathbf{Z}} - \frac{1}{K}\sum_{k=1}^{K}\frac{\partial L_{CE_k}}{\partial \mathbf{Z}} \\ &= \mathbf{P} - \mathbf{Y_{gt}} - \frac{1}{K}\sum_{k=1}^{K}(\mathbf{P} - \mathbf{Y_k}) \\ &= \frac{1}{K}\mathbf{1} - \mathbf{Y}_{gt}.\end{aligned} \tag{9}$$

where $\mathbf{1}$ indicates a vector with all values being 1.

A.5 CW AND RCE ARE SPECIAL CASES OF CE

Derivative of the temperature scaled CE-loss. The derivative of the CE-Loss with temperature scaling can be written as:

$$\frac{\partial L_{CE(Temp)}}{\partial \mathbf{Z}} = \frac{1}{T_e}(\mathbf{P}_e - \mathbf{Y}_{gt}), \tag{10}$$

This derivation unfolds similarly to the one previously presented for the CE Loss without temperature. Each entry of the logit vector $\mathbf{Z}$ is indicated with index i, while each entry of $\mathbf{P}$ is indicated with

index j. Again first looking at the softmax output with T_e ($\mathbf{P}_e$) with respect to the logit vector $\mathbf{Z}$ with $i = j$ we arrive at:

$$\begin{aligned}\frac{\partial p_e^i}{\partial z_i} &= \frac{\partial(\frac{e^{z_i/T_e}}{\sum_{k=1}^{K} e^{z_k/T_e}})}{\partial z_i} \\ &= \frac{1}{T_e}(p_e^i(1-p_e^i)).\end{aligned} \tag{11}$$

For the case where $i \neq j$ we arrive at the following derivative:

$$\begin{aligned}\frac{\partial p_e^j}{\partial z_i} &= \frac{\partial(\frac{e^{z_j/T_e}}{\sum_{k=1}^{K} e^{z_k}})}{\partial z_i} \\ &= \frac{1}{T_e}(-p_e^j p_e^i).\end{aligned} \tag{12}$$

Analogous to Eq. (3) and Eq. (5), we can calculate the derivatives for the CE Loss with T_e. For the case $i = gt$ we arrive at:

$$\begin{aligned}\frac{\partial L_{CE(Temp)}}{\partial z_{gt}} &= \frac{\partial(-\log p_e^{gt})}{\partial z_{gt}} \\ &= \frac{1}{T_e}(p_e^{gt} - 1).\end{aligned} \tag{13}$$

For the case $i \neq gt$ we arrive at:

$$\begin{aligned}\frac{\partial L_{CE(Temp)}}{\partial z_i} &= \frac{\partial(-\log p_e^{gt})}{\partial z_i} \\ &= \frac{1}{T_e}p_e^i,\end{aligned} \tag{14}$$

With Eq. (13) and Eq. (14), we finally arrive at Eq. (10).

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 (Hinton et al., 2015). 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 (Hinton et al., 2015; Cho & Hariharan, 2019) as well as a defense method (Papernot et al., 2016). With the temperature taken into account, the derivative of the CE is derived as follows:

$$\frac{\partial L}{\partial \mathbf{Z}} = \frac{1}{T_e}(\mathbf{P}_e - \mathbf{Y}_{gt}), \tag{15}$$

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. Empirically, we demonstrate the influence of temperature on the semantic strength in Table 7. The results validate that increasing/decreasing the T_e shifts the performance close to RCE/CW loss.

Table 7: Influence of different temperature values on 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

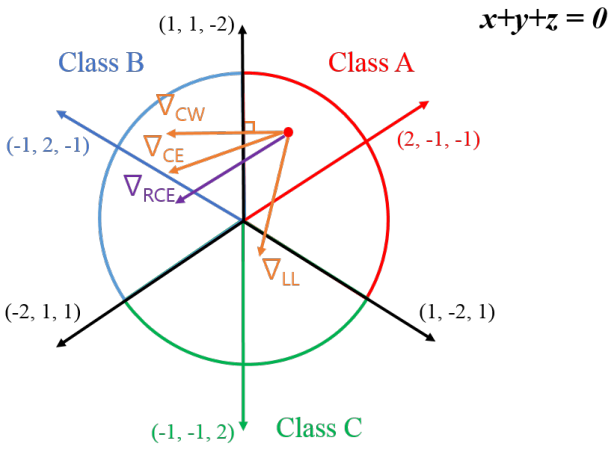

Figure 4: Geometric illustration of the logit gradient for different losses.

A.6 Geometric Illustration of the Logit Gradient.

After establishing the gradient directions of common loss functions and the introduction of our loss function and its corresponding gradient (A.4 and A.5), we now provide a geometric perspective, to illustrate why the proposed loss increases semantic 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.

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 (A.3) 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 derived derivatives for CE, CW and CE(LL), are:

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

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 semantically 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. Refer to A.8 for more details.

Scale-invariant Property of the Gradient Derivative. For FGSM, only the direction of the derivative matters and the scale is irrelevant because the attack is adopted as the basic method for all approaches to get the sign of the derivative. Without losing generality, we compare two losses L_A and L_B by setting $\frac{\partial L_B}{\partial \mathbf{Z}} = s\frac{\partial L_A}{\partial \mathbf{Z}}$ where s is a positive scale factor. We can derive:

$$\begin{aligned} sign(\frac{\partial L_B}{\partial \mathbf{X}}) &= sign(\frac{\partial L_B}{\partial \mathbf{Z}}\frac{\partial \mathbf{Z}}{\partial \mathbf{X}}) \\ &= sign(s\frac{\partial L_A}{\partial \mathbf{Z}}\frac{\partial \mathbf{Z}}{\partial \mathbf{X}}) \\ &= sign(\frac{\partial L_A}{\partial \mathbf{Z}}\frac{\partial \mathbf{Z}}{\partial \mathbf{X}}) \\ &= sign(\frac{\partial L_A}{\partial \mathbf{X}}) \end{aligned} \tag{16}$$

Relationship to other loss functions. The probability of the i-th class in $\mathbf{P}_e$ is shown as:

$$p_e^i = \frac{e^{z_i/T_e}}{\sum_{k=1}^{k=K} e^{z_k/T_e}} \tag{17}$$

Note that T_e ranges from $(0, \infty)$. Without losing generality, by assuming $z_x > z_y$, we can derive:

$$\begin{aligned} \frac{p_e^x}{p_e^y} &= \frac{\frac{e^{z_x/T_e}}{\sum_{k=1}^{k=K} e^{z_k/T_e}}}{\frac{e^{z_y/T_e}}{\sum_{k=1}^{k=K} e^{z_k/T_e}}} \\ &= \frac{e^{z_x/T_e}}{e^{z_y/T_e}} \\ &= e^{(z_x - z_y)/T_e} \\ &> 1 \end{aligned} \tag{18}$$

RCE loss can be seen as a special case of CE loss. For $z_x > z_y$ and $T_e \to \infty$, we can derive:

$$\begin{aligned} \lim_{T_e \to \infty} \frac{p_e^x}{p_e^y} &= \lim_{T_e \to \infty} e^{(z_x - z_y)/T_e} \\ &= 1 \end{aligned} \tag{19}$$

With the above equation and $\sum_{k=1}^{k=K} p_e^i = 1$, it can be concluded that $\mathbf{P}_e = \frac{1}{K}\mathbf{1}$ when $T_e \to \infty$ or when T_e is set to a large value. Thus, in this case, Eq. (10) can be further derived as follows:

$$\begin{aligned} \frac{\partial L_{CE(Temp)}}{\partial \mathbf{Z}} &= \frac{1}{T_e}(\mathbf{P}_e - \mathbf{Y}_{gt}) \\ &= \frac{1}{T_e}(\frac{1}{K}\mathbf{1} - \mathbf{Y}_{gt}) \end{aligned} \tag{20}$$

Given the scale-invariant property indicated by Eq. (16), Eq. (20) is equivalent to the derived gradient in Eq. (9) for the RCE loss. Thus, we conclude that the RCE loss can be seen as a special case of the CE loss by setting T_e to a large value.

CW loss can be seen as a special case of CE loss. We will now show the behavior of $\mathbf{P_e^i}$ when $T_e \to 0$. Given $z_x > z_y$, we can derive:

$$\begin{aligned} \lim_{T_e \to 0} \frac{P_e^x}{P_e^y} &= \lim_{T_e \to 0} e^{(z_x - z_y)/T_e} \\ &= \infty \end{aligned} \tag{21}$$

If i_{max} is the index of the class with the largest logit, $\lim_{T_e \to 0} p_e^{i_{max}} = 1$. Otherwise, $\lim_{T_e \to 0} p_e^{i} = 0$ ($i \neq i_{max}$). Given the definition $j = \arg\max_{i \neq gt} z_i$, we know that the class with the highest logit in $\mathbf{Z}$ is either the j-th class or the gt class. Thus, for small enough T_e ($T_e \to 0$), $p_e^{gt} + p_e^{j} = 1$. Let us denote $p_j = m$ and $p_{gt} = 1 - m$. Then, Eq. 10 can be rewritten as

$$\frac{\partial L_{CE(Temp)}}{\partial \mathbf{Z}} = \frac{m}{T_e}(\mathbf{Y}_j - \mathbf{Y}_{gt}), \tag{22}$$

Given the scale-invariant property indicated by Eq. 16, Eq. 22 is equivalent to the derived gradient in Eq. 8 for the CW loss. Thus, we conclude that the CW loss can be seen as a special case of the CE loss by setting T_e to a very small value.

A.7 Additional Transferability Results.

Table 8: Non-targeted transferability of I-FGSM (top), and MI-DI-TI-FGSM (bottom) attacks for source network DenseNet121. Each entry represents the ICR/non-targeted success rate (%).

	RN50	DN121	VGG16bn	RN152	MNv2	IncV3
CW	27.49/86.10	636.75/100.00	22.98/80.20	16.96/73.90	26.88/76.40	8.36/42.30
CE	68.10/86.70	851.94/100.00	58.14/84.90	39.74/75.30	51.90/79.20	14.22/45.70
RCE (Ours)	128.89/85.30	1000.00/100.00	100.56/83.50	72.46/75.10	85.67/82.90	19.32/44.10
CW	70.29/96.90	632.11/100.00	58.03/96.40	46.42/92.30	80.04/92.90	39.95/76.90
CE	206.55/98.50	883.52/100.00	192.35/97.80	138.87/95.50	168.07/96.70	99.17/84.30
RCE (Ours)	378.31/98.50	1000.00/100.00	337.00/98.10	254.97/95.20	288.82/97.50	144.69/82.80

Table 9: Targeted transferability of I-FGSM (Top), and MI-DI-TI-FGSM (bottom) attacks for source network DenseNet121. Each entry represents the ICR/targeted success rate (%).

	RN50	DN121	VGG16bn	RN152	MNv2	IncV3
CW	195.81/4.60	1.11/99.90	219.33/2.60	265.05/1.20	277.73/0.70	533.24/0.10
CE	295.22/0.90	1.12/97.50	322.68/0.60	341.30/0.60	347.49/0.30	586.48/0.00
RCE (Ours)	154.12/5.20	1.01/98.70	175.77/4.00	226.36/1.70	254.80/0.80	510.23/0.30
Po-Trip	245.60/2.30	1.00/100.00	282.81/1.20	304.62/0.60	319.42/0.50	562.21/0.00
CW	39.09/38.30	1.00/100.00	54.32/27.10	69.03/23.70	103.37/11.50	205.07/6.90
CE	79.21/15.70	1.00/100.00	107.42/10.90	118.32/7.80	163.35/4.50	280.00/2.70
Po-Trip	84.02/17.80	1.00/100.00	128.97/10.30	124.94/8.90	172.83/4.90	291.45/3.30
RCE (Ours)	16.95/45.50	1.01/98.80	22.67/39.80	41.85/29.50	71.47/14.30	165.36/8.90

Table 10: Non-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/non-targeted success rate (%).

	RN50	DN121	VGG16bn	RN152	MNv2	IncV3
CW	316.46/100.00	558.95/100.00	26.06/90.50	30.95/94.30	30.02/86.10	8.87/55.50
CE	705.61/100.00	827.97/100.00	94.82/93.40	111.01/96.30	76.93/90.20	24.24/57.50
RCE (Ours)	1000.00/100.00	1000.00/100.00	194.79/92.50	262.02/93.30	153.23/91.00	39.06/56.40
CW	343.13/100.00	571.41/100.00	87.43/99.80	90.96/99.70	106.95/98.40	66.65/90.00
CE	769.42/100.00	870.79/100.00	346.65/99.90	368.33/99.80	293.47/99.30	216.17/94.00
RCE (Ours)	999.78/100.00	1000.00/100.00	632.32/99.50	681.50/99.80	538.56/99.50	301.55/93.10

Additional Source Model(s). In the main manuscript, we present the targeted transferability results with ResNet50 as the source model. Additionally, we choose DenseNet121 as the source white-box model. The results are shown in Table 8 for non-targeted attack and in Table 9 for targeted attack. Moreover, we also ensemble ResNet50 and DenseNet121 and the results are shown in Table 10 and in Table 11 for the non-targeted and targeted attack, respectively. We find that the trend mirrors that of choosing ResNet50 as the source white-box model. Specifically, our proposed RCE loss outperforms

Table 11: 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 success rate (%).

	RN50	DN121	VGG16bn	RN152	MNv2	IncV3
CW	1.00/100.00	1.00/99.90	121.00/12.60	75.19/22.90	173.41/4.20	449.05/0.50
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
CW	1.00/100.00	1.00/100.00	9.64/74.00	5.60/81.80	26.12/41.80	89.31/23.40
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

the existing losses for achieving a semantically stronger attack. For the more challenging targeted attack, our RCE loss also results in a significantly higher targeted success rate.

CIFAR results. Additionally, we compare the transferability performance for different losses on the

Table 12: Non-targeted (top) and targeted (bottom) transferability of the MI-FGSM attack for source network ResNet50 trained on CIFAR-10. Each entry represents the ICR/non-targeted (top) and targeted (bottom) success rate (%).

	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 13: Non-targeted (top) and targeted (bottom) transferability of the MI-FGSM attack for source network ResNet50 trained on CIFAR-100. Each entry represents the ICR/non-targeted (top) and targeted (bottom) success rate (%).

	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

CIFAR datasets. The non-targeted and targeted results in Table 12 show that our proposed RCE loss also outperforms other loss functions on the CIFAR10 dataset. The higher ICR and lower ICR for the untargeted and targeted attack, respectively, indicate our proposed RCE loss leads to a semantically stronger attack. The results on the CIFAR100 in Table 13 resemble the results on CIFAR10. For the non-targeted attack, our RCE loss leads to comparable (slightly inferior on the CIFAR10 dataset and slightly superior on CIFAR100) performance for the metric of non-targeted attack success rate. For the targeted attack, our RCE loss leads to visibly better performance for the metric of targeted attack success rate.

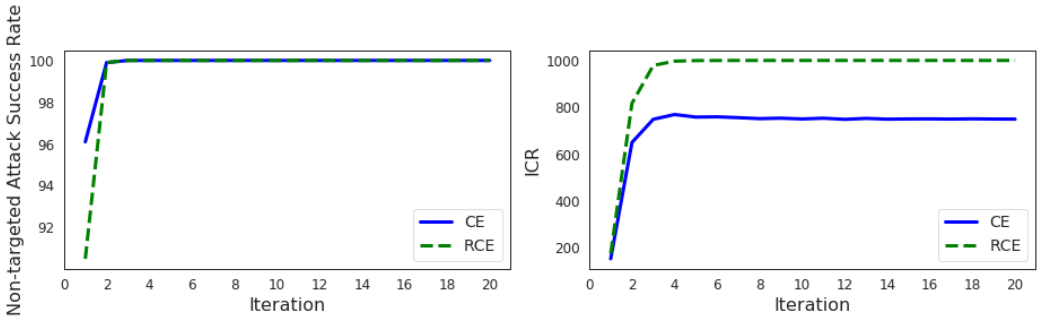

Figure 5: Non-targeted attack success rate and ICR on the white-box ResNet50.

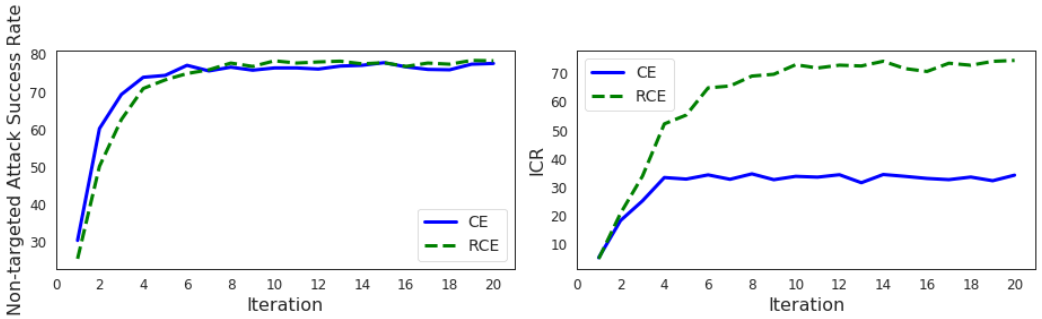

Figure 6: Non-targeted attack success rate and ICR on the black-box DenseNet50.

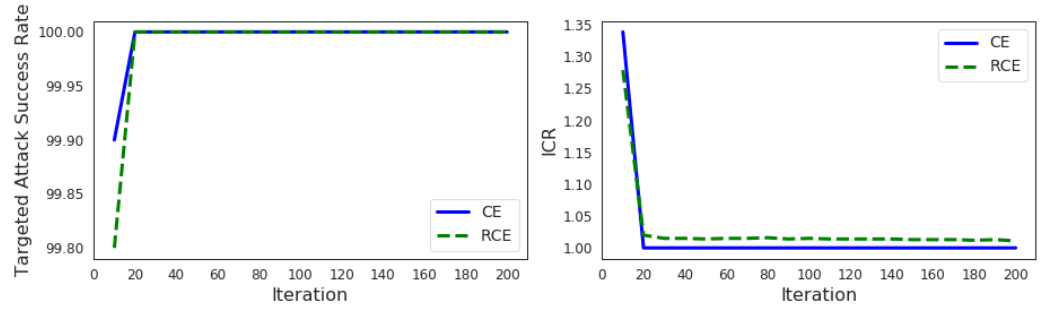

Figure 7: Targeted attack success rate and ICR on the white-box ResNet50.

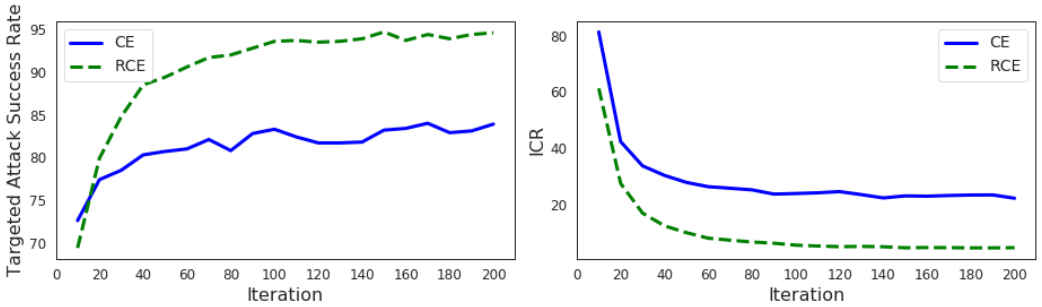

Figure 8: Targeted attack success rate and ICR on the black-box DenseNet121.

A.8 LIMITATION OF RCE LOSS IN THE EARLY ITERATIONS

As indicated in section A.6, the proposed RCE loss might converge slower than the existing CE loss due to its position-agnostic property. Transferring from ResNet50 to DenseNet121 on the ImageNet, we provide the white-box results and black-box results in Figure 5 and Figure 6, respectively. We observe that in the early iterations, CE outperforms our proposed RCE loss, especially for the metric of attack success rate. Similar trend can be observed for targeted attack results (see Figure 7 and Figure 8). Moreover, for the targeted attack, we observe that a small T, such as 20, is not sufficient for achieving satisfactory performance. Setting T to a larger value, such as 200, can boost both strength and transferability.

A.9 SEMANTICALLY STRONGER ATTACK UNDER IMAGE TRANSFORMATIONS.

Following (Kurakin et al., 2016), we apply image transformations to the generated adversarial examples to test whether our loss still achieves semantically stronger attack under image transformation. Note that such a setup constitutes to test the robustness of adversarial examples. The results are shown in Table 14. We observe that our proposed loss also achieves superior performance than the widely used CE loss, which suggests that our loss results in semantically stronger and robust adversarial examples.

Table 14: 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

Its derivative gradient to the logit vector $\mathbf{Z}$ is as follows:

$$\begin{aligned} g_i &= \frac{1-K}{K}, i = gt; \\ g_i &= \frac{1}{K}, i \neq gt, \end{aligned} \tag{23}$$

which can be rewritten as

$$\frac{\partial L_{RCE}}{\partial \mathbf{Z}} = \frac{1}{K} - \mathbf{Y}_{gt} \tag{24}$$