Supplementary Material for ICCV 2025 Paper #14577
IAP: Invisible Adversarial Patch Attack through Perceptibility-Aware Localization and Perturbation Optimization

Subrat Kishore Dutta Xiao Zhang
CISPA Helmholtz Center for Information Security
{subrat.dutta, xiao.zhang}@cispa.de

A. Detailed Experimental Settings

In this section, we give more details on the setup of our experiments. We evaluate the performance of our adversarial patch attack on image classification and face recognition tasks, with comparisons to state-of-the-art attack methods, such as Google Patch [1], LaVAN [8], GDPA [10], and Masked Projected Gradient Descent (MPGD), which is an extension of the standard PGD attack introduced in Madry et al. [13]. In addition, we evaluate the effectiveness of our attack against existing defense methods designed specifically against adversarial patch attacks [2–4, 7, 11, 18]. For GDPA, we balance attack efficacy and imperceptibility by setting the visibility parameter α to 0.4, while for MPGD, we set the l_∞ perturbation bound to $\epsilon = 16/255$.

Dataset and Model Setup. We consider a subset of the ILSVRC 2012 validation set [15] consisting of 1000 correctly classified images, one from each class, for image classification. For face recognition tasks, following Li and Ji [10], we use the test set of the VGG face dataset [10, 14], consisting of a total of 470 images across 10 classes. We consider four target network architectures: ResNet-50 [5], VGG16 [17], Swin Transformer Tiny, and Swin Transformer Base [12]. For image classification, we use their pre-trained weights. For face recognition, we re-train them on the VGG Face dataset's train set, which comprised $3,178$ images across 10 classes. The retraining procedure follows the same specifications as used by [10]. All the images in both tasks are resized to a dimension of 224×224 before being attacked.

Attack Configuration. In our experiments, we optimize the patch until the target class confidence reaches 0.9 or for a maximum of $1,000$ iterations. The patch size is fixed at 84×84, covering 14% of the image. While we use a square patch following prior works, our optimization framework can be generalized to other shapes. If the attack fails, we reinitialize the step size up to three times. All experiments are conducted on a single NVIDIA A100 GPU (80 GB), using PyTorch as the deep learning framework.

Attack Success Rate (ASR). We evaluate the effectiveness of different attack methods based on targeted attack success rate, denoted as ASR, which characterizes the ratio of instances that can be successfully attacked using the evaluated method. Let $\mathcal{A}$ be the evaluated attack, f_θ be the victim model, and $\mathcal{S}$ be a test set of correctly classified images. The ASR of $\mathcal{A}$ with respect to f_θ and $\mathcal{S}$ is defined as:

$$\mathrm{ASR}(\mathcal{A}; f_\theta, \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x} \in \mathcal{S}} \mathbb{1}\big(f_\theta(\hat{\boldsymbol{x}}) = y_{\mathrm{targ}}\big), \tag{1}$$

where $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$, and $\hat{\boldsymbol{x}}$ is the adversarial example generated by $\mathcal{A}$ for $\boldsymbol{x}$.

Imperceptibility. To measure patch imperceptibility, we use similarity matrices, incorporating both traditional statistical methods and convolutional neural network (CNN) based measures. The former measures involve Structural Similar Index Measure (SSIM) [20], Universal Image Quality index (UIQ) [19], and Signal to Reconstruction Error ratio (SRE) [9], while the latter involves CLIPScore [6], and Learned Perceptual Image Patch Similarity (LPIPS) metric [21]. SSIM measures structural similarity, while UIQ evaluates distortion based on correlation, luminance, and contrast, yielding a single index within $[-1, 1]$. SRE, akin to PSNR, measures error relative to the signal's power, ensuring consistency across different brightness levels. CLIPScore and LPIPS assess perceptual similarity using pre-trained DNNs, capturing subtle visual features. We evaluate similarity between adversarial and original samples on two scales: globally, by comparing entire images, and locally, by analyzing the similarity within the attacked region.

B. Other Experiments

B.1. Additional Results on ImageNet

In this section, we present additional detailed results corresponding to every victim model considered, with "Toaster" as the target class. The results include a comprehensive analysis of our attack's stealthiness, including all the imperceptibility metrics considered and mentioned earlier.

The evaluations across victim models, VGG16, ResNet-50, Swin Transformer Tiny, and Swin Transformer Base, are presented in Tables 3-6 respectively. The results account for the stability of IAP across architectures in terms of ASR, which is either on par with or exceeds the baseline methods considered. As evident from the analysis, we achieve state-of-the-art performance in imperceptibility, further demonstrating its stability. The adversarial samples created corresponding to each victim architecture are shown along with their target class confidence in Figures 3-6.

Cross-Class Attack Stability. To assess the effectiveness of our attack across multiple target classes, we extend our evaluation beyond the "Toaster" class to include "Baseball" and "Iron" as additional targets. We employ ResNet-50 as the victim model while maintaining all other attack configurations consistent with previous experiments. The results, summarized in Table 7, demonstrate that IAP achieves consistently high ASR across different target classes while preserving its imperceptibility, as illustrated in Figure 7. Notably, our approach exhibits stability across classes, achieving an ASR of 99.47 ± 0.13 and maintaining high imperceptibility, exemplified by a local SSIM of 0.94 ± 0.005.

B.2. Additional Results on VGG Face

Here, we present detailed results explicitly corresponding to every victim model corresponding to the three target classes considered. Tables 8-10, 11-13, 14-16, and 17-19 summarize the results for the three target classes using VGG16, ResNet-50, Swin Transformer Tiny, and Swin Transformer Base as victim models, respectively. The results show consistent attack performance across the criteria considered, as well as achieving state-of-the-art imperceptibility performance, which further demonstrates its efficiency. The adversarial samples corresponding to the target classes "A. J. Buckley", "Aamir Khan", and "Aaron Staton" are shown in Figures 8-10 respectively.

C. Further Analyses

C.1. Ablation Studies

We perform ablation studies to assess key components of IAP, including patch size, update iterations, and the regularization coefficient in the loss function (Equation 8). We compare our update rule with the Adam optimizer and test the assumption that adversarial patches attract the classifier's attention. All experiments use ImageNet with Swin Transformer Base as the victim model. Here, we detail the comprehensive evaluation corresponding to both attack efficacy and imperceptibility. Table 20 demonstrates that as the patch size increases, the imperceptibility improves. Figure 11 validates this as we see that the attack area becomes smoother with the increase in the size. Aligned with our hypothesis, the initial increase in w_3 improved the imperceptibility of the generated patches as presented in Table 21. We studied the impact of the update rule proposed by our method, IAP, by altering it with the update rule corresponding to the Adam optimizer. As shown in Table 22 and visualized in Figure 12, we achieve most of our imperceptibility because of the update rule we utilize for updating the perturbation. In addition, we also considered the effect of the number of optimization steps on the ASR and imperceptibility of the attack.

Effect of Patch Size. We evaluate the impact of patch size on attack efficacy and imperceptibility. We hypothesize that increasing the patch size would enhance attack performance and imperceptibility, as the perturbations would disperse over a larger area while remaining less salient. The results support this hypothesis, with a 99.4% attack success rate (ASR) for a patch covering 14% of the image, compared to 72.2% ASR for 2% coverage. For patch sizes of 4% or more, the ASR reached 90.7% or higher. These findings also show improved imperceptibility with larger patch sizes as highlighted in Table 16 and Figure 4.

Effect of Regularization Coefficient. We study the effect of the regularization coefficient w_3 in the human-oriented distance metric (Eq. 7), part of the total loss function (Eq. 8). We hypothesize that increasing w_3 would improve imperceptibility at the cost of slightly reducing attack performance. Our results support this hypothesis as shown in Table 17. As w_3 increases, the attack success rate slightly decreases while imperceptibility improves. However, beyond a certain point, the trend reverses due to the destabilizing effect of large w_3 values, which cause the loss function to be dominated by the regularization term, requiring more iterations for successful attacks and reducing imperceptibility.

Effect of Update Rule. We compare our proposed update rule, which allows for longer iterations with no perturbation magnitude constraints while maintaining imperceptibility, to the widely used Adam Optimizer update rule. We hypothesize that Adam, optimized for attack success, would yield a higher success rate. However, Adam's updates alter each color channel separately, potentially changing the pixel's base color, whereas our method preserves it. While Adam achieves a slightly higher attack success rate, IAP completely outperformed it in terms of imperceptibility as demonstrated in Table 18. Figure 6 visualizes and compares the adversarial patches generated by both approaches.

Effect of the number of Update Iterations. As the number of updates increases, the patch's appearance diverges from the original, even if the perturbations remain less salient. Despite the reduced saliency, more iterations typically improve attack success rates. In these experiments, we fix the patch size at 6% to evaluate the trade-off between attack efficacy and imperceptibility. The ASR increases as the number of update iterations increases, as shown by Table 23, with a slight reduction in imperceptibility as perturbations accumulate, as shown in Figure 13.

C.2. GradCAM analysis of Attention Overlap

To understand the change in the attention map induced by the adversarial samples generated by IAP, we analyze the shift in the highest attention location of the attention map generated in comparison to the one generated corresponding to the benign sample. We use GradCAM [16] to measure the attention maps. Analysis of the attention maps holds critical significance because of the defense implications that it can have on adversarial patch attacks. We measure the average proportion of the number of adversarial samples for which the location of highest attention in the attention map does not come within the attack surface area. We term this measure as “NoPatchLoc”, which is defined as follows:

$$\text{NoPatchLoc} = \frac{1}{N}\sum_{i=0}^{N}(1 - \mathbf{1}_A(x_i, y_i, Ox_i, Oy_i)), \quad (2)$$

where N is the total number of adversarial samples analyzed, and the indicator function is defined as follows:

$$\mathbf{1}_A(x_i, y_i, O_{x_i}, O_{y_i}) = \begin{cases} 1, & \text{if } O_{x_i} \leq x_i < O_{x_i} + s \text{ and } O_{y_i} \leq y_i < O_{y_i} + s \\ 0, & \text{otherwise} \end{cases}, \quad (3)$$

where s denotes the patch size, (O_{x_i}, O_{y_i}) is the optimal location identified by our method to locate the adversarial patch, and (x_i, y_i) is the coordinate of the highest attention location. Table 1 demonstrates the NoPatchLoc measures obtained from the generated adversarial samples corresponding to their respective victim models. As evident, except for ResNet-50, where the measure is 53.70%, the highest attention location remains consistently outside the attack region for more than 70% of the perturbed samples across all other architectures. The highest occurrence is observed for the Swin Transformer Base, achieving 81.30%. This provides strong evidence that accounts for the strong stealth capabilities of our method, as highlighted by the performance against defense methods.

C.3. Transferability

We assess the transferability of our general method in the untargeted scenario without incorporating any adaptations specifically aimed at enhancing attack transferability. Using a substitute model approach, we generate adversarial samples on the previously considered victim models and evaluate their transferability across a set of target models: SqueezeNet, ResNet-18, ResNet-34, VGG11, VGG13, and VGG19. Given that no specific adaptation scheme is used, our method achieves reasonable ASR as shown in Figure 1. The results indicate that transferability is influenced by the architectural similarity between the substitute and target models, as well as their relative model sizes.

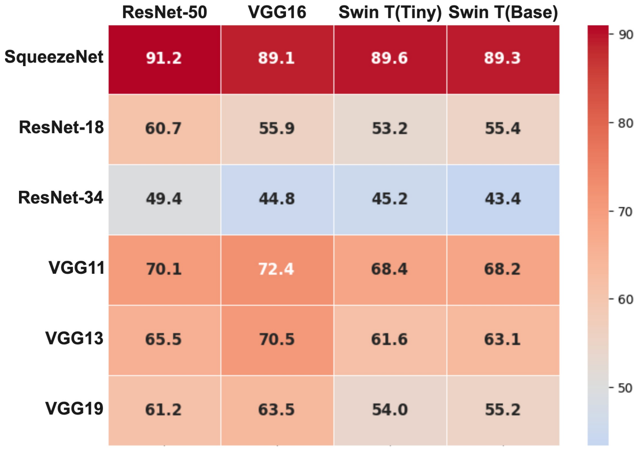

Figure 1. Transferability of IAP measured by ASR (%) on ImageNet. The first row represents the substitute model, and the first column represents the target models.

Model	NoPatchLoc(%)
VGG16	72.15
ResNet-50	53.70
Swin Transformer Tiny	73.11
Swin Transformer Base	81.30
Average	**70.07**

Table 1. Assessment of whether the GradCAM’s highest attention location overlaps with the adversarial patch location.

C.4. Black-box Adaptation

While IAP is initially designed as a white-box method, it can be successfully adapted to black-box settings. We ran additional experiments on ImageNet using the following black-box variation of IAP. Specifically, we first approximate the Grad-CAM localization map using a surrogate model (i.e., ResNet-50) for patch placement. Subsequently, we employ a hybrid approach for perturbation optimization, where we initialize the perturbations based on the same surrogate model and refine them using NES, a query-based attack algorithm. The results are shown in Table 3 in the main paper, where our black-box IAP variant achieves high (untargeted) attack success rates across different target models. We test 500 samples per model with a patch size of 84 and other parameters fixed. Based on white-box convergence trends, we run 400 surrogate iterations followed by 200 query-based steps, requiring at most $12,000$ queries.

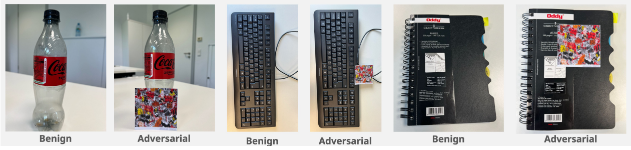

Figure 2. Illustrative images of physical-world applications of IAP.

Shape	ASR	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
Circle	99.2%	Local	0.95	0.88	27.10	91.46	0.085
		Global	0.99	0.98	37.79	99.13	0.016

Table 2. Performance of IAP in ASR and various imperceptibility metrics with a circular patch shape and patch size of 11%.

C.5. Physical-World Applicability

Additionally, we examine IAP's generalizability to physical-world, untargeted attack settings on 5 object classes. Patches are generated using our optimization scheme, initialized from a reference sticker image like PS-GANs. To ensure location invariance, each patch is trained by randomly placing it across four proposed "optimal" positions from different models. Printed patches are tested on 5 images per object under varying viewpoints (see Figure 2 for illustrative examples), achieving an average ASR of 70%, showing the potential of IAP's adaptability to physical domains.

C.6. Flexibility in Patch Shape

We also run experiments to study whether our method for generating invisible adversarial patches is shape-agnostic. Results are shown in Table 2, where we evaluate the performance of IAP using a circular patch with a diameter of 84 pixels (11% image area). Under our best-performing setup with Swin Transformer Base as the target model, IAP achieves a high ASR of 99.2% while preserving imperceptibility with LPIPS as low as 0.085. We believe similar results can also be achieved for other typical patch shapes, since our attack framework supports arbitrary binary masks.

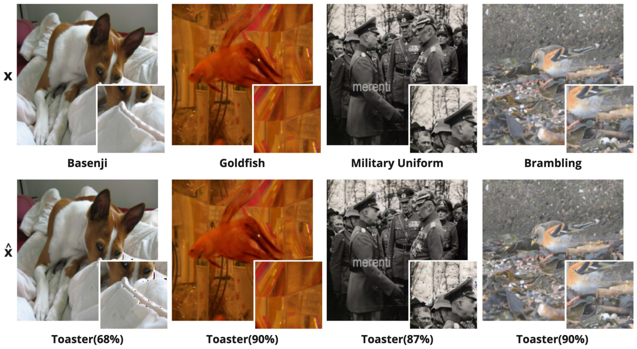

Figure 3. Visualizations of the original images and their adversarial counterparts produced by IAP corresponding to the target class on the ImageNet Dataset with **VGG16** as the victim model. x represents the benign sample, and $\hat{x}$ represents the adversarial samples with the generated adversarial patch corresponding to the target class. The smaller images at the right-bottom corner correspond to the optimal location (i', j').

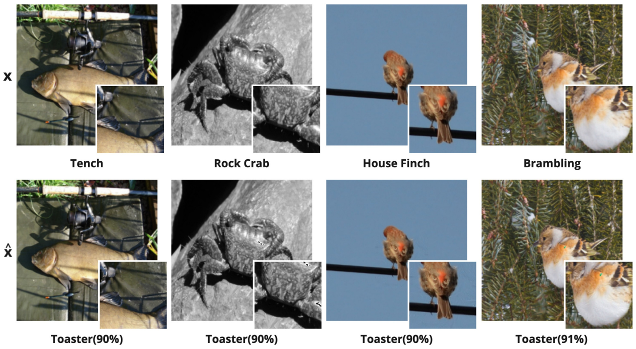

Figure 4. Visualizations of the original images and their adversarial counterparts produced by IAP corresponding to the target class on the ImageNet Dataset with **ResNet-50** as the victim model. x represents the benign sample, and $\hat{x}$ represents the adversarial samples with the generated adversarial patch corresponding to the target class. The smaller images at the right-bottom corner correspond to the optimal location (i', j').

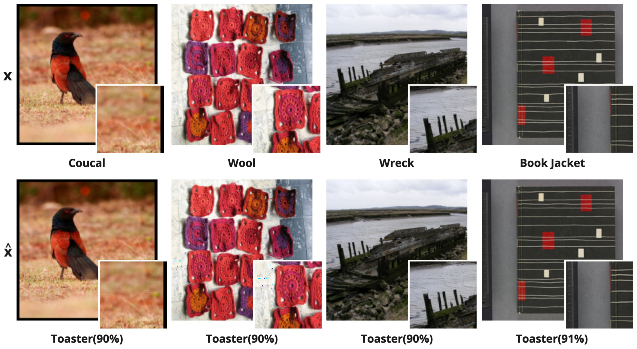

Figure 5. Visualizations of the original images and their adversarial counterparts produced by IAP corresponding to the target class on the ImageNet Dataset with **Swin Transformer Tiny** as the victim model. x represents the benign sample, and $\hat{x}$ represents the adversarial samples with the generated adversarial patch corresponding to the target class. The smaller images at the right-bottom corner correspond to the optimal location (i', j').

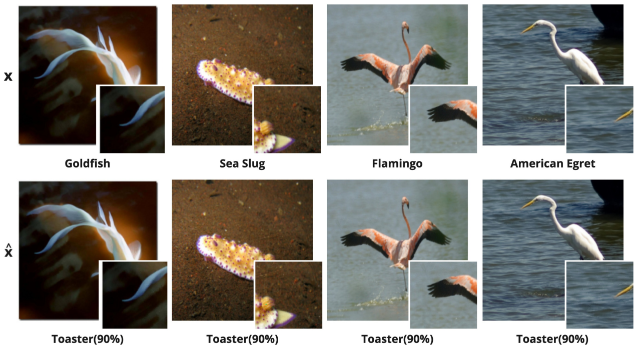

Figure 6. Visualizations of the original images and their adversarial counterparts produced by IAP corresponding to the target class on the ImageNet Dataset with **Swin Transformer Base** as the victim model. x represents the benign sample, and $\hat{x}$ represents the adversarial samples with the generated adversarial patch corresponding to the target class. The smaller images at the right-bottom corner correspond to the optimal location (i', j').

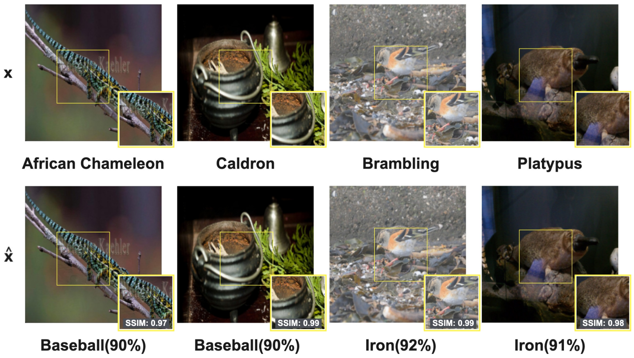

Figure 7. Visualizations of the original images and their adversarial counterparts produced by IAP corresponding to the target class on the ImageNet Dataset with **ResNet-50** as the victim model. x represents the benign sample, and $\hat{x}$ represents the adversarial samples with the generated adversarial patch corresponding to the target class. The smaller images at the right-bottom corner correspond to the optimal location (i', j').

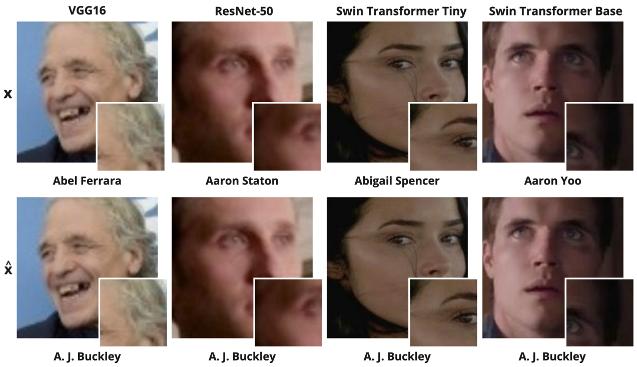

Figure 8. Visualizations of the original images and their adversarial counterparts with IAP and the target class **"A. J. Buckley"** on the VGG Face Dataset. x represents the benign sample, and $\hat{x}$ represents the adversarial samples with the generated adversarial patch corresponding to the target class. The smaller images at the right-bottom corner correspond to the optimal location (i', j').

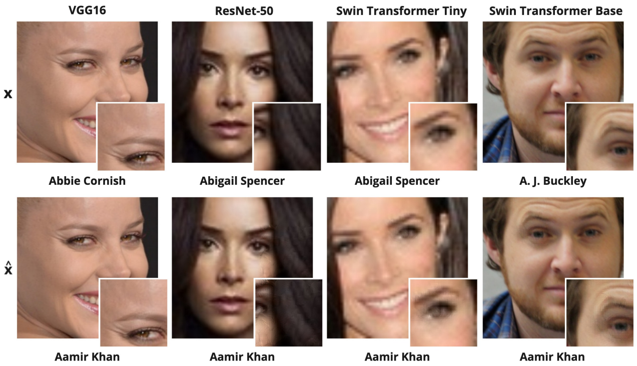

Figure 9. Visualizations of the original images and their adversarial counterparts with IAP and the target class **"Aamir Khan"** on the VGG Face Dataset. x represents the benign sample, and $\hat{x}$ represents the adversarial samples with the generated adversarial patch corresponding to the target class. The smaller images at the right-bottom corner correspond to the optimal location (i', j').

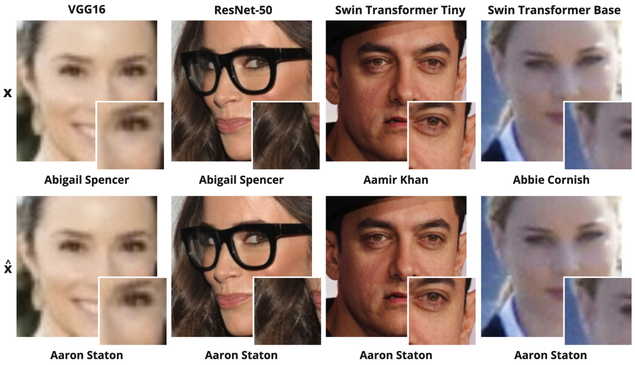

Figure 10. Visualizations of the original images and their adversarial counterparts with IAP and the target class **"Aaron Staton"** on the VGG Face Dataset. x represents the benign sample, and $\hat{x}$ represents the adversarial samples with the generated adversarial patch corresponding to the target class. The smaller images at the right-bottom corner correspond to the optimal location (i', j').

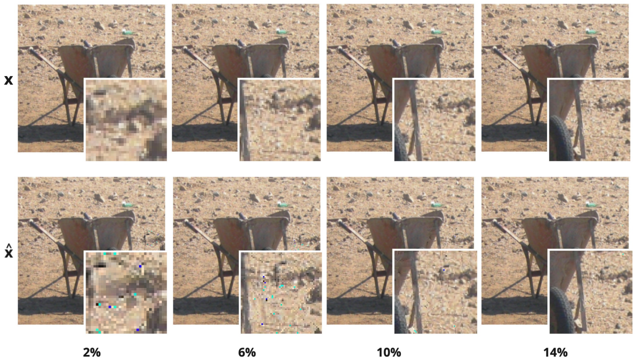

Figure 11. Visualizations of the impact of the patch sizes on attack imperceptibility. x represents the benign sample, and $\hat{x}$ represents the adversarial samples with the generated adversarial patch corresponding to the target class. The smaller images at the right-bottom corner correspond to the optimal location (i', j').

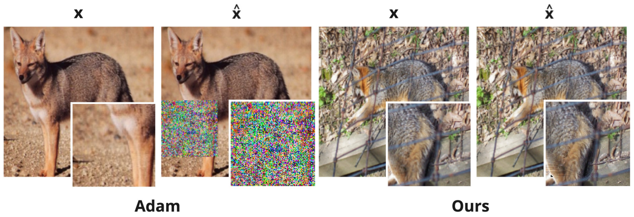

Figure 12. Visualizations of adversarial patch generated by update rule from Adam optimizer vs IAP. x represents the benign sample, and $\hat{x}$ represents the adversarial samples with the generated adversarial patch corresponding to the target class. The smaller images at the right-bottom corner correspond to the optimal location (i', j').

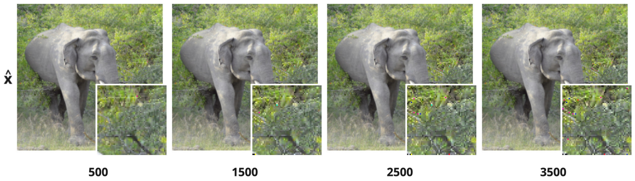

Figure 13. Visualizations of the impact of the number of update iterations on attack imperceptibility. $\hat{x}$ represents the adversarial samples with the generated adversarial patch. The smaller images at the right-bottom corner correspond to the optimal location (i', j'). The x-axis represents the number of update iterations.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
Google Patch	**100**	Local	0.002	0.000	11.93	32.50	0.760
		Global	0.830	0.820	18.73	73.10	0.190
LaVAN	93.6	Local	0.002	0.000	11.13	33.20	0.790
		Global	0.820	0.810	20.30	76.32	0.230
GDPA	89.2	Local	0.310	0.300	19.90	56.25	0.610
		Global	0.890	0.880	28.00	84.00	0.130
MPGD	96.5	Local	0.810	0.800	26.44	73.91	0.320
		Global	0.940	0.920	32.80	94.00	0.090
Ours	**99.1**	Local	**0.900**	**0.860**	**28.94**	**72.70**	**0.230**
		Global	**0.985**	**0.960**	**36.42**	**95.10**	**0.060**

Table 3. Detailed comparison of attack efficacy through ASR (%) and imperceptibility with **VGG16** as the victim model on the **ImageNet** dataset. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LIPIPS.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
Google Patch	99.1	Local	0.010	0.000	14.20	33.00	0.740
		Global	0.820	0.810	22.90	74.10	0.180
LaVAN	**100**	Local	0.010	0.000	14.20	33.30	0.780
		Global	0.820	0.810	23.40	76.10	0.180
GDPA	93.7	Local	0.350	0.330	19.80	65.20	0.570
		Global	0.920	0.910	28.40	87.10	0.090
MPGD	97.8	Local	0.790	0.780	25.30	76.20	0.240
		Global	0.950	0.930	33.60	93.30	0.050
Ours	**99.5**	Local	**0.940**	**0.910**	**28.34**	**84.54**	**0.120**
		Global	**0.990**	**0.970**	**37.23**	**96.52**	**0.020**

Table 4. Detailed comparison of attack efficacy through ASR (%) and imperceptibility with **ResNet-50** as the victim model on the ImageNet dataset. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LIPIPS.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
Google Patch	**99.8**	Local	0.002	0.000	11.80	32.80	0.770
		Global	0.830	0.820	18.94	73.90	0.150
LaVAN	99.7	Local	0.005	0.000	14.13	33.10	0.780
		Global	0.820	0.810	23.30	76.32	0.170
GDPA	83.7	Local	0.390	0.360	20.20	63.65	0.540
		Global	0.900	0.890	28.21	85.75	0.100
MPGD	98.8	Local	0.800	0.790	25.50	80.54	0.190
		Global	0.940	0.920	33.11	95.80	0.050
Ours	**99.6**	Local	**0.980**	**0.940**	**31.74**	**90.41**	**0.060**
		Global	**0.996**	**0.980**	**40.67**	**98.61**	**0.008**

Table 5. Detailed comparison of ASR (%) and imperceptibility with **Swin Transformer Tiny** as the victim model on the ImageNet dataset. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LIPIPS.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
Google Patch	97.9	Local	0.003	0.000	10.74	32.90	0.770
		Global	0.830	0.820	17.61	73.20	0.170
LaVAN	**100**	Local	0.004	0.000	13.10	33.19	0.780
		Global	0.820	0.810	23.30	76.35	0.180
GDPA	85.1	Local	0.360	0.345	20.40	61.25	0.540
		Global	0.880	0.870	28.00	85.10	0.110
MPGD	70.5	Local	0.800	0.800	25.30	74.30	0.200
		Global	0.940	0.920	33.00	92.10	0.050
Ours	**99.4**	Local	**0.970**	**0.910**	**31.30**	**89.33**	**0.070**
		Global	**0.994**	**0.970**	**40.10**	**98.43**	**0.010**

Table 6. Detailed comparison of ASR (%) and imperceptibility with **Swin Transformer Base** as the victim model on the ImageNet dataset. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LIPIPS.

$\mathbf{y}_{targ}$	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
Ipod	99.6	Local	0.95	0.92	28.9	86.8	0.115
		Global	0.99	0.97	37.8	97.1	0.018
Baseball	99.3	Local	0.94	0.91	28.5	85.0	0.118
		Global	0.99	0.97	37.4	96.7	0.019
Toaster	99.5	Local	0.94	0.91	28.3	84.5	0.120
		Global	0.990	0.970	37.23	96.52	0.020

Table 7. Detailed evaluation of attack efficacy through ASR (%) and imperceptibility for different target classes within the ImageNet Dataset. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑), the better, while the lower (↓), the better for LPIPS.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
Google Patch	**100**	Local	0.000	0.000	11.95	36.82	0.890
		Global	0.812	0.820	19.46	68.22	0.270
LaVAN	**100**	Local	0.006	0.000	15.85	36.55	0.865
		Global	0.820	0.825	24.18	71.84	0.220
GDPA	96.12	Local	0.240	0.220	21.00	57.96	0.660
		Global	0.870	0.865	29.00	75.66	0.151
MPGD	88.9	Local	0.620	0.533	28.30	65.30	0.400
		Global	0.960	0.935	36.70	86.70	0.087
Ours	**100**	Local	**0.930**	**0.880**	**31.81**	**66.50**	**0.207**
		Global	**0.990**	**0.980**	**40.11**	**88.57**	**0.039**

Table 8. Detailed evaluation and comparison of attack efficacy through ASR (%) and imperceptibility with **VGG16** as the victim model on the VGG Face dataset for the Target class **"A. J. Buckley"**. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
Google Patch	**99.9**	Local	0.000	0.000	11.76	36.43	0.860
		Global	0.810	0.820	19.36	68.22	0.270
LaVAN	99.5	Local	0.005	0.000	15.64	36.52	0.850
		Global	0.820	0.825	24.06	71.56	0.220
GDPA	99.50	Local	0.220	0.190	21.46	55.50	0.685
		Global	0.850	0.840	55.50	63.41	0.190
MPGD	86.85	Local	0.650	0.550	27.80	65.20	0.420
		Global	0.950	0.930	36.10	86.60	0.090
Ours	**98.8**	Local	**0.924**	**0.870**	**31.94**	**68.24**	**0.200**
		Global	**0.990**	**0.980**	**40.08**	**88.70**	**0.039**

Table 9. Detailed evaluation and comparison of attack efficacy through ASR (%) and imperceptibility with **VGG16** as the victim model on the VGG Face dataset for the Target class **"Aamir Khan"**. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
Google Patch	**100**	Local	0.000	0.000	10.76	36.65	0.860
		Global	0.810	0.820	18.27	68.70	0.290
LaVAN	**100**	Local	0.003	0.000	11.89	36.45	0.870
		Global	0.820	0.824	20.30	71.67	0.260
GDPA	91.50	Local	0.476	0.465	22.85	60.48	0.53
		Global	0.900	0.890	29.45	76.00	0.125
MPGD	84.95	Local	0.680	0.564	27.30	65.10	0.440
		Global	0.940	0.924	35.88	85.10	0.094
Ours	**99.53**	Local	**0.904**	**0.850**	**31.40**	**65.80**	**0.217**
		Global	**0.985**	**0.980**	**39.61**	**87.72**	**0.042**

Table 10. Detailed evaluation and comparison of attack efficacy through ASR (%) and imperceptibility with **VGG16** as the victim model on the VGG Face dataset for the Target class **"Aaron Staton"**. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
Google Patch	98.0	Local	0.010	0.000	17.52	38.81	0.730
		Global	0.830	0.820	24.25	63.13	0.210
LaVAN	**100**	Local	0.007	0.000	16.80	36.81	0.840
		Global	0.840	0.826	25.12	71.64	0.200
GDPA	99.5	Local	0.310	0.250	22.00	53.00	0.660
		Global	0.880	0.860	29.00	59.00	0.170
MPGD	78.1	Local	0.620	0.560	26.99	61.78	0.380
		Global	0.950	0.930	35.56	85.42	0.080
Ours	**98.8**	Local	**0.920**	**0.880**	**32.11**	**69.40**	**0.170**
		Global	**0.990**	**0.980**	**40.66**	**90.55**	**0.030**

Table 11. Detailed evaluation and comparison of attack efficacy through ASR (%) and imperceptibility with **ResNet-50** as the victim model on the VGG Face dataset for the Target class **"A. J. Buckley"**. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
Google Patch	99.5	Local	0.001	0.000	16.47	38.80	0.800
		Global	0.830	0.820	21.89	63.13	0.270
LaVAN	**100**	Local	0.007	0.000	16.89	36.82	0.830
		Global	0.840	0.826	25.30	71.51	0.210
GDPA	99.70	Local	0.280	0.230	21.99	56.73	0.600
		Global	0.870	0.850	56.73	59.32	0.200
MPGD	70.74	Local	0.610	0.550	26.60	59.87	0.390
		Global	0.940	0.930	35.30	84.63	0.080
Ours	**93.0**	Local	**0.890**	**0.830**	**30.88**	**65.75**	**0.226**
		Global	**0.980**	**0.970**	**39.37**	**87.30**	**0.040**

Table 12. Detailed evaluation and comparison of attack efficacy through ASR (%) and imperceptibility with **ResNet-50** as the victim model on the VGG Face dataset for the Target class **"Aamir Khan"**. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
Google Patch	80.3	Local	0.010	0.000	17.52	38.81	0.730
		Global	0.830	0.820	24.25	63.13	0.210
LaVAN	**97.0**	Local	0.010	0.000	17.45	41.54	0.750
		Global	0.830	0.820	22.32	62.68	0.240
GDPA	98.00	Local	0.330	0.280	22.10	55.68	0.60
		Global	0.880	0.850	29.12	57.54	0.200
MPGD	52.50	Local	0.610	0.550	26.83	60.25	0.380
		Global	0.940	0.930	35.25	83.42	0.080
Ours	**91.80**	Local	**0.890**	**0.840**	**30.89**	**65.70**	**0.216**
		Global	**0.980**	**0.970**	**39.33**	**88.32**	**0.040**

Table 13. Detailed evaluation and comparison of attack efficacy through ASR (%) and imperceptibility with **ResNet-50** as the victim model on the VGG Face dataset for the Target class **"Aaron Staton"**. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
Google Patch	98.9	Local	0.040	0.000	10.12	36.10	0.820
		Global	0.830	0.830	16.87	66.88	0.260
LaVAN	**100**	Local	0.007	0.000	16.49	36.50	0.850
		Global	0.840	0.825	24.75	71.87	0.210
GDPA	92.9	Local	0.330	0.270	21.85	62.10	0.570
		Global	0.880	0.870	29.30	71.76	0.140
MPGD	95.5	Local	0.630	0.540	27.65	62.48	0.380
		Global	0.950	0.930	35.72	86.66	0.070
Ours	**99.3**	Local	**0.860**	**0.800**	**29.22**	**63.28**	**0.275**
		Global	**0.980**	**0.970**	**38.00**	**87.83**	**0.048**

Table 14. Detailed evaluation and comparison of attack efficacy through ASR (%) and imperceptibility with **Swin Transformer Tiny** as the victim model on the VGG Face dataset for the Target class **"A. J. Buckley"**. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	**UIQ (↑)**	**SRE (↑)**	**CLIP (↑)**	**LPIPS (↓)**
Google Patch	99.2	Local	0.000	0.000	10.11	37.23	0.780
		Global	0.830	0.820	17.22	67.50	0.230
LaVAN	**100**	Local	0.006	0.000	16.31	36.57	0.850
		Global	0.840	0.825	24.71	71.73	0.210
GDPA	100	Local	0.340	0.300	19.85	60.84	0.600
		Global	0.910	0.910	29.82	80.01	0.100
MPGD	94.87	Local	0.640	0.550	27.68	62.69	0.370
		Global	0.950	0.930	35.80	86.97	0.070
Ours	**99.3**	Local	**0.870**	**0.820**	**29.80**	**63.00**	**0.240**
		Global	**0.980**	**0.970**	**38.60**	**88.20**	**0.043**

Table 15. Detailed evaluation and comparison of attack efficacy through ASR (%) and imperceptibility with **Swin Transformer Tiny** as the victim model on the VGG Face dataset for the Target class **“Aamir Khan”**. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	**UIQ (↑)**	**SRE (↑)**	**CLIP (↑)**	**LPIPS (↓)**
Google Patch	99.3	Local	0.000	0.000	12.60	38.84	0.820
		Global	0.830	0.820	17.48	63.13	0.290
LaVAN	**100**	Local	0.007	0.000	16.45	36.67	0.850
		Global	0.840	0.825	24.82	72.06	0.210
GDPA	92.4	Local	0.310	0.260	20.19	54.86	0.65
		Global	0.860	0.840	27.21	60.54	0.220
MPGD	96.2	Local	0.640	0.550	27.76	61.90	0.360
		Global	0.950	0.930	35.85	87.30	0.070
Ours	**98.60**	Local	**0.860**	**0.800**	**29.64**	**62.00**	**0.260**
		Global	**0.980**	**0.970**	**38.34**	**87.90**	**0.046**

Table 16. Detailed evaluation and comparison of attack efficacy through ASR (%) and imperceptibility with **Swin Transformer Tiny** as the victim model on the VGG Face dataset for the Target class **“Aaron Staton”**. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	**UIQ (↑)**	**SRE (↑)**	**CLIP (↑)**	**LPIPS (↓)**
Google Patch	98.2	Local	0.000	0.000	11.23	36.51	0.835
		Global	0.830	0.820	18.23	67.65	0.240
LaVAN	**100**	Local	0.005	0.000	15.47	36.52	0.850
		Global	0.840	0.825	23.80	71.82	0.220
GDPA	77.24	Local	0.410	0.360	21.59	58.14	0.56
		Global	0.910	0.900	29.66	72.23	0.110
MPGD	97.9	Local	0.600	0.520	27.45	61.22	0.390
		Global	0.940	0.920	35.57	85.00	0.080
Ours	**99.0**	Local	**0.860**	**0.780**	**29.8**	**63.00**	**0.300**
		Global	**0.980**	**0.960**	**38.10**	**86.00**	**0.055**

Table 17. Detailed evaluation and comparison of attack efficacy through ASR (%) and imperceptibility with **Swin Transformer Base** as the victim model on the VGG Face dataset for the Target class **“A. J. Buckley”**. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	**UIQ (↑)**	**SRE (↑)**	**CLIP (↑)**	**LPIPS (↓)**
Google Patch	97.2	Local	0.000	0.000	10.78	36.85	0.900
		Global	0.830	0.820	18.10	69.36	0.260
LaVAN	**99.3**	Local	0.004	0.000	15.00	36.49	0.850
		Global	0.840	0.824	23.40	71.68	0.220
GDPA	55.1	Local	0.160	0.140	18.36	65.87	0.700
		Global	0.920	0.920	30.10	84.10	0.090
MPGD	80.81	Local	0.610	0.530	27.35	61.22	0.390
		Global	0.940	0.930	35.47	85.28	0.080
Ours	**97.0**	Local	**0.840**	**0.760**	**29.63**	**61.00**	**0.300**
		Global	**0.970**	**0.960**	**37.84**	**86.00**	**0.055**

Table 18. Detailed evaluation and comparison of attack efficacy through ASR (%) and imperceptibility with **Swin Transformer Base** as the victim model on the VGG Face dataset for the Target class **“Aamir Khan”**. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS.

Method	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	**UIQ (↑)**	**SRE (↑)**	**CLIP (↑)**	**LPIPS (↓)**
Google Patch	98.3	Local	0.000	0.000	11.86	35.58	0.920
		Global	0.830	0.820	17.78	68.63	0.280
LaVAN	**99.8**	Local	0.006	0.000	15.73	36.70	0.850
		Global	0.840	0.825	24.12	71.96	0.210
GDPA	84.9	Local	0.290	0.260	19.76	57.25	0.620
		Global	0.910	0.910	29.72	78.20	0.100
MPGD	94.9	Local	0.630	0.550	27.75	61.50	0.370
		Global	0.950	0.930	35.84	86.54	0.070
Ours	**98.6**	Local	**0.880**	**0.810**	**30.50**	**63.50**	**0.250**
		Global	**0.980**	**0.970**	**38.84**	**88.40**	**0.045**

Table 19. Detailed evaluation and comparison of attack efficacy through ASR (%) and imperceptibility with **Swin Transformer Base** as the victim model on the VGG Face dataset for the Target class **“Aaron Staton”**. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS.

Patch Size(%)	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	**UIQ (↑)**	**SRE (↑)**	**CLIP (↑)**	**LPIPS (↓)**
2	72.2	Local	0.640	0.530	21.07	70.00	0.413
		Global	0.992	0.985	38.10	98.20	0.014
4	90.7	Local	0.784	0.683	23.32	70.77	0.308
		Global	0.991	0.972	37.68	98.11	0.014
6	94.2	Local	0.854	0.756	25.02	74.74	0.024
		Global	0.991	0.970	37.86	98.18	0.013
8	97.3	Local	0.896	0.810	26.77	78.05	0.183
		Global	0.991	0.970	38.14	98.15	0.012
10	98.1	Local	0.920	0.840	27.90	80.91	0.152
		Global	0.992	0.970	38.31	97.97	0.011
12	99.0	Local	0.934	0.860	28.90	83.43	0.126
		Global	0.992	0.965	38.46	98.03	0.011
14	**99.4**	Local	**0.970**	**0.910**	**31.30**	**89.33**	**0.070**
		Global	**0.994**	**0.970**	**40.10**	**98.43**	**0.010**

Table 20. Impact of **patch size** on attack performance represented through ASR (%) and imperceptibility with **Swin Transformer Base** as the victim model on the ImageNet dataset. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS.

w_3	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
0	99.0	Local	0.943	0.873	29.76	85.54	0.111
		Global	0.992	0.964	38.59	97.95	0.017
1	98.9	Local	0.944	0.874	29.79	85.65	0.110
		Global	0.992	0.965	38.62	97.97	0.017
4	98.9	Local	0.945	0.875	29.83	85.66	0.109
		Global	0.992	0.965	38.67	97.99	0.017
7	98.8	Local	0.946	0.876	29.84	85.68	0.108
		Global	0.992	0.966	38.69	98.01	0.016
10	99.1	Local	0.945	0.875	29.83	85.71	0.108
		Global	0.992	0.965	38.66	97.98	0.017
13	99.0	Local	0.944	0.874	29.78	85.56	0.110
		Global	0.992	0.965	38.60	97.98	0.017

Table 21. Impact of distance term regularization coefficient w_3 on attack performance represented through ASR (%) and imperceptibility with **Swin Transformer Base** as the victim model on the ImageNet dataset. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS.

Update Rule	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
Adam	**100**	Local	0.130	0.157	17.13	36.15	0.662
		Global	0.867	0.848	25.98	80.94	0.130
Ours	99.4	Local	**0.970**	**0.910**	**31.30**	**89.33**	**0.070**
		Global	**0.994**	**0.970**	**40.10**	**98.43**	**0.010**

Table 22. Impact of the **update rule** on attack performance represented through ASR (%) and imperceptibility with **Swin Transformer Base** as the victim model on the ImageNet dataset. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS.

No. Iters	ASR(%)	Scale	Imperceptibility metric				
			SSIM (↑)	UIQ (↑)	SRE (↑)	CLIP (↑)	LPIPS (↓)
500	86.0	Local	**0.870**	**0.770**	**25.88**	**75.87**	**0.223**
		Global	**0.992**	**0.972**	**38.48**	**98.23**	**0.015**
1000	94.2	Local	0.854	0.756	25.02	74.74	0.024
		Global	0.991	0.970	37.86	98.18	0.013
1500	96.2	Local	0.850	0.755	24.91	73.81	0.024
		Global	0.991	0.969	37.70	98.10	0.016
2000	97.3	Local	0.843	0.749	24.77	73.29	0.246
		Global	0.990	0.968	37.59	98.04	0.017
2500	98.0	Local	0.840	0.746	24.67	72.87	0.249
		Global	0.990	0.967	37.50	98.02	0.017
3000	98.5	Local	0.836	0.743	24.48	72.78	0.252
		Global	0.990	0.966	37.41	97.98	0.017
3500	**98.6**	Local	0.834	0.741	24.65	72.49	0.254
		Global	0.990	0.969	37.40	97.92	0.017

Table 23. Impact of **number of update iterations** on attack performance, represented through ASR (%) and imperceptibility with **Swin Transformer Base** as the victim model on the ImageNet dataset. For SSIM, UIQ, SRE, and CLIP scores, the higher (↑) the better, while the lower (↓) the better for LPIPS. Patch size is kept fixed at 6%.

References

[1] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. 1

[2] Zitao Chen, Pritam Dash, and Karthik Pattabiraman. Jujutsu: A two-stage defense against adversarial patch attacks on deep neural networks, 2022. 1

[3] Jia Fu, Xiao Zhang, Sepideh Pashami, Fatemeh Rahimian, and Anders Holst. Diffpad: Denoising diffusion-based adversarial patch decontamination. *arXiv preprint arXiv:2410.24006*, 2024.

[4] Jamie Hayes. On visible adversarial perturbations & digital watermarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1597–1604, 2018. 1

[5] 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 (CVPR)*, pages 770–778, 2016. 1

[6] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 1

[7] Caixin Kang, Yinpeng Dong, Zhengyi Wang, Shouwei Ruan, Yubo Chen, Hang Su, and Xingxing Wei. Diffender: Diffusion-based adversarial defense against patch attacks. In *European Conference on Computer Vision*, pages 130–147. Springer, 2024. 1

[8] Danny Karmon, Daniel Zoran, and Yoav Goldberg. Lavan: Localized and visible adversarial noise. In *International Conference on Machine Learning*, pages 2507–2515. PMLR, 2018. 1

[9] Charis Lanaras, José Bioucas-Dias, Silvano Galliani, Emmanuel Baltsavias, and Konrad Schindler. Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:305–319, 2018. 1

[10] Xiang Li and Shihao Ji. Generative dynamic patch attack. *arXiv preprint arXiv:2111.04266*, 2021. 1

[11] Jiang Liu, Alexander Levine, Chun Pong Lau, Rama Chellappa, and Soheil Feizi. Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14973–14982, 2022. 1

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 1

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1

[14] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 1

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. 1

[16] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

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

[18] Bilel Tarchoun, Anouar Ben Khalifa, Mohamed Ali Mahjoub, Nael Abu-Ghazaleh, and Ihsen Alouani. Jedi: entropy-based localization and removal of adversarial patches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4087–4095, 2023. 1

[19] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE signal processing letters*, 9(3):81–84, 2002. 1

[20] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[21] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1