TOWARDS SIMPLE YET EFFECTIVE TRANSFERABLE TARGETED ADVERSARIAL ATTACKS

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

ABSTRACT

Transfer-based targeted adversarial attacks against deep image classifiers remain an open issue. Depending on which parts of the deep neural network are explicitly incorporated into the loss function, the existing methods can be divided into two categories: (a) feature space attack and (b) output space attack. One recent work has shown that attacking the feature space outperforms attacking the output space by a large margin. However, the elevated attack success comes at the cost of requiring to train layer-wise auxiliary classifiers for each corresponding target class together with a greedy search to find the optimal layers. In this work, we revisit the output space attack and improve it from two perspectives: First, we identify over-fitting as one major factor that hinders transferability, for which we propose to augment the network input and/or feature layers with noise. Second, we propose a new cross-entropy loss with two ends: one for pushing the sample far from the source class, *i.e*. ground-truth class, and the other for pulling it close to the target class. We find that given sufficiently large iterations, our approach can outperform the state-of-the-art feature space method by a large margin.

1 INTRODUCTION

Since the discovery of adversarial examples, numerous attack methods, both image-specific ones (Szegedy et al., 2013; Moosavi-Dezfooli et al., 2016; Carlini & Wagner, 2017; Goodfellow et al., 2015; Kurakin et al., 2017; Madry et al., 2018; Dong et al., 2018) and universal adversarial perturbations (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) have been proposed. Regardless of the attack methods, one property of adversarial examples (Szegedy et al., 2013; Goodfellow et al., 2015) is their transferability (Kurakin et al., 2017; Zhou et al., 2018; Dong et al., 2018), which allows adversarial examples to be applied in a black-box scenario. Specifically, an adversarial example crafted on one model is also often effective in fooling another, previously unknown model (Liu et al., 2017). This property further increases the concern of the applicability of DNNs for security-sensitive applications in the real world. The body of research for transferable attacks has grown rapidly in the past few years (Dong et al., 2019; Wu et al., 2020). However, most of the proposed transferable attacks address the non-targeted scenario, and only a limited number of works investigate the targeted case, where the adversarial example is strictly required to be classified as a predefined target class. One recent work (Inkawhich et al., 2020b) has shown that optimizing the loss on the feature layers outperforms the baselines that only optimize the loss on the model output layer by a large margin with the assumption that feature space attack is more transferable than the output space. However, the feature space (targeted) attack in (Inkawhich et al., 2020b) requires training layer-wise auxiliary networks for each target class and a greedy search to find the optimal combination of feature layers. On the other hand, the widely used output space attack that optimizes the loss directly on the model output space is much more simple and straightforward, however, the resulting targeted transferability performance seems to be very limited. In this work, we revisit the output space attack method and identify over-fitting on the input and/or the feature space as one major factor that limits its transferability. To this end, we propose to augment the network input as well as the feature space with noise, which significantly improves the transferability. Moreover, we conjecture that it is not

*Equal contribution

enough to just minimize the distance to the target class without getting the sample far away from the ground-truth class. Thus, we propose a new loss that explicitly pulls the sample close to the target class and pushes the sample far from the ground-truth class, which even outperforms the SOTA output space targeted attack method with the Poincaré distance-based loss (Li et al., 2020a). The performance boost caused by our approach is significant, *e.g.* an increase from 4.6% to 42.6% targeted success rate for a black-box attack transferred from a ResNet50 to Inception-V3.

2 Related Work

Transfer-based Black-box Non-targeted Attacks. Early works (Goodfellow et al., 2015; Kurakin et al., 2017) in the transfer-based black-box attack are built directly upon the white-box attacks. Follow-up works enhance the transferability by ensembles of white-box models (Liu et al., 2017; Tramèr et al., 2018). (Dong et al., 2018) introduce a momentum term in the perturbation update to smooth the gradient, called MI-FGSM. Low frequency constraint and translation-invariance are considered in (Dong et al., 2019) to evade defended models. (Xie et al., 2019) improve the transferability by resizing the adversarial examples to encourage input diversity. Fine-tuning adversarial examples with the intermediate level attack has been studied in (Huang et al., 2019; Li et al., 2020b). Recently, adjusting the gradients in the backward propagation has been shown to improve the transferability by weighting the gradients through the shortcut and residual module (Wu et al., 2020) or backpropagating linearly (Guo et al., 2020). The above methods are all based on the optimization through the model output layer, showing reasonable performance for non-targeted attacks with very limited success in the more challenging targeted scenario.

Transfer-based Targeted Attacks. Overall, so far the investigation on the transfer-based targeted attack is still limited. (Inkawhich et al., 2019) achieve transferable targeted attack by generating perturbation in feature space, representing the "target" as a single point at a certain feature layer. The intuition is to minimize the distance between the adversarial image and an image of the target class in feature space with a simple L_2 loss. This method is shown to be sensitive to the choice of the target image and also difficult to scale to larger models or datasets. To address this concern, (Inkawhich et al., 2020a) improve the descriptive representation of the target class by modeling layer-wise and class-wise feature distributions. Recently, the same group of authors further extended this from a single layer to multiple layers including the final output layer with the Cross-Entropy loss (Inkawhich et al., 2020b). This line of work of generating the perturbations in the feature space is much more cumbersome for two major reasons: (a) a decision has to be made which layers to include together with a relatively cumbersome loss design (b) training an auxiliary classifier for each target class. Recently, a work (Li et al., 2020a) that conducts optimization in the output space also achieves reasonable transfer-based targeted performance with the Poincaré ball distance and the triplet loss. Our work is inspired by both of the two lines of work but prefers simple techniques that can non-trivially boost the performance.

3 Methodology

Output Space White-box Adversarial Attacks. Suppose, we are given a data distribution $\mathcal{D}$ of input pairs (x, y) of samples $x \in \mathbb{R}^d$, and their corresponding ground-truth label $y \in \{1, \ldots, k\}$. Additionally, consider a pre-trained classifier $f_\theta(x) : \mathcal{X} \to \mathcal{Y}$, parameterized through the weights θ (from here on omitted), which outputs a label as the prediction given an input sample. We wish to craft an adversarial example $\tilde{x} = x + \delta$ to fool the classifier, *i.e.* $f(\tilde{x}) \neq y$. Moreover, the adversarial example is required to be visually close to the original sample, *i.e.* $||\delta|| \leq \epsilon$, where $|| \cdot ||$ indicates the l_p-norm. Throughout this work, following prior works we adopt the l_∞-norm constraint and set $l_\infty = 16/255$ for all our experiments. To generate adversarial examples, the objective is to maximize a certain loss $\mathcal{L}(f(\tilde{x}), y)$, for which the cross-entropy (CE) is a common choice.

3.1 Techniques for Improving Transferability

Input Noise Perturbation. Augmentation with noise is a common technique to increase the generalization capabilities of DNNs (Hussein et al., 2017; Akbiyik, 2019). We propose to leverage noise augmentation on the input space as a method to increase the transferability of adversarial examples. To perturb the input space we choose uniform noise $\nu_I \in \mathcal{U}(-\tau_I, \tau_I)$.

Feature Noise Perturbation. Straightforwardly, the noise can also be added to the intermediate feature layers for further improving the network generalization. We are not the first to inject noise into feature layers, such as VAE (Kingma & Welling, 2013) adding Gaussian noise to hidden layers, ANL (You et al., 2019) perturbing the feature layers with adversarial noise, and PNI (He et al., 2019) injects noise to the network parameter weight to improve adversarial robustness. To our best knowledge, we are the first to perturb the feature layer with noise to improve the transferability of adversarial examples. Specifically, we apply uniform random noise $\nu_F \in \mathcal{U}(-\tau_F, \tau_F)$ to the output of each convolutional layer in the network.

Note, that the input perturbation and the one on the intermediate feature maps are orthogonal to each other. As our later experiments will show (refer to Table 1), the non-targeted attack success achieves on average above 95% with the proposed noise augmentation techniques. Such high attack rates indicate a certain saturation of this task. Hence, in the following, we focus on the transferability of adversarial examples in a targeted attack scenario. Note, that both proposed noise perturbation techniques can also be applied in this more challenging setting.

Novel Loss Function. The objective of a targeted attack is to craft an adversarial example that is misclassified as a predefined target class, $f(\tilde{x}) = y_t$. The CE loss with the target class as the label is a common choice in this scenario. However, this targeted CE loss only maximizes the probability of the target class without explicitly encouraging the sample to decrease the probability of the ground-truth class. We conjecture that the effectiveness of adversarial examples can be increased by not only increasing the logit of the target class but also decreasing the probability of the ground-truth class. Inspired by this motivation, we design a simple loss that combines the two CE losses: one for pushing the sample far from the source class and one for pulling the sample close to the target class. We term this CE loss with two ends *push-pull* (PP) loss and it is formulated as follows:

$$PP(Z, y_{gt}, y_{tar}) = CE_{tar}(Z, y_{tar}) - \beta CE_{gt}(Z, y_{gt}) \tag{1}$$

where Z indicates the model output and β is a weight for balancing the two CE loss terms. Empirically, we find that it achieves satisfactory performance when β is set to 1.

4 Experiments

4.1 Experimental Setup

Dataset. 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. Additional implementation details are presented in the Appendix.

Baseline and Metric. Since MI, DI, TI have been shown in numerous works to individually improve the transferability and their effect is compatible with each other, in this work, we use the combination of them as a strong baseline in our work. We indicate this combination as MI-DI-TI. We evaluate the transferability with the non-targeted or targeted attack success rate (ASR). The non-targeted ASR indicates the percentage of samples that are classified differently from the ground-truth label. The targeted ASR indicates the percentage of samples that are classified as the target class.

4.2 Transfer-based Black-box Attacks

Non-Targeted Attacks. We evaluate our proposed techniques of adding noise to the network input and/or feature layers for the relatively simpler scenario of the non-targeted attack. The results are shown in Table 1. We note that simply adding noise to the network input can significantly improve the transfer rate but it is not as effective as adding the noise to the feature layers. Combining the two noise perturbation techniques leads to the best performance. On average, from ResNet50 to 5 common DNN architectures, our simple technique can improve the non-targeted attack success rate from 95.9% to 97.6%. Due to this saturation for the non-targeted black-box attack case, we shift our focus to the more challenging task

Table 1: Non-targeted ASR for the non-targeted MI-DI-TI-FGSM attack with ResNet50 as the substitute model.

Attack	DN121	VGG16	RN152	MNv2	IncV3	Avg.
CE	99.2	98.4	99.2	98.4	84.2	95.9
CE + I_{Aug}	99.5	98.3	99.4	97.4	86.5	96.2
CE + F_{Aug}	99.4	99.2	99.9	98.9	87.8	97.0
CE + I_{Aug} + F_{Aug}	99.6	99.2	99.7	98.7	90.7	97.6

of the transfer-based targeted attack. Nonetheless, we will additionally report the non-targeted attack success rates.

Targeted Attacks. In the following the detailed performances of the three introduced techniques to improve targeted adversarial transferability are illustrated. The quantitative results are presented in Table 2. Comparing only the PP-loss with the CE loss or the recently introduced Po-Trip (Li et al., 2020a) loss it becomes apparent that already the proposed PP loss function improves the targeted attack success rate by a significant margin. For example, using a ResNet50 as the substitute model, the PP loss function improves the targeted success rate on average from 25.3% and 33.5 for the CE and Po-Trip, respectively to 50.6%. Applying the PP loss does also outperform the result reported in Inkawhich et al. (2020b), here indicated as FDA$^{(5)}$+xent, which is considered as the SOTA approach among the feature space attacks. This approach incorporates the information of 5 intermediate feature layers into the loss function, as well as applying the cross-entropy loss. The PP loss can further be combined with both introduced noise augmentation techniques to improve adversarial transferability. Among the two noise augmentation techniques, the noise perturbation on the intermediate feature layers leads to higher performance gains for the targeted attack success rate. The best targeted success rates can be achieved when all three introduced techniques are applied jointly. With an average targeted attack success rate of 73.8%, the adversarial examples crafted with our techniques constitute a significantly more dangerous threat to DNNs than the ones generated with the CE or Po-Trip loss.

Table 2: Non-targeted ASR/targeted ASR for a targeted MI-DI-TI attack with a single substitute model (ResNet50) in the targeted attack scenario.

Attack	DN121	VGG16	RN152	MNv2	IncV3	Avg.
CE	84.2/40.2	88.6/28.0	82.6/43.1	84.7/10.4	52.9/4.6	78.6/25.3
Po-Trip	84.0/56.7	86.0/33.1	83.0/55.5	81.5/15.1	51.0/7.1	77.1/33.5
FDA$^{(5)}$+xent	90.9/57.9	88.8/43.5	89.7/51.6	86.4/22.9	-	-
PP	97.6/73.1	97.8/62.5	98.2/78.2	95.0/28.5	71.8/10.8	92.1/50.6
PP + I_{Aug}	98.8/78.3	98.5/68.5	99.3/82.3	96.7/38.4	79.6/21.4	94.6/57.8
PP + F_{Aug}	99.8/87.6	99.7/82.6	99.8/90.5	97.7/56.0	80.1/28.7	95.4/69.1
PP + I_{Aug} + F_{Aug}	99.9/87.2	99.8/81.0	100.0/90.8	99.0/67.2	87.4/42.6	97.2/73.8

Ensemble-based Attacks. A common approach to further improve the transferability of adversarial examples is to leverage multiple substitute models instead of a single one. We adopt ResNet50 and DenseNet121 as substitute models and follow the ensemble-based strategy proposed in (Liu et al., 2017) to craft adversarial examples. The performance of the ensemble-based attack evaluated on various DNN architectures is presented in Table 3. Overall similar trends for the attack with a single model can be observed. Using the PP loss results in an average targeted attack success rate of 56%, a significant performance gain over the CE and the Po-Trip loss. Additionally, applying uniform noise augmentation on the input layer or the intermediate feature layers leads to further performance gains. The best results can be achieved with all three techniques combined, with an average targeted success rate of 81%. The beneficial effect of ensemble models can also clearly be observed. The best targeted success rate for ResNet50 is on average improved by about 7% through an ensemble with DenseNet121.

Table 3: Non-targeted ASR/Targeted ASR for a targeted MI-DI-TI attack with an ensemble of two substitute models.

Source	Attack	VGG16	RN152	DN201	MNv2	IncV3	IncV4	IncRes	Avg.
RN50 + DN121	CE	93.9/54.9	91.1/68.5	95.4/86.2	90.8/24.4	64.7/15.6	66.5/14.2	49.6/8.3	78.86/39.0
	Po-Trip	88.2/44.9	84.7/63.5	92.6/82.9	83.9/21.7	59.3/14.6	61.2/13.1	46.1/7.2	73.71/35.0
	PP	99.2/81.6	99.3/87.2	100.0/93.7	97.7/51.0	84.8/33.2	85.9/29.4	68.1/17.5	90.71/56.0
	PP + I_{Aug}	99.8/82.2	99.8/90.7	100.0/94.3	98.8/61.2	90.8/50.4	91.9/45.0	78.9/33.0	94.29/65.0
	PP + F_{Aug}	100.0/92.4	100.0/94.3	100.0/95.3	99.9/81.6	94.2/63.9	94.6/62.9	81.5/41.4	95.74/76.0
	PP + I_{Aug} + F_{Aug}	**100.0/90.9**	**100.0/94.3**	**100.0/94.6**	**99.9/85.3**	**97.8/74.2**	**97.3/73.0**	**88.6/55.0**	**97.66/81.0**

5 CONCLUSION

We address the challenging task of the transfer-based targeted attack. Specifically, we adopt the output space attack method as the baseline and improve it with noise perturbation and PP-loss. Despite its simplicity, our approach leads to a significant performance boost, resulting in a new SOTA record. From ResNet50 to Inception-v3, our approach improves the performance from 4.6% to 42.6%.

REFERENCES

Murtaza Eren Akbiyik. Data augmentation in training cnns: Injecting noise to images. 2019.

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.

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.

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

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

Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *CVPR*, 2019.

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.

Sarfaraz Hussein, Robert Gillies, Kunlin Cao, Qi Song, and Ulas Bagci. Tumornet: Lung nodule characterization using multi-view convolutional neural network with gaussian process. In *ISBI*, 2017.

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.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

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.

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

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

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016.

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

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, 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.

Zhonghui You, Jinmian Ye, Kunming Li, and Ping Wang. Adversarial noise layer: Regularize neural network by adding noise. In *ICIP*, 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.

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

A Appendix

A.1 Implementation Details

If not otherwise indicated, we use a step size $\alpha = 4/255$ and a total number of $N = 320$ iterations. Consistent with previous methods, we set the maximum perturbation magnitude to $\epsilon = 16$. To compare and use previous approaches, we try to follow the previous hyper-parameter settings as close as possible. Following (Dong et al., 2018) we set the momentum to 1, following (Xie et al., 2019) we set the probability of the stochastic input diversity p to 0.7, and adopt a kernel length of value 5 as in (Dong et al., 2019). For comparisons with (Li et al., 2020a) by reproducing their results, we set the weight of the triplet loss λ to 0.01 and the margin value γ to 0.007 as they reported.

A.2 Additional Experimental Results

Non-targeted Attack. Additionally to the results presented in the main manuscript, we report experimental results, when only MI-FGSM, TI-FGSM, and DI-FGSM are used as the baseline attack technique. Additionally, the results for the substitute model DenseNet-121 are presented. The results can be found in Table 4. Resembling the results in the main manuscript, adding noise to the input or feature layers during the adversarial example generation process improves the transfer rate of the adversarial examples. The best performances are obtained for the combination of the two proposed perturbation techniques. Using DenseNet-121 as the substitute model, the attack achieves an average attack success rate of 97.8% for the 5 target DNN architectures.

Targeted Attack. Supplementary to the results presented in Table 2, we present the targeted transferability results with DenseNet-121 as the substitute model in Table 5. Similar results as for ResNet-50

Table 4: Non-targeted ASR (%).

Substitute	FGSM variant	Attack	DN121	VGG16	RN152	MNv2	IncV3	Avg.
RN50	MI-FGSM	CE	86.0	83.3	90.8	84.5	50.4	79.0
		CE + I_{Aug}	93.4	89.6	96.1	89.1	62.8	86.2
		CE + F_{Aug}	100.0	96.5	92.6	95.5	73.7	91.7
		CE + I_{Aug} + F_{Aug}	100.0	97.5	95.4	96.3	83.0	94.4
	TI-FGSM	CE	79.6	77.7	88.3	78.2	39.1	72.6
		CE + I_{Aug}	88.1	83.5	93.8	82.8	50.4	79.7
		CE + F_{Aug}	90.7	87.7	95.1	84.8	47.9	81.2
		CE + I_{Aug} + F_{Aug}	92.6	90.3	97.1	89.4	56.9	85.3
	DI-FGSM	CE	97.4	96.9	97.9	94.6	57.9	88.9
		CE + I_{Aug}	97.1	96.8	97.6	95.0	66.6	90.6
		CE + F_{Aug}	97.6	98.2	98.5	96.3	60.7	90.3
		CE + I_{Aug} + F_{Aug}	98.0	97.8	98.3	96.7	69.1	92.0
Substitute	FGSM variant	Attack	RN50	VGG16	DN201	MNv2	IncV3	Avg.
DN121	MI-FGSM	CE	88.8	85.9	95.1	84.0	58.1	82.4
		CE + I_{Aug}	94.1	91.8	97.6	88.0	70.3	88.4
		CE + F_{Aug}	98.1	96.5	99.4	95.5	73.7	92.6
		CE + I_{Aug} + F_{Aug}	98.2	97.5	99.5	96.3	83.0	94.9
	TI-FGSM	CE	86.2	82.7	93.2	78.9	46.7	77.5
		CE + I_{Aug}	90.8	87.7	96.3	82.8	59.3	83.4
		CE + F_{Aug}	96.2	93.3	98.5	92.8	65.4	89.2
		CE + I_{Aug} + F_{Aug}	97.8	95.2	99.0	93.5	74.2	91.9
	DI-FGSM	CE	96.3	96.3	98.1	91.7	62.1	88.9
		CE + I_{Aug}	96.6	96.4	97.9	91.7	71.4	90.8
		CE + F_{Aug}	98.9	98.5	99.4	97.1	74.6	93.7
		CE + I_{Aug} + F_{Aug}	98.6	98.4	99.0	96.8	78.1	94.2
	MI-DI-TI-FGSM	CE	98.3	97.2	99.3	95.6	83.6	94.8
		CE + I_{Aug}	98.2	97.7	99.1	96.3	86.8	95.6
		CE + F_{Aug}	99.7	99.2	99.8	98.6	91.6	97.8
		CE + I_{Aug} + F_{Aug}	99.5	99.4	99.7	99.0	91.6	97.8

as the substitute model can be observed. Utilizing only the proposed PP-loss function, the targeted performance already increases by a significant margin, from 15.2 and 15.0 for the loss functions CE and Po-trip, respectively, to a targeted attack success rate of 37.6. Further combination of the loss with any of the proposed noise augmentations further boosts the targeted attack success rate. The best results are achieved when the PP loss is combined with the input and noise augmentation with an average targeted attack success rate of 70.9. In this case, only applying the PP loss does not outperform the results reported in Inkawhich et al. (2020b). However, using the PP-loss and noise augmentation on the feature layers or the combination of input and feature noise leads to a higher attack success rate than the one reported in Inkawhich et al. (2020b).

Table 5: Non-targeted ASR / targeted ASR (%) with a single substitute model. All Experiments are performed for MI-DI-TI).

	Attack	RN50	VGG16bn	DN201	MNv2	IncV3	Avg.
DN121	CE	78.4/16.4	80.3/11.3	79.2/40.5	83.6/5.3	51.6/2.3	74.6/15.2
	Po-Trip	72.3/16.4	74.5/9.4	72.9/41.6	78.6/4.6	48.0/3.2	69.3/15.0
	FDA$^{(5)}$+xent	92.2/50.1	92.1/48.0	95.6/77.1	88.8/24.4	-	92.2/49.9
	PP	95.6/48.0	95.7/39.8	97.7/76.6	93.1/13.3	72.7/10.5	91.0/37.6
	PP + I_{Aug}	96.5/57.4	97.1/47.1	98.9/81.0	94.2/19.5	78.5/17.1	93.0/44.4
	PP + F_{Aug}	99.4/82.8	99.5/72.3	100.0/91.3	97.4/40.3	86.5/29.7	96.6/63.3
	PP + I_{Aug} + F_{Aug}	99.5/84.7	99.5/76.8	99.9/92.3	98.4/55.0	90.8/45.9	97.6/70.9

A.3 Comparison with Inkawhich et al.

In Table 2 and Table 5 we compared our result with the results reported in Inkawhich et al. (2020b). Their approach can be categorized as a feature space attack since it incorporates multiple layers in the optimization function. Their approach requires a greedy search for the optimal layers and a careful choice of hyper-parameters to balance multiple losses as well as the training of class-wise

auxiliary classifiers for each target class (introducing additional hyper-parameter choices). Thus, it is non-trivial to reproduce their results, especially, since the authors did not open-source their code. Finally, we were not able to reproduce their results and therefore directly compared to their reported results. We note that the comparison might not be absolutely fair. For the evaluation in Inkawhich et al. (2020b) the authors randomly chose 15.000 images from the ImageNet validation set, while we evaluated our approach on the NeurIPS challenge dataset.