[CVPR2022]

Supplementary Material

A. Experimental Setup

A.1. General Setup

Following previous works [2, 3, 6], we evaluate our proposed approach on an ImageNet-compatible dataset composed of 1000 images. Consistent with prior works, we set the maximum perturbation magnitude to $L_\infty = 16/255$. 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.

A.2. 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.

B. Additional Results for supporting our proposed RCE loss

Table 1. Results of transfering from CNN (ResNet50) to ViT and MLP backbones. Each entry represents the ICR/targeted ASR@1 (%).

	non-targeted Acc.	ICR	OLNR	NLOR	NRT	CosSim
CE	100.00	629.11	604.28	74.50	255.81	0.37
CW	100.00	276.59	258.51	5.28	237.44	0.48
LL	100.00	409.08	398.66	963.84	300.79	0.10
FDA	99.00	504.42	498.90	430.81	297.03	0.13
RCE(Ours)	100.00	1000.00	984.15	477.37	346.03	-0.15
RCE(LL)	100.00	567.13	555.13	996.23	342.56	-0.12

Smaller ϵ in white-box attack. Table 1 reports the white-box results with an allowed ϵ=4/255 on ResNet50. The trend mirrors that with ϵ set $16/255$ in the main manuscript.

Transfer from CNN to ViT and MLP. The results of transfering from CNN (ResNet50) to ViT and MLP are reported in Table 2. We observe that our proposed RCE loss outperforms existing ones by a large margin.

DenseNet121 as the surrogate model. 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 3 for non-targeted attack and in Table 4 for targeted attack.

Table 2. Results of transfering from CNN (ResNet50) to ViT and MLP. Each entry represents the ICR/targeted ASR@1 (%). ResNet50 is trained by l_2-PGD atttack with ϵ set to 0.5.

		ViT B16	ViT L16	MLP-M B16	MLP-M L16
I-FGSM	CE	187.03/3.0	218.54/7.0	147.16/9.0	211.69/6.0
	Po-Trip	140.45/23.0	159.86/21.0	83.11/20.0	139.43/12.0
	RCE	48.53/28.0	66.58/25.0	43.98/31.0	71.57/17.0
MI-DI-TI	CE	42.74/42.0	51.73/41.0	19.00/38.0	96.81/13.0
	Po-Trip	41.86/39.0	41.37/37.0	44.77/37.0	98.00/18.0
	RCE	23.35/49.0	32.67/54.0	10.90/44.0	46.62/28.0

C. Top-k Attack Strength is Transferable

Transferability and Strength As was discussed in the main manuscript, source ASR@1 (source ICR) and target ASR@1 (target ICR) both increase over iterations. Now, we show this relationship more explicitly in Figure 1.

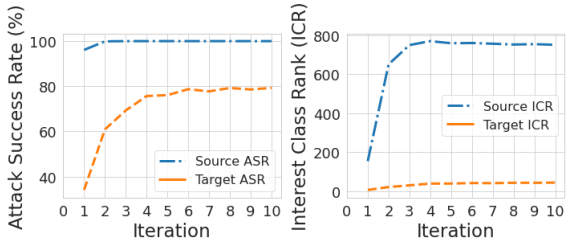

Figure 1. **Left:** Source ASR@1 and Target ASR@1 (average of several target networks) over iterations. **Right:** Source ICR and Target ICR (average of several target networks) over iterations.

Single Sample Analysis. Figure 2 shows the results of ICR in untargeted and targeted settings over T iterations for both white-box and black-box models for a (randomly chosen) single sample. The results show that top-k attack strength based on the metric of ICR is transferable on a single sample (see the similar trend of ICR with more iterations on the white-box and black-box models).

D. Zero Sum Constraint

D.1. Phenomenon

As introduced in the main manuscript, 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 differ-

Table 3. 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 4. 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

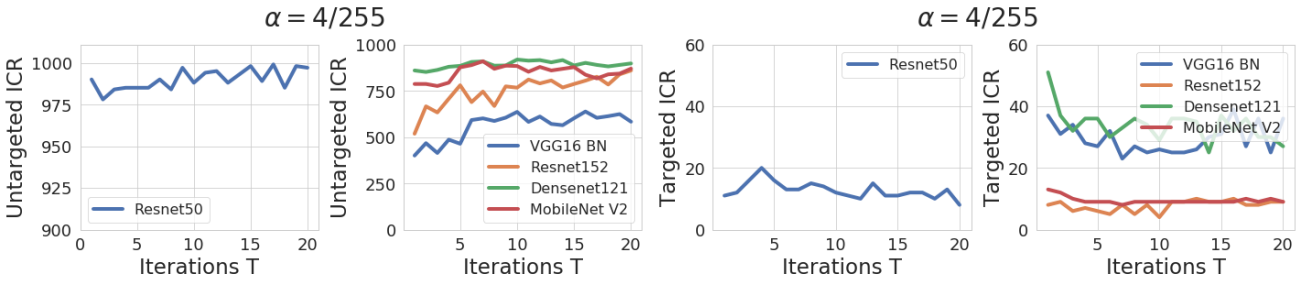

Figure 2. Untargeted (left two) and targeted (right two) ICR over T iterations in both white-box (ResNet50) and black-box scenarios on a (randomly chosen) single sample.

Table 5. 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

ent network architectures. From the results in Table 5 and Table 6, 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 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 occurring, not just due to very small values in $\mathbf{Z}$, we further

Table 6. 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

present the absolute sum, *i.e.* $\sum_{i=1}^{i=K} |z_i|$. Additionally, the relatively large values for the standard deviation, the mini-

Table 7. 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

Table 8. Zero sum experiment results on the CIFAR10 dataset with various network architectures. The metrics reported are the average (with standard deviation) over the validation samples. Weights of the network were initialized with $\sim \mathcal{N}(0,\ 1^2)$.

Network	Sum	Abs. Sum	Std	Min	Max	Accuracy
RN56	0.00±0.00	29.93±4.99	4.38±0.70	-3.79±0.78	11.75±2.62	93.59%
VGG19bn	0.00±0.00	19.12±1.57	3.05±0.22	-2.24±0.37	8.71±1.06	93.38%

mum, 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 7. The results show that the zero-sum constraint also holds for adversarial examples.

D.2. Influence of Network Weights Initialization

We investigate the influence of the network weight initialization on the zero sum phenomenon. We observe that zero sum constraint is still present even with different weight initialization parameters suggesting that there is no influence of the network weight initialization on the phenomenon. The results are reported in Table 8, Table 9, Table 10, Table 11, and Table 12.

Table 9. Zero sum experiment results on the CIFAR10 dataset with various network architectures. The metrics reported are the average (with standard deviation) over the validation samples. Weights of the network were initialized with $\sim \mathcal{N}(1,\ 1^2)$.

Network	Sum	Abs. Sum	Std	Min	Max	Accuracy
RN56	0.00±0.00	32.13±5.81	4.59±0.78	-4.24±0.92	12.01±2.81	92.48%
VGG19bn	0.00±0.00	22.86±2.73	3.44±0.40	-3.63±0.55	9.19±1.41	93.04%

D.3. Influence of Unbalanced Dataset

Additionally, we confirm that the zero sum constraint is valid when the dataset classes are unbalanced. For 10 classes in CIFAR10, we test two variants of unbalance. In the first setup, we set the number of samples for each class

Table 10. Zero sum experiment results on the CIFAR10 dataset with various network architectures. The metrics reported are the average (with standard deviation) over the validation samples. Weights of the network were initialized with $\sim \mathcal{N}(3,\ 3^2)$.

Network	Sum	Abs. Sum	Std	Min	Max	Accuracy
RN56	0.00±0.00	39.15±8.08	5.44±1.21	-6.86±2.69	13.09±3.76	91.77%
VGG19bn	0.00±0.00	25.40±3.94	3.87±0.54	-6.35±2.01	9.20±1.28	92.70%

Table 11. Zero sum experiment results on the CIFAR10 dataset with various network architectures. The metrics reported are the average (with standard deviation) over the validation samples. Weights of the network were initialized with $\sim \mathcal{N}(5,\ 5^2)$.

Network	Sum	Abs. Sum	Std	Min	Max	Accuracy
RN56	0.00±0.00	31.46±6.77	4.30±0.93	-4.50±1.09	10.58±3.20	90.20%
VGG19bn	0.00±0.00	20.87±2.44	3.24±0.35	-3.28±0.55	8.81±1.28	92.62%

Table 12. Zero sum experiment results on the CIFAR100 dataset with various network architectures. The metrics reported are the average (with standard deviation) over the validation samples. Weights of the network were initialized with $\sim \mathcal{N}(5,\ 5^2)$.

Network	Sum	Abs. Sum	Std	Min	Max	Accuracy
RN56	0.01±0.00	377.77±75.17	4.79±0.96	-10.98±2.36	14.10±4.35	63.83%
VGG19bn	0.00±0.00	558.74±83.81	6.78±0.98	-15.31±2.52	17.14±2.80	68.32%

Table 13. Zero sum experiment results on the unbalanced CIFAR10 dataset (linear change) with various network architectures. The metrics reported are the average (with standard deviation) over the validation samples.

Network	Sum	Abs. Sum	Std	Min	Max	Accuracy
RN56	0.00±0.00	25.70±4.81	3.82±0.73	-2.99±0.73	10.41±2.60	91.22%
VGG19bn	0.00±0.00	20.55±2.31	3.24±0.37	-3.60±0.73	8.70±1.35	89.89%

to change linearly (from 5000 to 500). The result is in Table 13. In the second setup, we adopt the unbalanced class setting as the common long-tail problem setup [1] with the imbalance factor set to 50 or 100. The results for this setup are in Table 14 and Table 15. The result for CIFAR100 dataset with imbalance factor of 50 is in Table 16.

Table 14. Zero sum experiment results on the unbalanced CIFAR10 dataset (imbalance factor 50) with various network architectures. The metrics reported are the average (with standard deviation) over the validation samples.

Network	Sum	Abs. Sum	Std	Min	Max	Accuracy
RN56	0.00±0.00	26.37±5.27	3.77±0.92	-3.21±0.80	9.84±3.32	77.70%
VGG19bn	0.00±0.00	23.38±5.01	3.39±0.59	-3.38±1.07	8.79±1.85	78.69%

D.4. Possible Explanation

Admittedly, 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

Table 15. Zero sum experiment results on the unbalanced CIFAR10 dataset (imbalance factor 100) with various network architectures. The metrics reported are the average (with standard deviation) over the validation samples.

Network	Sum	Abs. Sum	Std	Min	Max	Accuracy
RN56	0.01±0.01	25.92±4.79	3.68±0.84	-3.10±0.73	9.57±3.12	70.61%
VGG19bn	0.00±0.00	23.70±5.30	3.44±0.69	-3.85±1.53	8.63±2.04	70.59%

Table 16. Zero sum experiment results on the unbalanced CIFAR100 dataset (imbalance factor 50) with various network architectures. The metrics reported are the average (with standard deviation) over the validation samples.

Network	Sum	Abs. Sum	Std	Min	Max	Accuracy
RN56	0.07±0.04	246.12±49.72	3.21±0.65	-6.49±1.48	11.88±3.92	44.46%
VGG19bn	0.01±0.01	225.33±32.34	2.93±0.39	-4.26±0.64	12.08±2.40	45.40%

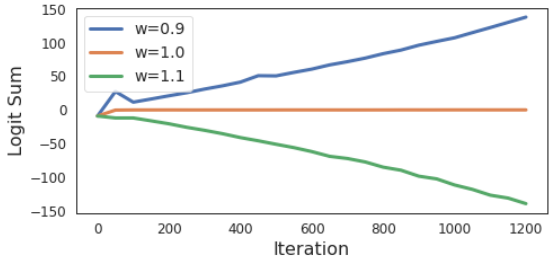

Figure 3. Influence of w on the $\sum_{i=1}^{i=K} z_i$ in the training stage.

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.

E. 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, shown in the main manuscript.

E.1. 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:

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

with $p_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}$. Eq. 1 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.* $i = j = gt$:

$$\begin{aligned}\frac{\partial L_{CE}}{\partial z_{gt}} &= \frac{\partial(-\log p_{gt})}{\partial z_{gt}} \\ &= -\frac{1}{p_{gt}}\frac{\partial p_{gt}}{\partial z_{gt}} \\ &= -\frac{1}{p_{gt}}(p_{gt}(1-p_{gt}) \\ &= p_{gt} - 1.\end{aligned} \tag{2}$$

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:

$$\begin{aligned}\frac{\partial p_j}{\partial z_i} &= \frac{\partial(\frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}})}{\partial z_i} \\ &= -\frac{e^{z_j}e^{z_i}}{(\sum_{k=1}^{K} e^{z_k})^2} \\ &= -p_j p_i.\end{aligned} \tag{3}$$

With Eq. (3), 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.* $i \neq gt$:

$$\begin{aligned}\frac{\partial L_{CE}}{\partial z_i} &= \frac{\partial(-\log p_{gt})}{\partial z_i} \\ &= -\frac{1}{p_{gt}}\frac{\partial p_{gt}}{\partial z_i} \\ &= -\frac{1}{p_{gt}}(-p_{gt}p_i) \\ &= p_i.\end{aligned} \tag{4}$$

From Eq. 2 and Eq. 4, we arrive at the formulation presented in the main manuscript:

$$\frac{\partial L_{CE}}{\partial \mathbf{Z}} = \mathbf{P} - \mathbf{Y}_{gt}, \tag{5}$$

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 as:

$$\frac{\partial L_{CE(LL)}}{\partial \mathbf{Z}} = \mathbf{Y}_{LL} - \mathbf{P}. \tag{6}$$

E.2. CW Loss

CE and CW loss are the two most widely used losses for the white-box attack [4, 5]. In the above, we derive the gradient for CE loss and we further conduct a similar derivation for CW loss which is denoted as $L_{CW} = z_j - z_{gt}$ [4, 5] 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:

$$\frac{\partial L_{CW}}{\partial \mathbf{Z}} = \mathbf{Y}_j - \mathbf{Y}_{gt}. \tag{7}$$

E.3. Relative Cross-Entropy (RCE) Loss

With Eq. 5, we can calculate the derivative of the proposed RCE loss:

$$\begin{aligned}
\frac{\partial L_{RCE}}{\partial \mathbf{Z}} &= \frac{\partial (L_{CE_{gt}} - \frac{1}{K}\sum_{k=1}^{K} L_{CE_k})}{\partial \mathbf{Z}} \\
&= \frac{\partial L_{CE_{gt}}}{\partial \mathbf{Z}} - \frac{1}{K}\sum_{k=1}^{K}\frac{\partial L_{CE_k}}{\partial \mathbf{Z}} \\
&= \mathbf{P} - \mathbf{Y_{gt}} - \frac{1}{K}\sum_{k=1}^{K}(\mathbf{P} - \mathbf{Y_k}) \\
&= \frac{1}{K}\mathbf{1} - \mathbf{Y}_{gt}.
\end{aligned} \tag{8}$$

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

F. CW and RCE are Special Cases of CE

F.1. Derivative of the Temperature Scaled CE-loss

The derivative of the CE-Loss with temperature scaling can be written as:

$$\frac{\partial L_{CE(Temp)}}{\partial \mathbf{Z}} = \frac{1}{T_e}(\mathbf{P}_e - \mathbf{Y}_{gt}), \tag{9}$$

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:

$$\begin{aligned}
\frac{\partial p_e^i}{\partial z_i} &= \frac{\partial (\frac{e^{z_i/T_e}}{\sum_{k=1}^{K} e^{z_k/T_e}})}{\partial z_i} \\
&= \frac{1}{T_e}(p_e^i(1 - p_e^i)).
\end{aligned} \tag{10}$$

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

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

Analogous to Eq. (2) and Eq. (4), we can calculate the derivatives for the CE Loss with T_e. For the case $i = gt$ we arrive at:

$$\begin{aligned}
\frac{\partial L_{CE(Temp)}}{\partial z_{gt}} &= \frac{\partial(-\log p_e^{gt})}{\partial z_{gt}} \\
&= \frac{1}{T_e}(p_e^{gt} - 1).
\end{aligned} \tag{12}$$

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

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

With Eq. (12) and Eq. (13), we finally arrive at Eq. (9).

F.2. Scale-invariant Property of the Gradient Derivative

As highlighted in the main manuscript, only the direction of the derivative matters and the scale is irrelevant because FGSM 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 scale factor. We can derive:

$$\begin{aligned}
sign(\frac{\partial L_B}{\partial \mathbf{X}}) &= sign(\frac{\partial \mathbf{Z}}{\partial \mathbf{X}}\frac{\partial L_B}{\partial \mathbf{Z}}) \\
&= sign(s\frac{\partial \mathbf{Z}}{\partial \mathbf{X}}\frac{\partial L_A}{\partial \mathbf{Z}}) \\
&= sign(\frac{\partial \mathbf{Z}}{\partial \mathbf{X}}\frac{\partial L_A}{\partial \mathbf{Z}}) \\
&= sign(\frac{\partial L_A}{\partial \mathbf{X}})
\end{aligned} \tag{14}$$

F.3. Relationship to Other Loss Functions

The probability of the i-th class in $\mathbf{P}_e$ is shown as:

$$p_e^i = \frac{e^{z_i/T_e}}{\sum_{k=1}^{k=K} e^{z_k/T_e}} \tag{15}$$

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

$$\begin{aligned} \frac{p_e^x}{p_e^y} &= \frac{\frac{e^{z_x/T_e}}{\sum_{k=1}^{k=K} e^{z_k/T_e}}}{\frac{e^{z_y/T_e}}{\sum_{k=1}^{k=K} e^{z_k/T_e}}} \\ &= \frac{e^{z_x/T_e}}{e^{z_y/T_e}} \\ &= e^{(z_x - z_y)/T_e} \\ &> 1 \end{aligned} \tag{16}$$

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:

$$\begin{aligned} \lim_{T_e \to \infty} \frac{p_e^x}{p_e^y} &= \lim_{T_e \to \infty} e^{(z_x - z_y)/T_e} \\ &= 1 \end{aligned} \tag{17}$$

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. (9) can be further derived as follows:

$$\begin{aligned} \frac{\partial L_{CE(Temp)}}{\partial \mathbf{Z}} &= \frac{1}{T_e}(\mathbf{P}_e - \mathbf{Y}_{gt}) \\ &= \frac{1}{T_e}(\frac{1}{K}\mathbf{1} - \mathbf{Y}_{gt}) \end{aligned} \tag{18}$$

Given the scale-invariant property indicated by Eq. (14), Eq. (18) is equivalent to the derived gradient in Eq. (8) 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:

$$\begin{aligned} \lim_{T_e \to 0} \frac{P_e^x}{P_e^y} &= \lim_{T_e \to 0} e^{(z_x - z_y)/T_e} \\ &= \infty \end{aligned} \tag{19}$$

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. 9 can be rewritten as

$$\frac{\partial L_{CE(Temp)}}{\partial \mathbf{Z}} = \frac{m}{T_e}(\mathbf{Y}_j - \mathbf{Y}_{gt}), \tag{20}$$

Given the scale-invariant property indicated by Eq. 14, Eq. 20 is equivalent to the derived gradient in Eq. 7 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.

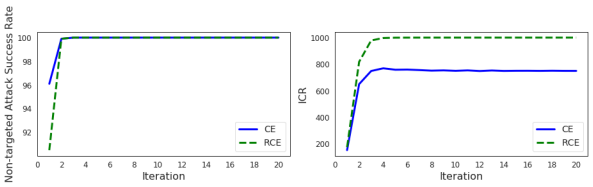

Figure 4. Non-targeted attack success rate and ICR on the white-box ResNet50.

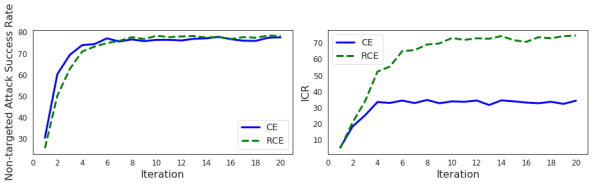

Figure 5. Non-targeted attack success rate and ICR on the black-box DenseNet121.

G. Limitation of RCE Loss in the Early Iterations

As indicated in the main manuscript, 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 4 and Figure 5, respectively. We observe that in the early iterations, CE outperforms our proposed RCE loss, especially for the metric of attack success rate.

References

[1] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 3

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

[3] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*, 2019. 1

[4] 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. 5

[5] Saehyung Lee, Hyungyu Lee, and Sungroh Yoon. Adversarial vertex mixup: Toward better adversarially robust generalization. In *CVPR*, 2020. 5

[6] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. In *CVPR*, 2020. 1