Frequency-driven Imperceptible Adversarial Attack on Semantic Similarity Supplementary Materials

Cheng Luo* Qinliang Lin* Weicheng Xie† Bizhu Wu Jinheng Xie Linlin Shen
[1]Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University
[2]Shenzhen Institute of Artificial Intelligence & Robotics for Society
[3]Guangdong Key Laboratory of Intelligent Information Processing
{luocheng2020, linqinliang2021}@email.szu.edu.cn, {wcxie, llshen}@szu.edu.cn

A. Methodology for Targeted Attack

As mentioned above, we define the attack in the targeted scenario as:

$$\boldsymbol{x}_i^{adv} = \arg\min_{\boldsymbol{x}_i'} [s'_{i,i} - s'_{i,t}]_+, \tag{4}$$

where t denotes the index of the target image in a minibatch, $s'_{i,i} = \text{sim}(f(\boldsymbol{x}_i'), f(\boldsymbol{x}_i))$ and $s_{i,t} = \text{sim}(f(\boldsymbol{x}_i'), f(\boldsymbol{x}_t))$ are similarity scores. It aims to encourage the adversarial example $\boldsymbol{x}_i'$ to be close to the target $\boldsymbol{x}_t$ in terms of the feature representation.

The self-weighting scheme that is specific to this targeted attack is defined as follows:

$$\begin{aligned} \boldsymbol{x}_i^{adv} &= \arg\min_{\boldsymbol{x}_i'} \mathcal{L}_{SSA}(\boldsymbol{x}_i, \boldsymbol{x}_i') \\ &= \arg\min_{\boldsymbol{x}_i'} [\alpha_i s'_{i,i} - \beta_i s'_{i,t}]_+. \end{aligned} \tag{11}$$

As for the weighting factors α_i and β_i, we set them as:

$$\begin{cases} \alpha_i = [s'_{i,i} - m]_+, \\ \beta_i = [1 + m - s'_{i,t}]_+. \end{cases} \tag{12}$$

The targeted attack follows a similar attack design and weight factor setting as the untargeted attack. The only difference is that SSA in the targeted scenario does not need the selection of the most dissimilar example in the minibatch but requires a certain example of the target category.

B. Implementation details

For compared attack approaches, the parameters κ and c in C&W [1] are set to 40 and 0.1, respectively; the parameters α_l and α_c in PerC-AL [8] are initialized to 1 and 0.5 and gradually reduced to 0.01 and 0.05, respectively, with cosine annealing. For evaluating the robustness of attacks against defense approaches or the attack success rates of attacks against online models, the ℓ_∞ bound of 8/255 is used for perturbation generation.

*Equal Contribution
†Corresponding Author

C. Additional Experimental Results

C.1. Targeted Attack in the White-box Setting

In this section, we investigate the attack performance of SSAH in the targeted scenario in Tab. 6. It shows that our SSAH remains effective at generating imperceptible perturbations for the targeted scenario.

C.2. Parameter Sensitivity Analyses

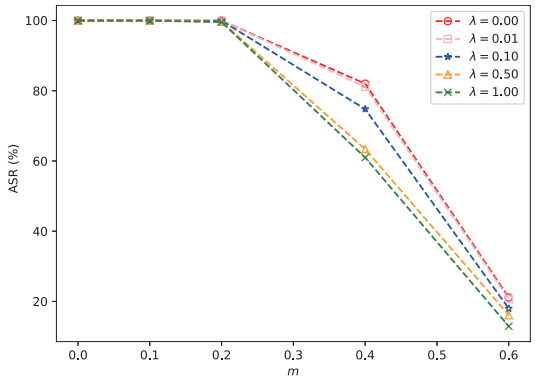

Figure 8. Sensitive analyses of λ and m in terms of attack success rate (ASR).

There are two hyperparameters in SSAH, *i.e.*, the margin m in Eq. (6) for adjusting self-paced weighting and λ in Eq. (10) for weighing the low-frequency constraint. The sensitivity analyses of these two hyperparameters are performed on CIFAR-10 and ImageNet-1K. The quantitative results on CIFAR-10 are presented in Figs. 8 and 9, and the visualization results on ImageNet-1K are presented in

Dataset	Attack	Iteration	RunTime (s) ↓	ASR (%) ↑	ℓ_2 ↓	ℓ_∞ ↓	FID ↓	LF ↓
CIFAR-10	BIM [5]	10	33	99.94	0.67	**0.03**	13.36	0.18
	PGD [6]	10	38	99.94	1.19	**0.03**	27.23	0.31
	MIM [2]	10	35	99.92	1.88	**0.03**	26.67	0.47
	AdvDrop [3]	150	389	99.11	1.13	0.08	16.52	0.42
	C&W ℓ_2 [1]	1000	978	100	**0.45**	0.07	10.34	0.13
	SSA (ours)	150	176	99.90	0.55	0.04	6.13	0.15
	SSAH (ours)	150	178	99.92	0.48	0.04	**5.14**	**0.07**
CIFAR-100	BIM [5]	10	32	99.41	**0.74**	**0.03**	13.59	0.27
	PGD [6]	10	36	99.34	1.22	**0.03**	25.64	0.39
	MIM [2]	10	32	99.11	1.84	**0.03**	25.49	0.64
	AdvDrop [3]	150	308	97.70	1.09	0.08	15.56	0.43
	C&W ℓ_2 [1]	1000	743	99.99	0.94	0.09	17.59	0.64
	SSA (ours)	150	134	99.15	1.27	0.08	10.57	0.52
	SSAH (ours)	150	138	99.01	0.99	0.07	**8.88**	**0.07**
ImageNet-1K	BIM [5]	10	2166	98.31	25.18	0.03	39.61	10.44
	PGD [6]	10	2973	98.65	53.84	0.03	37.21	16.87
	MIM [2]	10	2358	99.98	92.86	0.03	81.62	39.93
	AdvDrop [3]	150	46968	99.76	14.95	0.06	11.28	5.67
	C&W ℓ_2 [1]	1000	$>$ 100000	97.83	**1.85**	0.04	12.93	0.84
	SSA (ours)	200	31742	98.64	4.31	**0.02**	8.64	1.94
	SSAH (ours)	200	30050	98.06	3.38	**0.02**	**6.42**	**0.47**

Table 6. Results of the attack success rate (ASR) and four metrics related with perceptual similarity by seven attack approaches in the targeted scenario. The best results are marked in bold.

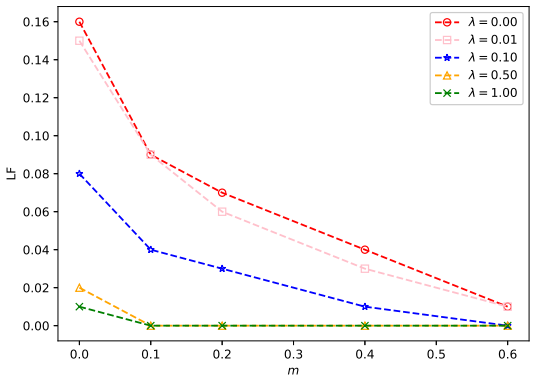

Figure 9. Sensitive analyses of λ and m in terms of LF.

Fig. 10.

Fig. 8 shows that SSAH becomes sensitive to λ when the margin m is large (*e.g.*, $m \geq 0.2$). Taking the case of $m = 0.4$ for example, along with increasing weighting λ, the performance of SSAH decreases fast from 82.02% when $\lambda = 0.00$ to 60.98% when $\lambda = 1.00$. The reason is that we actually impose a weak attack strength on images when m value is large. In this condition, a large λ results in a strong constraint on our attack objective in SSAH, leading to a low attack success rate.

Based on the above analysis, it can be concluded that an appropriate selection of m and λ is necessary. The sensitivity analyses in Fig. 8 and Fig. 9 show that SSAH can achieve relatively stable and satisfying performances when $m \in [0.0, 0.2]$ and $\lambda \in [0.1, 1.0]$.

C.3. Batch Size

Tab. 7 reports the results with batch size from 32 to 10000. Our attack works reasonably well over this wide range of batch sizes. The results are similarly good when the batch size is from 32 to 10000, and the differences are at the level of random variations.

Batch Size	ASR ↑	ℓ_2 ↓	ℓ_∞ ↓	LF ↓
32	99.94	**0.25**	**0.02**	**0.02**
64	99.90	**0.25**	**0.02**	**0.02**
128	99.91	**0.25**	**0.02**	0.03
256	**99.95**	**0.25**	**0.02**	0.03
512	99.94	**0.25**	**0.02**	0.03
1024	99.91	**0.25**	**0.02**	0.03
2048	99.89	0.26	**0.02**	0.03
4096	99.90	0.26	**0.02**	0.03
10000	99.94	0.26	**0.02**	0.03

Table 7. Effect of batch sizes (CIFAR-10 evaluation attack success rate and four metrics related with perceptual similarity).

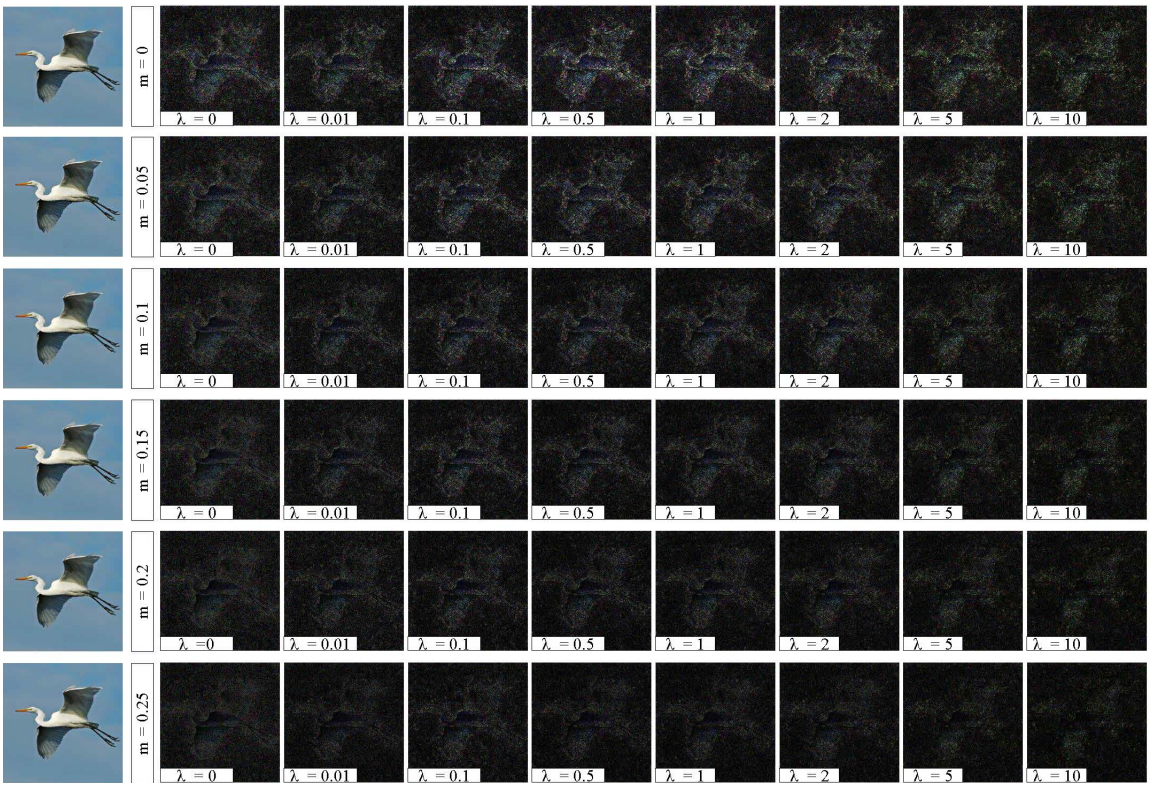

Figure 10. Adversarial perturbations to an ImageNet-1K image using different hyperparameters of m and λ in SSAH.

C.4. Additional Evaluation of Imperceptibility

In this section, we evaluate perturbation imperceptibility in terms of perceptual colour difference ($\overline{C_2}$). This metric is used in PerC-AL [8] to measure human colour perception.

Tab. 8 shows that our SSAH without the constrant of ℓ_2 ($\overline{C_2}$) distance still achieves competitive performances in terms of ℓ_2 ($\overline{C_2}$) . It further demonstrates the superiority of our attack in perturbation imperceptibility.

Attack	Iter.	ASR ↑	ℓ_2 ↓	LF ↓	$\overline{C_2}$ ↓
C&W	1000	99.27	**1.51**	0.67	152.51
PerC-AL	1000	98.78	4.35	1.59	**90.62**
SSAH (ours)	200	98.01	1.81	**0.06**	124.32

Table 8. Results of the attack success rate (ASR) and three metrics related with perceptual similarity by three attack approaches on ImageNet-1K in the untargeted scenario.

C.5. Additional Robustness Evaluations

To evaluate the performance of our attack against image transformation-based defense, we test the robustness of the adversarial examples against bit-depth reduction [4, 7]. Fig. 11 shows our imperceptible attack on high-frequency components is still robust against this image transformation-based defense (*i.e.*, bit-depth reduction).

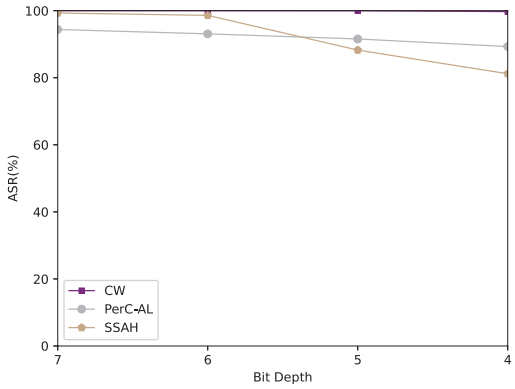

Figure 11. Evaluation of robustness against bit-depth reduction on CIFAR-10.

C.6. Analysis of SPW

To further measure the effect of applying the self-paced weighting in SSAH, we randomly sample images from the testing set of CIFAR-10, and present the distributions of perturbation intensities generated by SSAH with and without self-paced weighting in Fig. 12. This figure shows that the perturbation intensities generated by SSAH with SPW appear to be smaller than those by SSAH without SPW. It

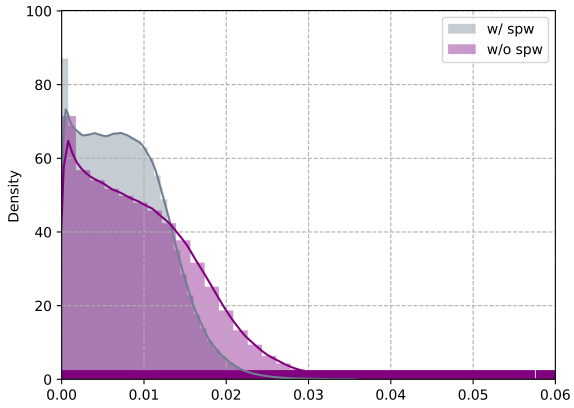

Figure 12. Distributions of perturbation intensities by SSAH with self-paced weighting (w/ SPW) and without self-paced weighting (w/o SPW) on a subset (1024 random samples) of CIFAR-10.

indicates that the weighting scheme can well reduce the redundant perturbations caused by over-optimization.

C.7. Additional Visualization Results

In this section, we present more visualization results, including adversarial examples and perturbations generated in the white-box setting (*i.e.*, Fig. 13, Fig. 14, Fig. 15 and Fig. 16) and transferable adversarial examples across architectures and datasets (*i.e.*, Fig. 17).

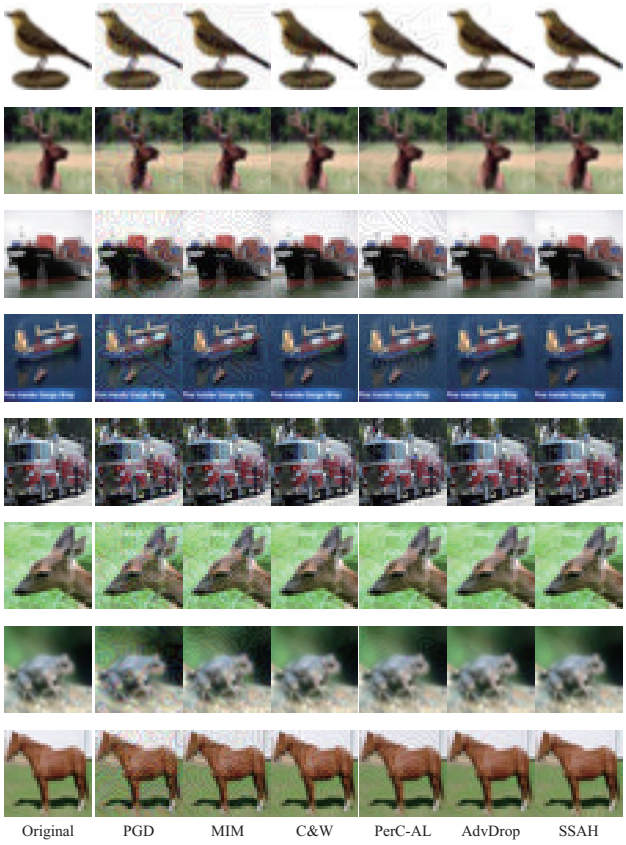

Figure 13. Adversarial examples generated by six different attack approaches on CIFAR-10.

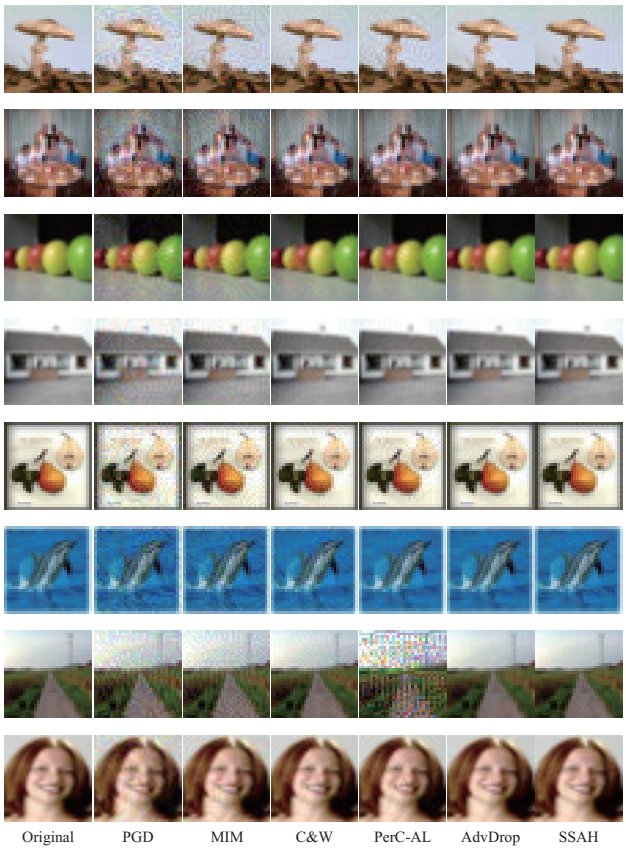

Figure 14. Adversarial examples generated by six different attack approaches on CIFAR-100.

Figure 15. Adversarial examples generated by three different attack approaches on ImageNet-1K.

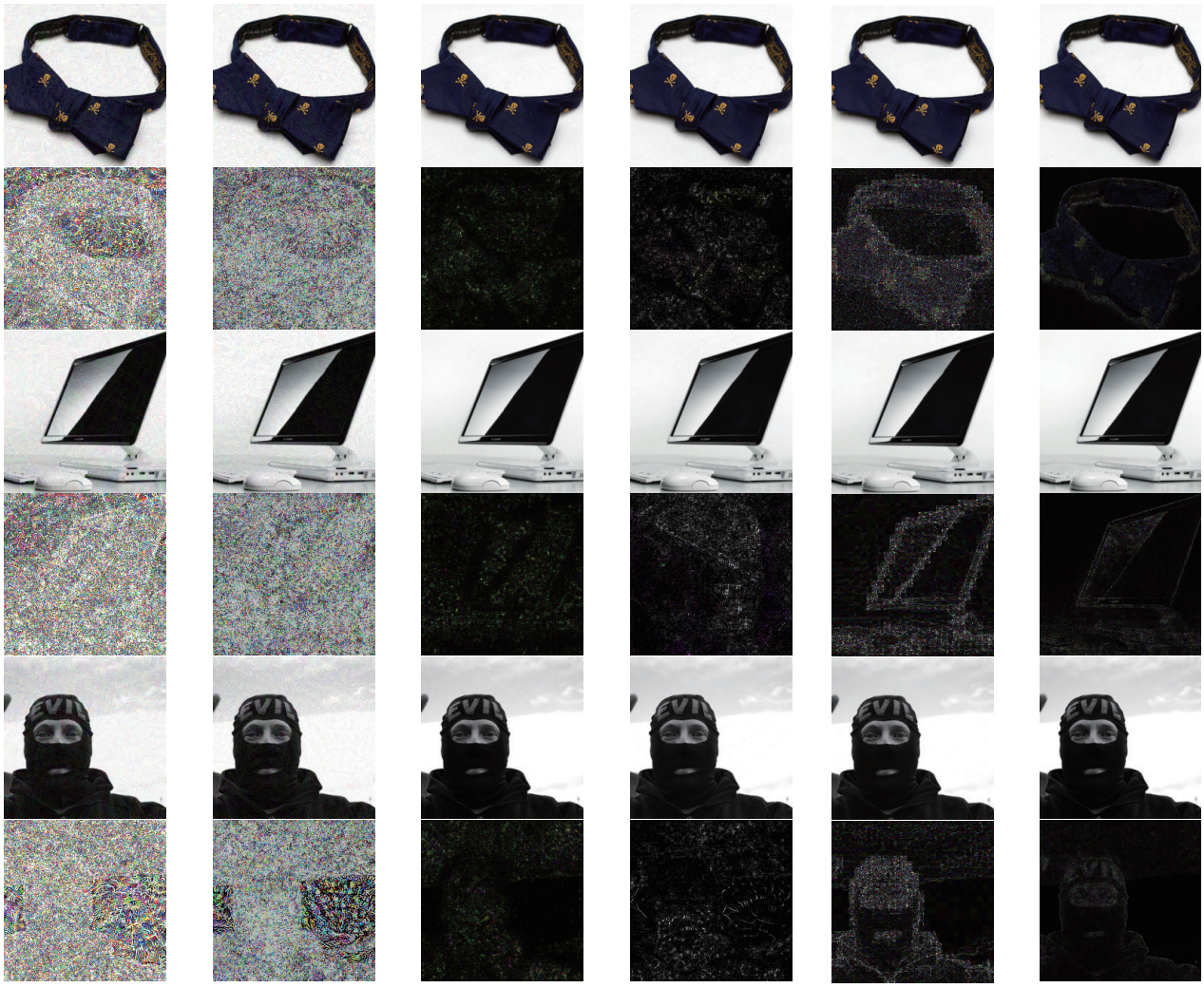

Figure 16. Adversarial examples and their corresponding perturbations generated by six different attack approaches on ImageNet-1K. The 1st, 3rd and 5th rows are adversarial examples, while the 2nd, 4th and 6th rows are their perturbations.

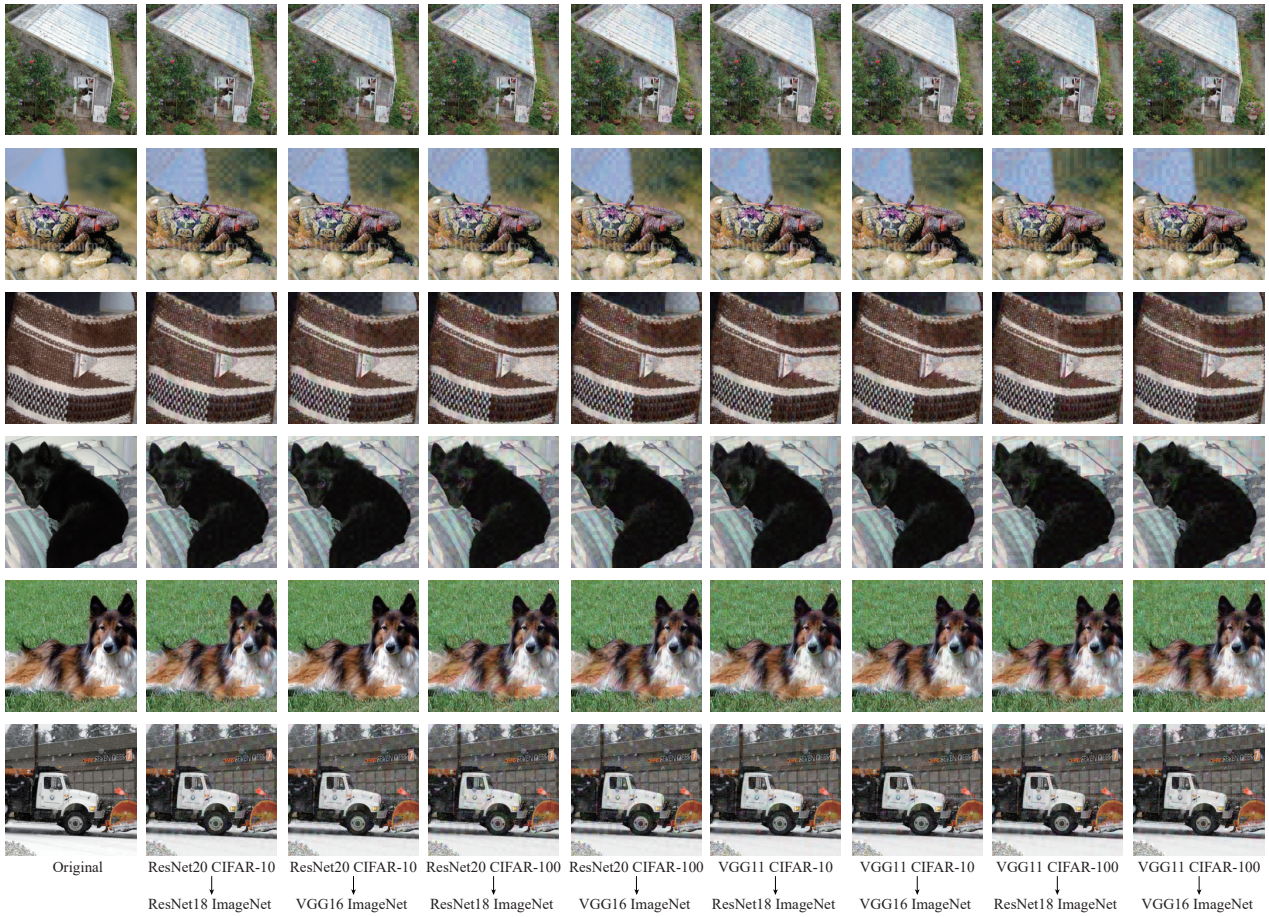

Figure 17. Adversarial examples generated by SSAH based on a surrogate model trained on the source domain to a target model trained on the target domain. The original examples are selected from the validation set of ImageNet-1K. A B $\rightarrow$ C D denotes that model A trained on dataset B is used to craft perturbations to fool model C trained on dataset D.

D. Image Samples in Transferable Attack

In the experiment of attacking online models, we randomly sample 200 images from the ImageNet-1K validation set. For reproducibility, we list the names of these used images as follows:

n02108089/val_00016416, n01917289/val_00047927, n03877845/val_00026991, n02125311/val_00024128, n03661043/val_00018698, n02085620/val_00031154, n03691459/val_00023671, n03888257/val_00025121, n02276258/val_00032773, n01704323/val_00038815, n02342885/val_00027868, n01632458/val_00031521, n03770439/val_00013597, n02231487/val_00009943, n03478589/val_00040225, n02111277/val_00046591, n02840245/val_00031888, n04398044/val_00041709, n07892512/val_00017146, n01694178/val_00025511, n03873416/val_00047900, n02114367/val_00033884, n02092002/val_00022857, n03976467/val_00033699, n02342885/val_00041792, n02457408/val_00041441, n03773504/val_00022433, n03930313/val_00002877, n09193705/val_00038734, n02804414/val_00008402, n03124170/val_00026251, n01806143/val_00023973, n01818515/val_00021663, n03376595/val_00040795, n02226429/val_00045770, n02655020/val_00008184, n02484975/val_00045387, n03478589/val_00000035, n02951585/val_00023091, n01692333/val_00033244, n02281406/val_00046852, n02389026/val_00014369, n04310018/val_00038429, n03956157/val_00038501, n04200800/val_00018851, n01968897/val_00030526, n01608432/val_00043085, n03444034/val_00026796, n02342885/val_00047769, n02814533/val_00005978, n01843383/val_00037244, n02422106/val_00035337, n02963159/val_00030024, n13052670/val_00039616, n01829413/val_00036044, n02823750/val_00041752, n03602883/val_00039677, n12985857/val_00038482, n04005630/val_00004526, n04487394/val_00033068, n03127747/val_00011868, n02701002/val_00028205, n03124170/val_00033776, n03355925/val_00006767, n03042490/val_00025402, n02787622/val_00019653, n03837869/val_00036456, n04523525/val_00019004, n04409515/val_00014503, n13052670/val_00033352, n02074367/val_00023213, n03075370/val_00021941, n02971356/val_00023953, n03126707/val_00043871, n02641379/val_00008847, n01440764/val_00017699, n03457902/val_00049086, n03180011/val_00043506, n01980166/val_00005836, n04392985/val_00003756, n07747607/val_00005888, n04417672/val_00046924, n02165105/val_00030936, n03290653/val_00015959, n03933933/val_00030150, n03709823/val_00001860, n02028035/val_00010037, n04118538/val_00045773, n02361337/val_00037480, n03529860/val_00022699, n04204347/val_00036559, n02606052/val_00014777, n02104365/val_00035049, n02494079/val_00002579, n01877812/val_00022693, n01687978/val_00029055, n02107683/val_00035628, n02804414/val_00047642, n03891332/val_00036752, n02113186/val_00009045, n02105162/val_00030963, n02172182/val_00045513, n01560419/val_00018672, n02134084/val_00016340, n02794156/val_00006960, n01484850/val_00016988, n03016953/val_00021529, n04404412/val_00049571, n03888257/val_00023107, n02909870/val_00036628, n02105641/val_00030831, n09468604/val_00028882, n02097658/val_00015119, n02172182/val_00031062, n02106030/val_00014342, n02493793/val_00013076, n03976467/val_00032721, n01924916/val_00037245, n01944390/val_00022381, n03954731/val_00005903, n02883205/val_00003045, n04004767/val_00008345, n01693334/val_00022937, n02443484/val_00018383, n04597913/val_00016472, n03670208/val_00010364, n02009912/val_00026861, n02326432/val_00002413, n04229816/val_00047467, n02971356/val_00000831, n04192698/val_00001173, n02791270/val_00044488, n03595614/val_00025275, n03126707/val_00038038, n04613696/val_00014904, n04235860/val_00006430, n01484850/val_00045730, n02835271/val_00033559, n02128385/val_00036527, n02395406/val_00033668, n02115913/val_00026888, n02930766/val_00049552, n02403003/val_00017987, n02655020/val_00005213, n04049303/val_00018090, n04553703/val_00048784, n03868863/val_00011676, n03769881/val_00001207, n04399382/val_00006985, n03394916/val_00003759, n07760859/val_00025973, n04069434/val_00001854, n02093647/val_00000883, n02690373/val_00030319, n01728920/val_00020274, n02132136/val_00006703, n04141327/val_00003236, n03743016/val_00038715, n02102480/val_00004185, n02974003/val_00021691, n03018349/val_00020100, n02074367/val_00030029, n04149813/val_00032339, n03594945/val_00037128, n04399382/val_00039275, n04372370/val_00003304, n03388549/val_00041363, n01984695/val_00002243, n02489166/val_00000206, n02037110/val_00011196, n04418357/val_00028155, n04251144/val_00005522, n02795169/val_00025956, n02113799/val_00016697, n04461696/val_00021814, n09468604/val_00041897, n02093991/val_00004687, n07802026/val_00040001, n03992509/val_00046414, n03662601/val_00023517, n02110958/val_00011495, n03467068/val_00020240, n02869837/val_00015695, n02127052/val_00036625, n11879895/val_00012320, n02747177/val_00002145, n02966687/val_00008803, n02704792/val_00030917, n07613480/val_00000620, n07930864/val_00039707, n04389033/val_00030415, n02128385/val_00024214, n02391049/val_00015124, n01616318/val_00006250, n04536866/val_00024173, n03467068/val_00049842, n03180011/val_00049615, n02099429/val_00040055, n03291819/val_00011445, n04447861/val_00038921.

References

[1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, pages 39–57, 2017. 1, 2

[2] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9193, 2018. 2

[3] Ranjie Duan, Yuefeng Chen, Dantong Niu, Yun Yang, AK Qin, and Yuan He. Advdrop: Adversarial attack to dnns by dropping information. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7506–7515, 2021. 2

[4] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[5] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR) Workshops*, 2017. 2

[6] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[7] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Annual Network and Distributed System Security Symposium (NDSS)*, 2018. 3

[8] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1039–1048, 2020. 1, 3