A Adversarial Examples

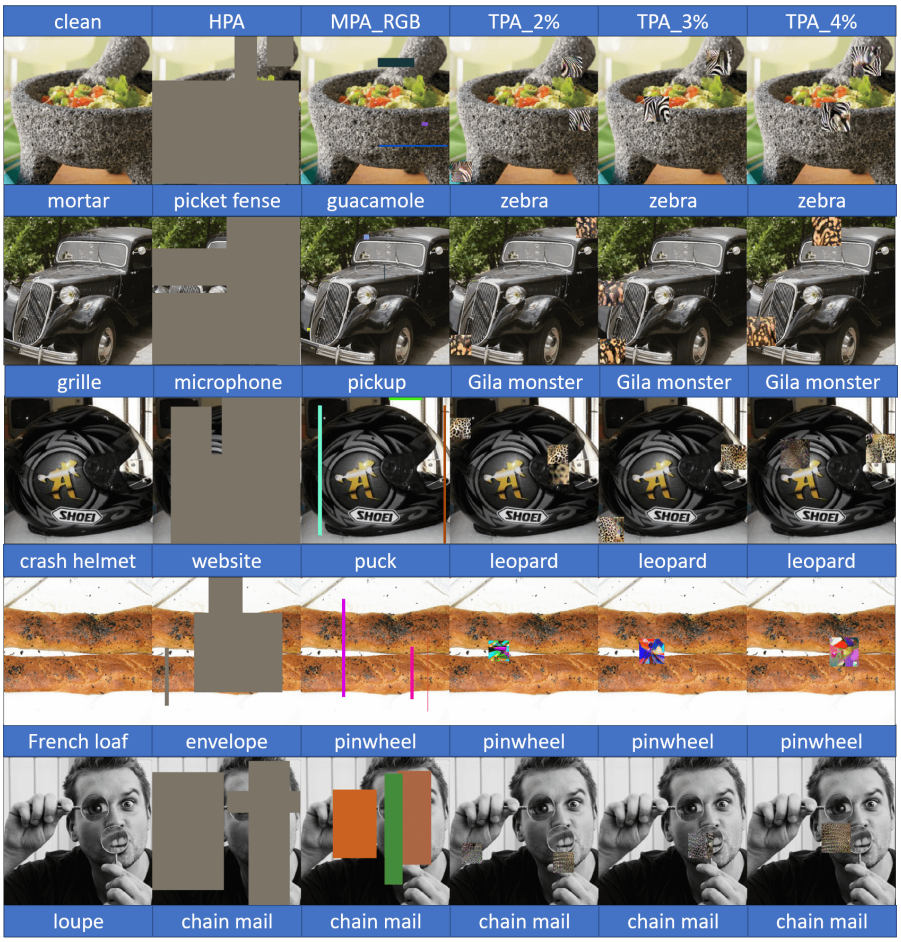

Fig. 5. Adversarial examples generated by targeted *PatchAttack* on ResNet50. The images in the same row are attacked with the same target class. The first three columns correspond to clean images, Hastings Patch Attack (HPA) and Monochrome Patch Attack (MPA), and the last three columns Texture-based Patch Attack (TPA) with the single patch area being 2%, 3% and 4%, respectively.

B Attention Maps of Adversarial Examples

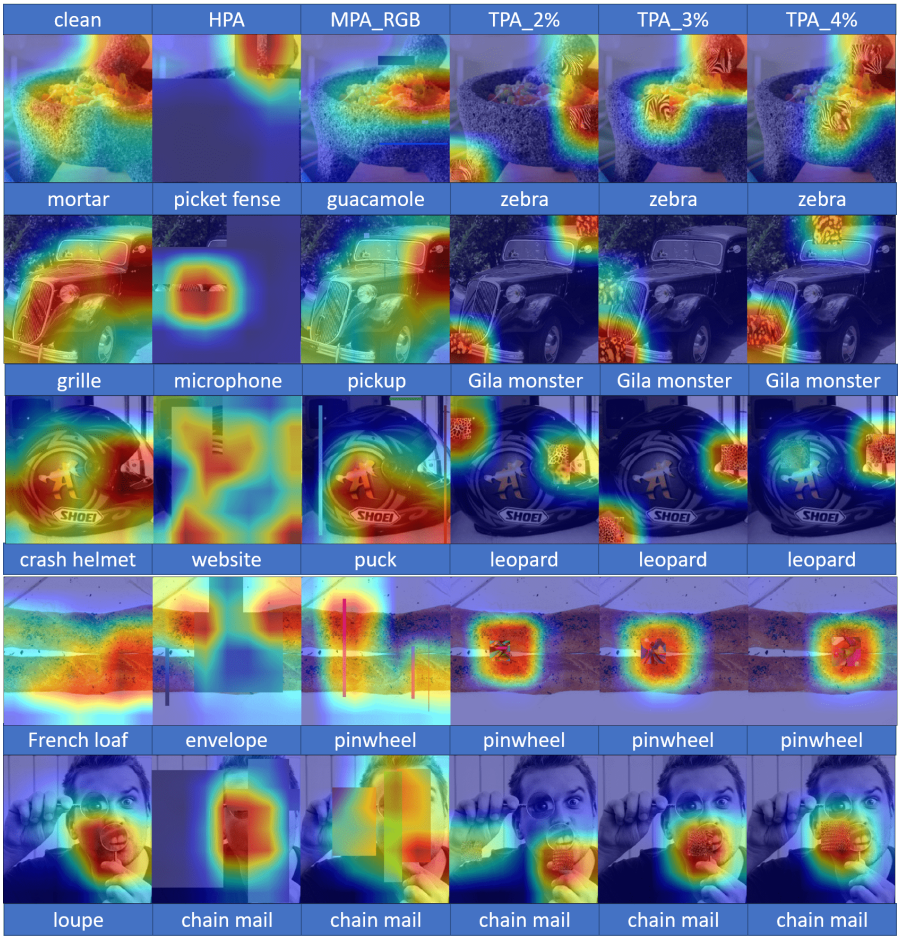

Fig. 6. Attention maps of the adversarial examples in Fig. 5 generated by Grad-CAM on ResNet50. The images in the same row are attacked with the same target class. The first three columns correspond to clean images, Hastings Patch Attack (HPA) and Monochrome Patch Attack (MPA), and the last three columns Texture-based Patch Attack (TPA) with the single patch area being 2%, 3% and 4%, respectively.

C Defense 3: against white-box patch attack defense

We evaluate our Texture-based Patch Attack (TPA) against Local Gradients Smoothing (LGS) [36] which is dedicated to defend against white-box patch attack on ImageNet. We perform the targeted attack on ResNet50 with the same setting in 4.2 and show the result in Table 5. While LGS leads to slightly higher patch area and slightly lower target accuracy, it clearly fails to defend against TPA.

Table 5. Experimental results on 1000 images randomly selected from the ILSVRC2012 validation set. T_acc. and Avg_area denote the classification accuracy on target labels and average area percentage occluded by the patches, respectively

Attack	Defense	T_acc.(%)	Avg_area(%)
TPA_N10_4%	–	99.70	9.97
TPA_N10_4%	LGS	97.50	13.25

D Comparison between Metropolis-Hastings sampling and Reinforcement Learning

We implement the Hastings Patch Attack (HPA) in the same RGB and texture search space used by Monochrome Patch Attack (MPA) and Texture-based Patch Attack (TPA) to compare this sampling method and Reinforcement Learning method (RL). The experiments are performed on ResNet50 with the standard setup in 4.2.

Table 6. Experimental results of the defenses on 1000 images randomly selected from the ILSVRC2012 validation set. The maximum allowed query number is 10000 and 50000 for the non-targeted and targeted settings. Acc., T_acc., Avg_area, and Avg_qry denote the classification accuracy on ground truth and target labels, average area percentage occluded by the patches, average query number, respectively

Non-targeted	Acc.(%)	Avg_area(%)	Avg_qry
HPA_RGB	0.20	16.88	10000
MPA_RGB	0.00	5.41	9681
targeted	T_acc.(%)	Avg_area(%)	Avg_qry
HPA_RGB	24.80	69.63	50000
MPA_RGB	25.90	18.45	28361

It is observed that MPA_RGB is better than HPA_RGB, because it achieves lower accuracy in the non-targeted setting and higher target accuracy in targeted setting, while also using a smaller area and less queries.

Table 7. Experimental results of the defenses on 1000 images randomly selected from the ILSVRC2012 validation set. The maximum allowed query number is 10000 and 50000 for the non-targeted and targeted settings. Acc., T_acc., Avg_area, and Avg_qry denote the classification accuracy on ground truth and target labels, average area percentage occluded by the patches, average query number, respectively

Non-targeted	Acc.(%)	Avg_area(%)	Avg_qry
HPA_N4_4%	1.10	5.42	3522.5
TPA_N4_4%	0.30	5.06	1137
targeted	T_acc.(%)	Avg_area(%)	Avg_qry
HPA_N10_4%	99.80	10.89	14345
TPA_N10_4%	99.70	9.97	8643

Here we can observe that RL still is much more query-efficient than the sampling algorithm, however, the methods are comparable in terms of accuracy and occlusion area. This can be attributed to our improved search space for performing the attacks, highlighting the importance of our texture dictionary.

E Transferability of adversarial patch dictionary generated by white-box method

We implement Adversarial Patch (AP) [8], the white-box patch attack. We first generate an adversarial patch dictionary (AdvPatchDict) consisting of 1000 classes by attacking VGG19 using AP on ImageNet dataset, and then attack the other 4 networks used in our experiments with those patches in the dictionary. The results are shown in the Table 8. In non-targeted settting, AdvPatchDict decreases accuracy to 0.20% on VGG19 but only to $56\% - 66\%$ on the other networks. In targeted setting, it increases target accuracy on VGG to 98.20% but basically fails to increase it for other networks. Clearly, AdvPatchDict generated by the white-box method overfits to the architecture used to generate them, highlighting the superiority of the design of our texture dictionary.

Table 8. Experimental results on 1000 images randomly selected from the ILSVRC2012 validation set. Acc., T_acc. and P_area, denote the classification accuracy on ground truth and target labels, area percentage occluded by the adversarial patch, respectively

AdvPatchDict	Non-targeted Acc.(%)	targeted T_acc.(%)	P_area(%)
VGG19	0.20	98.20	8.95
ResNet50	62.50	0.00	8.95
DenseNet121	57.80	1.70	8.95
ResNeXt50	65.30	0.10	8.95
MobileNet-V2	56.00	0.10	8.95