Appendix for Introducing Competition to Boost the Transferability of Targeted Adversarial Examples through Clean Feature Mixup

Algorithm 1 CFM-RDI-MI-TI

Input: A classifier f; a clean example $\mathbf{x}$; a target label y_t.
Input: Adversary's objective $\mathcal{L}$; the maximum iterations T; ℓ_∞ perturbation bounds ϵ; step size η; decay factor μ; Gaussian kernel $\boldsymbol{W}$ for TI.
Input: mixing probabilty p; upper bounds for mixing ratios α_{max} for CFM modules.
Output: An adversarial example $\mathbf{x}^{adv}$

```
f' = AttachCFM(f; p, α_max)        ▷ Attach CFM modules to conv and fc layers
Store clean features into CFM modules via f'(x)
g_1 = 0; x_1^adv = x
for t = 1 → T − 1 do
    Compute the gradients with RDI input transforms via f'
        ĝ_{t+1} = ∇_{x_t^adv} L(f'(RDI(x_t^adv)), y_t)        (1)
    g̃_{t+1} = μ · g_t + ĝ_{t+1} / ||ĝ_{t+1}||_1        ▷ Apply MI
    g_{t+1} = W ∗ g̃_{t+1}        ▷ Apply TI
    x_{t+1}^adv = x_t^adv − η · sign(g_{t+1})        ▷ Apply FGSM
    x_{t+1}^adv = Clip_x^ε(x_{t+1}^adv)
end for
x^adv = x_T^adv
return x^adv
```

$$\hat{\mathbf{g}}_{t+1} = \bigtriangledown_{\mathbf{x}_t^{adv}} \mathcal{L}(f'(RDI(\mathbf{x}_t^{adv})), y_t) \quad (1)$$

$$\tilde{\mathbf{g}}_{t+1} = \mu \cdot \mathbf{g}_t + \frac{\hat{\mathbf{g}}_{t+1}}{\|\hat{\mathbf{g}}_{t+1}\|_1}, \quad \mathbf{g}_{t+1} = \boldsymbol{W} * \tilde{\mathbf{g}}_{t+1}, \quad \mathbf{x}_{t+1}^{adv} = \mathbf{x}_t^{adv} - \eta \cdot \text{sign}(\mathbf{g}_{t+1}), \quad \mathbf{x}_{t+1}^{adv} = Clip_{\mathbf{x}}^{\epsilon}(\mathbf{x}_{t+1}^{adv})$$

A. Algorithm

The CFM method is compatible with many existing attack methods, and as an example, the pseudo-codes of the CFM-RDI-MI-TI method are described in Algorithm 1.

B. References to Pre-trained Models

B.1. Pre-trained Models on the ImageNet Dataset

We used a total of 16 models, and the sources of the pre-trained weights of the models are as follows.

The weights for the following six models are downloaded from TorchVision library[1]: VGG-16 [14], ResNet-18 (RN-18) [6], ResNet-50 (RN-50) [6], DenseNet-121 (DN-121) [8], MobileNet-v2 (MB-v2) [13], Inception-v3 (Inc-v3) [17].

The weights for the following nine models are downloaded from Pytorch Image Models (timm) library [20]: Xception (Xcep) [1], EfficientNet-B0 (EF-B0) [18], Inception ResNet-v2 (IR-v2) [16], Inception-v4 (Inc-v4) [16], Vision Transformer (ViT) [3], LeViT [5], ConViT [4], Twins [2], and Pooling-based Vision Transformer (PiT) [7]. The pre-trained weights for the adversarially trained RN-50 (adv-RN-50) [21] is provided by the official repository of [12].

The adv-RN-50 is adversarially trained on small ℓ_2-norm-constrained adversarial examples ($||\boldsymbol{\delta}||_2 \leq 0.1$), which is recently demonstrated to be effective in boosting the transfer success rate when used as a source model [15].

B.2. Pre-trained Models on the CIFAR-10 Dataset

The pre-trained weights for the following six models are provided by [11]: VGG-16 [14], ResNet-18 (RN-18) [6], ResNet-50 (RN-50) [6], DenseNet-121 (DN-121) [8], MobileNet-v2 (MB-v2) [13], and Inception-v3 (Inc-v3) [17].

We used four ensemble models composed of three ResNet-20 [6] networks (ens3-RN-20). They are trained under four defensive settings: standard training, ADP [10], GAL [9], and DVERGE [22]. The pre-trained weights for the four ensemble models are provided by [22].

C. Additional Experimental Results

C.1. Visualization of Generated Adversarial Examples

Figure 1, 2, 3, 4, 5 and 6 visualize the generated adversarial examples for qualitative comparison. We denoted the true and target classes below the clean images and computed the average attack success rates over the ten carefully selected pre-trained target models listed in Table 4. Note that all adversarial perturbations are constrained by the ℓ_∞-norm (i.e., $||\boldsymbol{\delta}||_\infty \leq \epsilon$ where we used $\epsilon = 16/255$).

[1] https://github.com/pytorch/vision

C.2. Extended Experimental Results With Additional Source Models and Baselines

Table 1 and Table 2 show the extended experimental results on the ImageNet-Compatible dataset with additional source models, i.e., adv-RN-50 and DN-121 in Table 1 and RN-50 and DN-121 in Table 2. For the additional source models, we used the same hyperparameters of CFM as in RN-50 (i.e., $\alpha_{max} = 0.75$ and $p = 0.1$). We also included the results of Admix with the number of scale copies of 5 (i.e., $m_1 = 5$ in [19]) and SI-CFM-RDI for more comprehensive comparisons. The Admix$_{m_1=5}$ follows the original setting of the Admix [19], which utilizes the SI technique in its internal loops.

C.3. Extended Experimental Results on the CIFAR-10 dataset

Table 3 shows the extended experimental results on the CIFAR-10 dataset, which additionally include the results of Admix$_{m_1=5}$ and SI-CFM-RDI with different source models (Inc-v3, VGG-16, and DN-121) for more comprehensive comparisons.

C.4. Experimental Results of Combined Attacks With Multiple Techniques

Table 4 shows the experimental results of various combinations of multiple attack techniques. The results demonstrate that CFM is compatible with existing attack methods, and various combinations with CFM can further improve the transferability of adversarial examples.

C.5. Experimental Results with Different Mixing Hyperparameters

Table 4 shows the experimental results on how the transfer success rates vary by changing the values of the mixing probability p and the upper bound of mixing ratios α_{max}. In this experiment, we used adv-RN-50 as the source model and evaluated the transfer success rates on the carefully selected ten target models. CFM achieves the highest success rate when $p = 0.1$ and $\alpha_{max} = 0.75$, but it also achieves comparable attack success rates at other values. This indicates that CFM is not very sensitive to the changes in hyperparameters and can achieve consistent performance improvement.

References

[1] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 1

[2] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[4] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021. 1

[5] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12259–12269, 2021. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[7] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11936–11945, 2021. 1

[8] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[9] Sanjay Kariyappa and Moinuddin K Qureshi. Improving adversarial robustness of ensembles with diversity training. *arXiv preprint arXiv:1901.09981*, 2019. 1

[10] Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, pages 4970–4979. PMLR, 2019. 1

[11] Huy Phan. huyvnphan/pytorch_cifar10, Jan. 2021. 1

[12] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *ArXiv preprint arXiv:2007.08489*, 2020. 1

[13] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 1

[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[15] Jacob Springer, Melanie Mitchell, and Garrett Kenyon. A little robustness goes a long way: Leveraging robust features for targeted transfer attacks. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[16] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the

Figure 1. Visualization of generated adversarial examples. The source model is RN-50. Each average targeted attack success rate was calculated over the ten carefully selected target models, which are more difficult to confuse. For example, an average targeted attack success rate of 50% means that 5 out of 10 target models recognize the adversarial example as the target class.

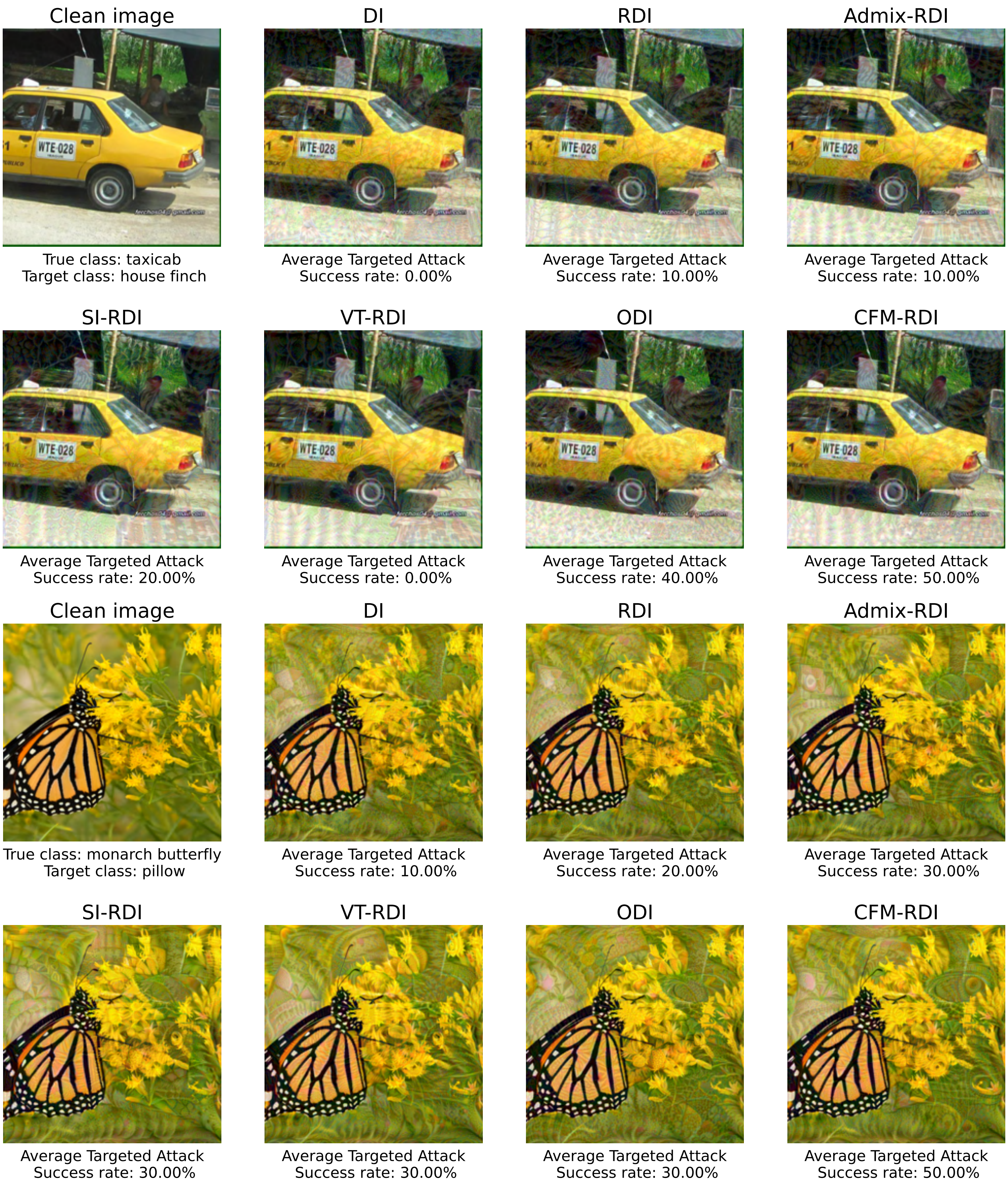

Figure 2. Visualization of generated adversarial examples. The source model is RN-50. Each average targeted attack success rate was calculated over the ten carefully selected target models, which are more difficult to confuse. For example, an average targeted attack success rate of 50% means that 5 out of 10 target models recognize the adversarial example as the target class.

Figure 3. Visualization of generated adversarial examples. The source model is adv-RN-50. Each average targeted attack success rate was calculated over the ten carefully selected target models, which are more difficult to confuse. For example, an average targeted attack success rate of 50% means that 5 out of 10 target models recognize the adversarial example as the target class.

Figure 4. Visualization of generated adversarial examples. The source model is adv-RN-50. Each average targeted attack success rate was calculated over the ten carefully selected target models, which are more difficult to confuse. For example, an average targeted attack success rate of 50% means that 5 out of 10 target models recognize the adversarial example as the target class.

Figure 5. Visualization of generated adversarial examples. The source model is Inc-v3. Each average targeted attack success rate was calculated over the ten carefully selected target models, which are more difficult to confuse. For example, an average targeted attack success rate of 50% means that 5 out of 10 target models recognize the adversarial example as the target class.

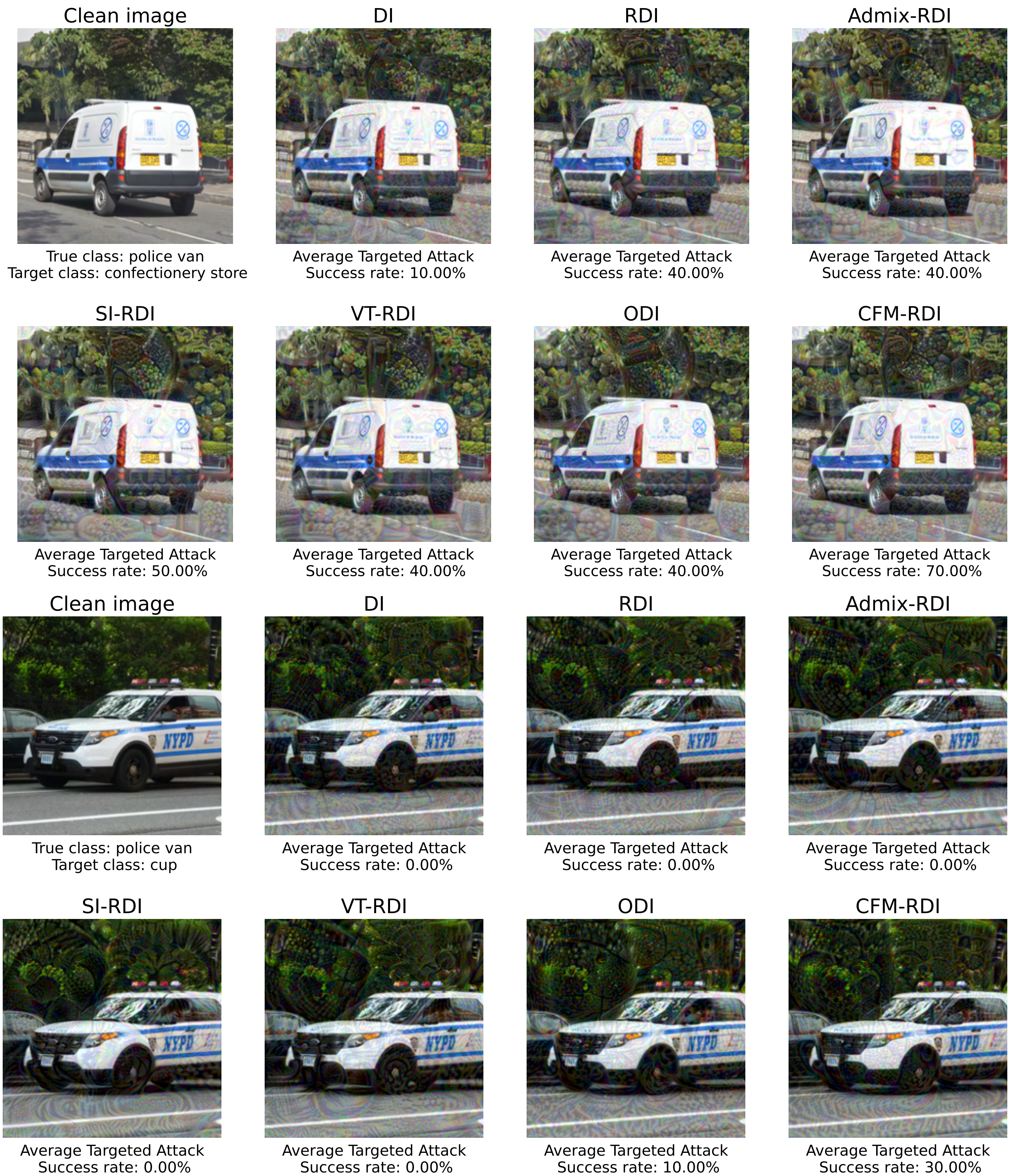

Figure 6. Visualization of generated adversarial examples. The source model is Inc-v3. Each average targeted attack success rate was calculated over the ten carefully selected target models, which are more difficult to confuse. For example, an average targeted attack success rate of 50% means that 5 out of 10 target models recognize the adversarial example as the target class.

impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 1

[17] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 1

[18] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 1

[19] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16158–16167, 2021. 2

[20] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 1

[21] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020. 1

[22] Huanrui Yang, Jingyang Zhang, Hongliang Dong, Nathan Inkawhich, Andrew Gardner, Andrew Touchet, Wesley Wilkes, Heath Berry, and Hai Li. Dverge: diversifying vulnerabilities for enhanced robust generation of ensembles. *Advances in Neural Information Processing Systems*, 33:5505–5515, 2020. 1

Source : RN-50	Target model										
Attack	VGG-16	RN-18	RN-50	DN-121	Xcep	MB-v2	EF-B0	IR-v2	Inc-v3	Inc-v4	Avg.
DI	62.5	56.6	**98.9**	72.3	5.7	28.2	29.3	4.5	9.2	9.9	37.7
RDI	65.4	71.8	98.0	81.3	13.1	46.6	46.6	16.8	30.7	23.9	49.4
Admix$_{m_1=1}$-RDI	74.2	80.7	98.7	86.8	20.9	59.4	56.1	26.7	42.7	34.1	58.0
Admix$_{m_1=5}$-RDI	75.2	83.0	98.4	89.6	36.5	64.7	66.4	44.7	62.5	50.5	67.2
SI-RDI	70.5	79.8	98.8	88.9	29.5	56.2	66.2	37.9	56.4	43.6	62.8
VT-RDI	68.8	78.7	98.2	82.5	27.9	54.5	56.1	32.8	45.8	37.9	58.3
ODI	78.3	77.1	97.6	87.0	43.8	67.3	70.0	49.5	65.9	55.4	69.2
CFM-RDI	84.7	88.4	98.4	90.3	51.1	81.5	78.8	48.0	65.5	59.3	74.6
SI-CFM-RDI	**85.9**	**88.5**	98.4	**92.3**	**62.5**	**81.6**	**82.7**	**61.5**	**74.5**	**69.7**	**79.8**
Source : adv-RN-50	Target model										
Attack	VGG-16	RN-18	RN-50	DN-121	Xcep	MB-v2	EF-B0	IR-v2	Inc-v3	Inc-v4	Avg.
DI	65.3	81.5	**91.5**	87.0	32.6	62.5	68.8	36.9	55.3	42.2	62.4
RDI	59.7	83.5	90.7	85.9	39.7	67.0	68.8	44.2	62.4	45.1	64.7
Admix$_{m_1=1}$-RDI	62.7	83.0	90.3	86.6	46.9	71.8	72.4	48.8	66.3	53.0	68.2
Admix$_{m_1=5}$-RDI	54.4	81.0	86.0	81.8	48.8	68.0	68.5	50.7	68.3	52.9	66.0
SI-RDI	53.9	79.4	87.1	83.8	46.6	66.5	69.5	52.0	69.1	52.2	66.0
VT-RDI	54.0	76.8	84.7	81.2	38.5	60.3	58.7	42.7	56.1	44.9	59.8
ODI	62.0	77.6	84.3	85.0	56.3	66.9	73.0	61.1	71.9	60.0	69.8
CFM-RDI	**76.7**	**86.3**	90.9	**87.6**	**67.1**	**82.4**	**83.4**	**64.7**	**77.1**	**67.4**	**78.4**
SI-CFM-RDI	70.0	82.3	86.8	85.7	63.4	79.2	79.4	61.8	76.2	63.9	74.9
Source : Inc-v3	Target model										
Attack	VGG-16	RN-18	RN-50	DN-121	Xcep	MB-v2	EF-B0	IR-v2	Inc-v3	Inc-v4	Avg.
DI	2.9	2.4	3.4	5.0	1.9	1.8	3.7	3.0	**99.2**	4.2	12.8
RDI	3.5	3.8	4.0	7.0	3.1	3.0	5.9	6.3	98.7	7.1	14.2
Admix$_{m_1=1}$-RDI	6.3	6.5	8.8	12.8	6.0	6.1	10.9	12.2	98.7	13.6	18.2
Admix$_{m_1=5}$-RDI	4.4	9.0	8.3	13.3	8.2	6.5	12.0	14.8	98.5	16.3	19.1
SI-RDI	4.0	5.2	5.7	11.0	6.3	4.6	8.2	11.6	98.8	12.1	16.8
VT-RDI	5.9	8.9	9.4	13.2	7.4	5.9	9.8	12.3	98.7	14.7	18.6
ODI	14.3	14.9	16.7	32.3	20.3	13.7	25.3	26.4	95.6	31.6	29.1
CFM-RDI	22.9	26.8	26.2	39.1	34.1	27.1	38.6	36.2	95.9	44.8	39.2
SI-CFM-RDI	**24.4**	**36.3**	**32.3**	**51.1**	**44.8**	**30.9**	**45.7**	**52.0**	97.5	**55.4**	**47.0**
Source : DN-121	Target model										
Attack	VGG-16	RN-18	RN-50	DN-121	Xcep	MB-v2	EF-B0	IR-v2	Inc-v3	Inc-v4	Avg.
DI	37.4	28.7	44.4	**98.7**	5.2	13.1	18.7	4.3	7.1	8.3	26.6
RDI	42.1	48.8	55.7	98.5	10.1	21.0	29.0	12.8	20.8	18.8	35.8
Admix$_{m_1=1}$-RDI	53.2	60.7	67.6	98.3	17.8	31.5	39.4	20.1	31.1	26.5	44.6
Admix$_{m_1=5}$-RDI	49.6	60.4	65.3	98.6	21.6	34.8	43.5	28.9	41.0	34.3	47.8
SI-RDI	45.4	53.0	60.1	98.6	16.1	27.8	37.3	22.0	34.3	25.8	42.0
VT-RDI	47.7	56.7	62.1	98.6	20.3	28.7	36.9	25.4	31.5	27.2	43.5
ODI	64.2	64.2	71.7	98.0	31.4	45.9	56.1	39.8	52.8	45.9	57.0
CFM-RDI	76.2	79.0	83.9	97.8	41.1	62.5	68.6	43.6	56.1	53.8	66.3
SI-CFM-RDI	**77.2**	**81.2**	**85.4**	97.8	**49.7**	**67.8**	**74.8**	**53.8**	**67.9**	**59.7**	**71.5**

Table 1. Extended experimental results on targeted attack success rates (%) against the ten target models on the ImageNet-Compatible dataset.

Source : RN-50	Target model							
Attack	adv-RN-50	ViT	LeViT	ConViT	Twins	PiT	Avg.	Computation time per image (sec)
DI	10.9	0.1	3.6	0.3	1.3	1.5	2.9	3.73
RDI	34.8	0.7	13.1	1.9	5.9	6.8	10.5	3.29
Admix$_{m_1=1}$-RDI	52.4	1.3	22.5	2.5	8.5	8.4	15.9	9.73
Admix$_{m_1=5}$-RDI	68.6	4.0	36.3	7.7	18.7	20.0	25.9	49.19
SI-RDI	59.9	2.9	29.4	6.3	15.5	17.9	22.0	16.16
VT-RDI	64.2	2.9	28.1	5.2	15.0	14.0	21.6	19.83
ODI	64.7	5.1	37.0	10.7	20.1	29.1	27.8	9.05
CFM-RDI	75.5	4.3	46.1	8.9	25.2	24.7	30.8	3.72
SI-CFM-RDI	**80.8**	**12.4**	**60.1**	**16.7**	**39.7**	**43.3**	**42.2**	18.34
Source : adv-RN-50	Target model							
Attack	adv-RN-50	ViT	LeViT	ConViT	Twins	PiT	Avg.	Computation time per image (sec)
DI	**98.9**	5.7	36.9	10.1	19.2	20.5	31.9	3.77
RDI	98.8	10.8	49.5	19.9	29.4	35.8	40.7	3.29
Admix$_{m_1=1}$-RDI	**98.9**	12.1	55.5	23.1	32.4	38.9	43.5	9.86
Admix$_{m_1=5}$-RDI	98.4	19.7	56.4	34.1	36.2	49.4	49.0	49.19
SI-RDI	98.7	19.4	57.6	35.3	35.2	52.1	49.7	16.34
VT-RDI	98.5	10.6	46.3	20.0	27.1	34.4	39.5	19.83
ODI	97.3	22.2	57.7	38.8	40.0	54.9	51.8	9.04
CFM-RDI	98.3	29.5	**69.8**	41.8	**52.7**	59.8	58.6	3.74
SI-CFM-RDI	98.2	**33.1**	68.9	**46.6**	52.2	**61.9**	**60.1**	18.47
Source : Inc-v3	Target model							
Attack	adv-RN-50	ViT	LeViT	ConViT	Twins	PiT	Avg.	Computation time per image (sec)
DI	0.2	0.1	0.3	0.0	0.0	0.1	0.1	2.84
RDI	0.8	0.2	1.8	0.2	0.4	0.7	0.7	2.47
Admix$_{m_1=1}$-RDI	2.0	0.1	4.1	0.6	1.4	1.4	1.6	7.27
Admix$_{m_1=5}$-RDI	5.0	0.8	6.4	1.6	1.6	3.9	3.2	36.30
SI-RDI	2.0	0.3	4.1	0.9	0.7	3.2	1.9	12.23
VT-RDI	3.2	0.4	5.2	0.8	1.6	1.8	2.2	14.74
ODI	6.5	0.8	12.4	1.7	3.5	6.7	5.3	6.74
CFM-RDI	8.6	2.1	21.9	3.2	6.1	11.6	8.9	2.96
SI-CFM-RDI	**19.3**	**6.1**	**33.7**	**6.8**	**12.4**	**22.5**	**16.8**	14.63
Source : DN-121	Target model							
Attack	adv-RN-50	ViT	LeViT	ConViT	Twins	PiT	Avg.	Computation time per image (sec)
DI	3.2	0.2	3.0	0.4	1.0	1.1	1.5	3.62
RDI	10.1	0.8	8.5	1.3	3.7	4.5	4.8	3.22
Admix-RDI	19.2	1.0	14.7	1.7	6.8	7.4	8.5	9.46
SI-Admix-RDI	26.7	2.4	21.8	3.4	10.5	14.2	13.2	46.61
SI-RDI	19.2	2.0	16.1	2.4	8.2	11.7	9.9	15.65
VT-RDI	26.6	2.2	19.2	3.5	8.3	11.7	11.9	18.87
ODI	35.6	3.3	26.9	7.4	14.7	21.9	18.3	9.06
CFM-RDI	43.2	3.6	32.8	6.4	17.3	21.1	20.7	3.69
SI-CFM-RDI	**54.3**	**8.0**	**46.5**	**11.8**	**28.4**	**35.5**	**30.8**	18.18

Table 2. Extended experimental results of targeted attack success rates (%) against one adversarially trained model and five Transformer-based classifiers with the ImageNet-Compatible dataset. We also report the average computation time to construct an adversarial example.

Source : RN-50	Target model										
						ens3-RN-20					
Attack	VGG-16	RN-18	MB-v2	Inc-v3	DN-121	Baseline	ADP	GAL	DVERGE	Avg.	Computation time per image (sec)
DI	66.4	71.5	62.7	71.1	84.2	77.9	56.5	14.3	15.6	57.8	0.64
RDI	66.4	70.9	64.1	73.4	82.8	76.3	55.8	13.5	14.9	57.6	0.59
SI-RDI	72.9	76.3	77.1	77.0	84.7	81.2	65.5	20.0	22.4	64.1	3.17
VT-RDI	89.8	87.1	92.6	92.9	93.7	94.4	82.3	24.3	31.3	76.5	3.82
Admix$_{m_1=1}$-RDI	74.2	78.8	76.2	82.7	89.2	85.2	66.4	17.3	18.4	65.4	1.98
Admix$_{m_1=5}$-RDI	79.9	82.3	81.3	83.4	90.0	86.0	69.7	22.8	25.6	69.0	9.06
CFM-RDI	98.3	97.7	99.0	**99.0**	**99.2**	**98.8**	97.2	54.9	59.3	89.3	0.72
SI-CFM-RDI	**98.5**	**98.1**	**99.2**	98.9	**99.2**	**98.8**	**97.3**	**61.3**	**65.8**	**90.8**	5.02

Source : Inc-v3	Target model										
						ens3-RN-20					
Attack	VGG-16	RN-18	MB-v2	Inc-v3	DN-121	Baseline	ADP	GAL	DVERGE	Avg.	Computation time per image (sec)
DI	22.8	12.8	32.1	78.7	14.7	32.8	21.7	3.8	3.5	24.8	1.74
RDI	21.3	14.7	33.9	86.7	16.6	37.6	22.3	4.4	5.0	26.9	1.88
SI-RDI	43.1	28.5	51.4	**99.8**	30.0	60.9	46.5	11.9	9.1	42.4	8.54
VT-RDI	53.7	31.1	72.0	93.3	38.5	69.6	55.7	8.9	9.4	48.0	10.43
Admix$_{m_1=1}$-RDI	29.6	16.9	43.3	90.5	19.9	47.4	30.9	5.3	5.1	32.1	5.60
Admix$_{m_1=5}$-RDI	47.2	32.0	58.1	99.7	34.8	64.0	50.8	14.9	11.2	45.9	30.94
CFM-RDI	45.3	29.1	57.2	94.7	33.9	55.7	42.2	8.8	8.3	41.7	2.02
SI-CFM-RDI	**62.3**	**45.5**	**67.5**	99.5	**49.9**	**71.6**	**60.9**	**19.1**	**15.5**	**54.6**	8.84

Source : DN-121	Target model										
						ens3-RN-20					
Attack	VGG-16	RN-18	MB-v2	Inc-v3	DN-121	Baseline	ADP	GAL	DVERGE	Avg.	Computation time per image (sec)
DI	44.3	46.5	39.0	45.6	92.8	41.0	30.9	9.3	9.1	39.8	0.95
RDI	43.3	44.8	37.6	46.0	92.9	38.7	28.1	9.4	9.9	39.0	1.01
SI-RDI	51.3	47.4	48.1	52.6	98.3	46.4	37.5	11.0	11.1	44.9	6.33
VT-RDI	67.9	62.9	67.1	69.2	91.3	61.3	52.6	13.9	16.8	55.9	7.45
Admix$_{m_1=1}$-RDI	50.7	53.9	45.7	52.8	93.1	48.4	37.0	9.9	10.4	44.7	2.86
Admix$_{m_1=5}$-RDI	62.6	58.5	57.9	62.8	98.3	56.4	44.1	14.0	13.8	52.0	14.70
CFM-RDI	97.0	**96.5**	95.9	97.6	**100.0**	95.8	91.9	43.4	45.1	84.8	1.28
SI-CFM-RDI	**97.3**	96.2	**97.1**	**97.7**	99.6	**96.2**	**92.6**	**49.1**	**52.0**	**86.4**	7.44

Table 3. Targeted attack success rates (%) against nine target models, including four ensemble-based defensive models on the CIFAR-10 dataset. We also evaluated the average computation time for crafting an adversarial example.

Source : RN-50	Target model											
Attack	Xcep	MB-v2	EF-B0	IR-v2	Inc-v4	ViT	LeViT	ConViT	Twins	PiT	Avg.	Comput. time per image (sec)
None (-MI-TI)	0.6	2.9	1.6	0.1	0.5	0.0	0.0	0.0	0.0	0.0	0.6	3.27
RDI	13.1	46.6	46.6	16.8	23.9	0.7	13.1	1.9	5.9	6.8	17.5	3.29
SI-RDI	29.5	56.2	66.2	37.9	43.6	2.9	29.4	6.3	15.5	17.9	30.5	16.16
VT-RDI	27.9	54.5	56.1	32.8	37.9	2.9	28.1	5.2	15.0	14.0	27.4	19.83
ODI	43.8	67.3	70.0	49.5	55.4	5.1	37.0	10.7	20.1	29.1	38.8	9.05
ODI-RDI	45.8	65.8	69.0	48.2	51.4	6.2	41.9	11.8	22.8	31.9	39.5	9.77
Admix$_{m_1=1}$-RDI	20.9	59.4	56.1	26.7	34.1	1.3	22.5	2.5	8.5	8.4	24.0	9.73
Admix$_{m_1=5}$-RDI	36.5	64.7	66.4	44.7	50.5	4.0	36.3	7.7	18.7	20.0	35.0	49.19
VT-Admix$_{m_1=1}$-RDI	33.5	61.2	58.9	37.5	43.0	4.9	35.0	6.1	16.4	17.9	31.4	58.08
CFM	6.3	35.2	31.9	4.9	9.4	0.0	3.1	0.2	0.8	1.2	9.3	3.35
CFM-RDI	51.1	81.5	78.8	48.0	59.3	4.3	46.1	8.9	25.2	24.7	42.8	3.72
CFM-ODI	55.1	72.5	73.4	55.4	60.7	8.6	48.7	**16.7**	30.1	39.0	46.0	9.13
SI-CFM-RDI	62.5	81.6	**82.7**	61.5	69.7	12.4	60.1	**16.7**	39.7	43.3	53.0	18.34
VT-CFM-RDI	57.3	77.4	74.6	55.2	62.0	11.2	53.0	15.7	33.6	36.3	47.6	20.69
Admix$_{m_1=1}$-CFM-RDI	56.6	**84.0**	81.7	51.1	64.8	6.3	52.3	10.7	28.1	29.5	46.5	9.99
Admix$_{m_1=5}$-CFM-RDI	**65.9**	81.6	82.6	**61.8**	**69.9**	**12.5**	**60.4**	**16.7**	**40.0**	**42.4**	**53.4**	52.57
Source : adv-RN-50	Target model											
Attack	Xcep	MB-v2	EF-B0	IR-v2	Inc-v4	ViT	LeViT	ConViT	Twins	PiT	Avg.	Comput. time per image (sec)
None (-MI-TI)	7.7	18.6	23.8	8.2	6.8	0.6	7.8	1.4	3.6	3.9	8.2	3.27
RDI	39.7	67.0	68.8	44.2	45.1	10.8	49.5	19.9	29.4	35.8	41.0	3.29
SI-RDI	46.6	66.5	69.5	52.0	52.2	19.4	57.6	35.3	35.2	52.1	48.6	16.34
VT-RDI	38.5	60.3	58.7	42.7	44.9	10.6	46.3	20.0	27.1	34.4	38.4	19.83
ODI	56.3	66.9	73.0	61.1	60.0	22.2	57.7	38.8	40.0	54.9	53.1	9.04
ODI-RDI	52.8	65.6	68.8	57.1	56.8	25.1	57.3	39.5	38.0	53.5	51.5	9.96
Admix$_{m_1=1}$-RDI	46.9	71.8	72.4	48.8	53.0	12.1	55.5	23.1	32.4	38.9	45.5	9.86
Admix$_{m_1=5}$-RDI	48.8	68.0	68.5	50.7	52.9	19.7	56.4	34.1	36.2	49.4	48.5	49.19
VT-Admix$_{m_1=1}$-RDI	42.8	63.4	59.2	45.1	45.4	13.6	46.7	21.9	28.2	38.1	40.4	58.08
CFM	54.3	80.3	80.5	50.5	57.9	11.3	51.7	18.5	32.4	33.6	47.1	3.39
CFM-RDI	**67.1**	**82.4**	**83.4**	**64.7**	**67.4**	29.5	**69.8**	41.8	**52.7**	59.8	**61.9**	3.74
CFM-ODI	55.4	68.6	69.9	56.2	57.3	27.6	57.7	41.7	39.7	55.8	53.0	9.12
SI-CFM-RDI	63.4	79.2	79.4	61.8	63.9	**33.1**	68.9	**46.6**	52.2	**61.9**	61.0	18.47
VT-CFM-RDI	58.4	75.7	73.2	58.1	59.4	25.4	61.3	40.0	45.2	53.7	55.0	20.71
Admix$_{m_1=1}$-CFM-RDI	63.6	81.8	81.6	62.1	64.2	29.0	67.1	39.9	49.4	57.7	59.6	9.97
Admix$_{m_1=5}$-CFM-RDI	59.0	75.9	75.5	58.6	58.0	30.3	64.9	43.9	45.1	57.6	56.9	53.05

Table 4. Targeted attack success rates (%) of the combined attacks with multiple techniques against the ten selected target models, which are more difficult to be disturbed. The experiment was conducted on the ImageNet-Compatible dataset.

Ablation		Target model										
p	α_{max}	Xcep	MB-v2	EF-B0	IR-v2	Inc-v4	ViT	LeViT	ConViT	Twins	PiT	Avg.
0.05	0.5	55.2	78.7	79.2	54.4	60.9	15.7	62.8	29.1	41.0	47.9	52.5
0.05	0.75	59.8	82.1	82.3	61.3	65.3	20.5	67.5	33.8	46.2	53.2	57.2
0.05	1.0	63.9	**83.3**	**83.9**	63.0	**68.0**	24.8	**69.8**	40.1	50.6	56.6	60.4
0.1	0.5	61.3	81.7	80.8	62.5	64.1	22.8	69.0	37.4	46.2	54.1	58.0
0.1	0.75	**67.1**	82.4	83.4	**64.7**	67.4	**29.5**	**69.8**	**41.8**	**52.7**	**59.8**	**61.9**
0.1	1.0	64.9	81.5	81.0	61.6	66.7	28.1	66.6	41.5	49.6	**59.8**	60.1
0.15	0.5	64.2	82.6	81.9	63.6	67.7	27.0	68.7	41.0	50.8	58.0	60.5
0.15	0.75	62.6	80.4	80.2	61.1	64.0	28.7	65.2	40.9	49.5	56.7	58.9
0.15	1.0	53.2	73.6	72.5	49.7	52.2	22.0	57.2	34.9	39.0	48.6	50.3

Table 5. Targeted attack success rates (%) of CFM-RDI with different mixing probability p and upper bound of mixing ratios α_{max}. The source model is adv-RN-50.