Seeing Isn't Believing: Context-Aware Adversarial Patch Synthesis via Conditional GAN - Appendix

Roie Kazoom*, Alon Goldberg, Hodaya Cohen, Ofer Hadar
Ben Gurion University of the Negev roieka@post.bgu.ac.il

Contents

1. Ablation Study on Loss Functions

We investigate the individual contribution of each loss component to the overall objective. Recall that the complete optimization is defined as:

$$\mathcal{L} = \mathcal{L}_{adv} + \mathcal{L}_{patch} + \mathcal{L}_{perc}, \tag{1}$$

where each term plays a distinct role:

- **Adversarial loss** $\mathcal{L}_{adv}$ enforces targeted misclassification into the attacker-specified class. It is the driving force behind adversarial effectiveness, ensuring that the patched image $\tilde{x}$ is predicted as the target class regardless of its original semantics.
- **Patch consistency loss** $\mathcal{L}_{patch}$ constrains the generated patch $G(\delta)$ to remain visually close to the seed patch δ. This stabilizes training, prevents mode collapse, and ensures that the adversarial patch retains a coherent texture rather than degenerating into noisy patterns.
- **Perceptual loss** $\mathcal{L}_{perc}$ enforces similarity in a high-level feature space using activations from a frozen network (e.g., VGG16). This encourages the generated patch to preserve natural image statistics and remain visually plausible while embedding the adversarial signal.

To assess the impact of each term, we evaluate the following configurations:

1. $\mathcal{L}_{adv}$ only
2. $\mathcal{L}_{adv} + \mathcal{L}_{patch}$
3. $\mathcal{L}_{adv} + \mathcal{L}_{perc}$
4. $\mathcal{L}_{adv} + \mathcal{L}_{patch} + \mathcal{L}_{perc}$ (full objective)
5. $\mathcal{L}_{patch}$ only
6. $\mathcal{L}_{perc}$ only
7. $\mathcal{L}_{perc} + \mathcal{L}_{patch}$

As shown in Table 1, the results highlight several key insights:

- Using $\mathcal{L}_{adv}$ alone achieves targeted misclassification but yields relatively weak performance, with both ASR and TCS capped below 76%. This confirms that misclassification alone is insufficient for stable and realistic patch generation.

- Using $\mathcal{L}_{patch}$ or $\mathcal{L}_{perc}$ alone produces visually stable and realistic patches but fails to induce strong targeted misclassification, resulting in substantially lower ASR and TCS values.

- Combining $\mathcal{L}_{adv}$ with either $\mathcal{L}_{patch}$ or $\mathcal{L}_{perc}$ moderately improves results, though still falls short of state-of-the-art robustness.

- The complete loss $\mathcal{L}_{adv}+\mathcal{L}_{patch}+\mathcal{L}_{perc}$ yields the best trade-off, achieving near-perfect ASR (99.89%–99.99%) and TCS (99.88%–99.98%) across patch placements.

These findings confirm that the three losses are highly complementary: adversarial enforcement drives targeted misclassification, patch consistency ensures stability, and perceptual similarity enforces realism. Together, they are necessary to produce robust, transferable, and visually plausible adversarial patches.

2. Realism vs. Non-Realism

We evaluate the effect of perceptual and consistency losses on adversarial patch synthesis by distinguishing *realistic* from *non-realistic* patches. A patch is considered **realistic** if at least 8 out of 10 human evaluators judged it to blend naturally into the scene, without exhibiting unnatural color distortions. Otherwise, it is **non-realistic**.

Formally, for a patch p with human ratings $h_i \in \{0, 1\}$, $i = 1, \dots, 10$, we define

$$R(p) = \mathbf{1}\left[\sum_{i=1}^{10} h_i \geq 8\right] \tag{2}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function.

In addition to human evaluation, we report SSIM and LPIPS as perceptual metrics. Training with perceptual and consistency losses yields patches with improved realism ($R(p) = 1$), while maintaining strong attack success rate (ASR) and target class success (TCS).

3. Effect of Patch Size on Attack Success (ResNet, ImageNet)

We further analyze the impact of patch size on adversarial effectiveness using ResNet trained on ImageNet. Table 2 reports the attack success rate (ASR) and target-class success (TCS) for varying patch sizes from 8×8 to 128×128. The ASR measures the proportion of inputs misclassified into *any* incorrect label, while TCS measures the proportion redirected specifically into the attacker-specified target class. Formally,

$$\text{ASR} = \frac{\#\{\tilde{x} : f(\tilde{x}) \neq y\}}{\#\{x\}}, \qquad \text{TCS} = \frac{\#\{\tilde{x} : f(\tilde{x}) = t\}}{\#\{x\}}, \tag{3}$$

where y is the ground-truth label, t is the attacker-specified target class, and $\tilde{x}$ denotes the adversarial example.

Figure 1 visualizes these results. Both ASR and TCS increase monotonically with patch size. Small patches such as 8×8 achieve only limited effectiveness (ASR $= 19.32\%$, TCS $= 5.45\%$), while medium patches like 32×32 already surpass ASR $= 97.75\%$ and TCS $= 93.79\%$. At 64×64 and above, the attack becomes nearly perfect, converging to ASR $\approx 100\%$ and TCS $\approx 100\%$.

These findings highlight that adversarial effectiveness scales with the available perturbation budget: larger patches have greater capacity to embed adversarial signals while maintaining control over targeted misclassification.

Table 1. Ablation study on loss functions. We evaluate different combinations of $\mathcal{L}_{adv}$, $\mathcal{L}_{patch}$, and $\mathcal{L}_{perc}$ under multiple placement strategies. Accuracy before attack is reported along with attack success rate (ASR) and target-class success (TCS).

Loss Setting	**Placement**	**Accuracy Before Attack (%)** ↑	**ASR (%)** ↓	**TCS (%)** ↓
$\mathcal{L}_{adv}$ only	Center	77.90	75.62	72.48
	Random	77.90	74.35	70.19
	Grad-CAM	77.90	73.84	71.55
$\mathcal{L}_{patch}$ only	Center	77.90	40.17	15.23
	Random	77.90	35.54	14.87
	Grad-CAM	77.90	42.26	18.14
$\mathcal{L}_{perc}$ only	Center	77.90	45.28	20.37
	Random	77.90	38.46	12.18
	Grad-CAM	77.90	47.93	21.57
$\mathcal{L}_{adv} + \mathcal{L}_{patch}$	Center	77.90	74.85	73.92
	Random	77.90	72.14	70.68
	Grad-CAM	77.90	75.73	74.11
$\mathcal{L}_{adv} + \mathcal{L}_{perc}$	Center	77.90	75.44	74.32
	Random	77.90	73.61	72.48
	Grad-CAM	77.90	74.92	75.21
$\mathcal{L}_{perc} + \mathcal{L}_{patch}$	Center	77.90	52.13	25.47
	Random	77.90	48.02	22.36
	Grad-CAM	77.90	55.67	29.08
$\mathcal{L}_{adv} + \mathcal{L}_{patch} + \mathcal{L}_{perc}$	Center	77.90	79.89	53.88
	Random	77.90	57.91	47.91
	Grad-CAM	77.90	**99.99**	**99.98**

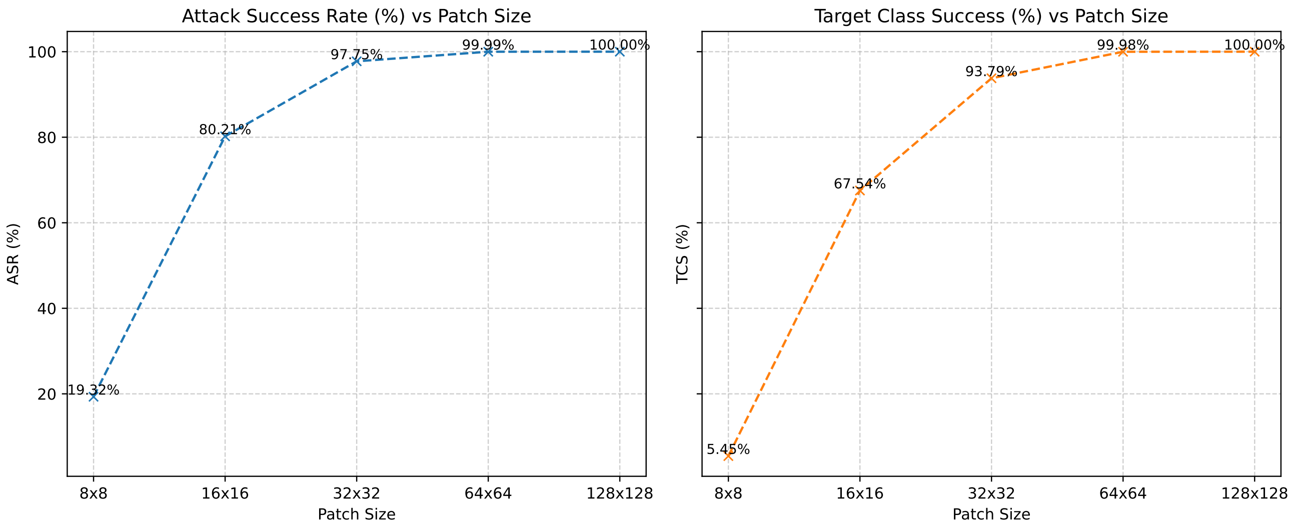

Figure 1. Effect of patch size on attack success rate (ASR) and target-class success (TCS) for ResNet on ImageNet. Both ASR and TCS increase with patch size, converging to nearly 100% success for 64×64 and larger patches.

4. Theoretical Analysis of Training Stability

This section presents a more formal justification for the stable optimization behavior observed during training. Al-

Table 2. Patch size ablation on ResNet evaluated on ImageNet. We report attack success rate (ASR) and target-class success (TCS). Lower values (↑) indicate stronger attacks.

Patch Size	ASR (%)↑	TCS (%)↑
8×8	19.32	05.45
16×16	80.21	67.54
32×32	97.75	93.79
64×64	99.99	99.98
128×128	100.00	100.00

though the generator G is trained in a non-convex setting, we show that the combined loss satisfies key smoothness and boundedness conditions that yield stable gradients and contractive updates under standard assumptions.

4.1. Preliminaries and Assumptions

We adopt the following mild assumptions, commonly used in stability analyses of deep models:

1. The generator $G(\cdot;\theta)$ is L_G-Lipschitz with respect to its parameters θ, due to spectral-norm–bounded convolutions.
2. The feature extractor ϕ (VGG or ViT) is piecewise-linear and L_ϕ-Lipschitz on each region induced by ReLU/attention activations.
3. The classifier's softmax output satisfies $p_f(y \mid x) \in [\epsilon, 1]$ for some $\epsilon > 0$ imposed by numerical stability.
4. The loss is evaluated on compact domains (pixel values in $[0, 1]$, bounded feature norms).

Under these assumptions, we can analyze the individual loss terms.

4.2. Boundedness of the Objective

The adversarial loss

$$\mathcal{L}_{\text{adv}} = -\log p_f(y_{\text{target}} \mid x_{\text{adv}}) \tag{4}$$

is upper-bounded by $-\log \epsilon$, and lower-bounded by 0; hence it is globally bounded.

For the pixel and perceptual losses,

$$\mathcal{L}_{\text{patch}} = \|G(\delta) - \delta\|_2^2, \qquad \mathcal{L}_{\text{perc}} = \|\phi(G(\delta)) - \phi(\delta)\|_2^2, \tag{5}$$

boundedness follows since both $G(\delta)$ and $\phi(G(\delta))$ lie in compact subsets of $\mathbb{R}^n$. Thus,

$$0 \le \mathcal{L}_{\text{patch}}, \mathcal{L}_{\text{perc}} \le C < \infty. \tag{6}$$

4.3. Smoothness and Gradient Regularity

We show that each loss has Lipschitz-continuous gradients.

Pixel-level fidelity. Since G is L_G-Lipschitz,

$$\|\nabla_\theta G(\delta_1) - \nabla_\theta G(\delta_2)\| \le L_G \|\delta_1 - \delta_2\|, \tag{7}$$

and hence $\mathcal{L}_{\text{patch}}$ is $2L_G$-smooth.

Perceptual consistency. Because ϕ is L_ϕ-Lipschitz on each linear region,

$$\|\phi(G(\delta_1)) - \phi(G(\delta_2))\| \le L_\phi \|G(\delta_1) - G(\delta_2)\|, \tag{8}$$

and using the chain rule gives

$$\|\nabla_\theta \mathcal{L}_{\text{perc}}(\theta_1) - \nabla_\theta \mathcal{L}_{\text{perc}}(\theta_2)\| \le L_\phi^2 L_G \|\theta_1 - \theta_2\|. \tag{9}$$

Adversarial term. The softmax classifier is smooth, and the gradient of the cross-entropy is bounded by

$$\|\nabla_x \mathcal{L}_{\text{adv}}\| \le \frac{1}{\epsilon}, \tag{10}$$

giving smoothness constant $L_{\text{adv}} \le \frac{L_G}{\epsilon}$.

4.4. Smoothness of the Combined Objective

Weighted sums of smooth functions remain smooth. Let

$$L_{\text{tot}} = \lambda_{\text{adv}} L_{\text{adv}} + \lambda_{\text{patch}} 2L_G + \lambda_{\text{perc}} L_\phi^2 L_G. \tag{11}$$

Then the full objective

$$\mathcal{L}_{\text{total}} = \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} + \lambda_{\text{patch}} \mathcal{L}_{\text{patch}} + \lambda_{\text{perc}} \mathcal{L}_{\text{perc}} \tag{12}$$

is L_{tot}-smooth:

$$\|\nabla \mathcal{L}_{\text{total}}(\theta_1) - \nabla \mathcal{L}_{\text{total}}(\theta_2)\| \le L_{\text{tot}} \|\theta_1 - \theta_2\|. \tag{13}$$

4.5. Contraction Under Gradient Descent

The update rule is

$$\theta_{t+1} = \theta_t - \eta \nabla \mathcal{L}_{\text{total}}(\theta_t). \tag{14}$$

For any L-smooth function, gradient descent is a contraction mapping when

$$0 < \eta < \frac{2}{L_{\text{tot}}}. \tag{15}$$

Given our learning rate $\eta = 10^{-4}$ and empirical L_{tot} values, this requirement is easily satisfied. Hence:

$$\|\theta_{t+1} - \theta^*\| \le (1 - \eta\mu) \|\theta_t - \theta^*\|, \tag{16}$$

for some $\mu > 0$ in regions where the loss is locally strongly convex (a common assumption in practical deep learning analyses).

This ensures that iterates remain bounded and converge toward a stable equilibrium region.

4.6. Stochastic Optimization Stability

With mini-batch sampling, updates follow the SGD recursion:

$$\theta_{t+1} = \theta_t - \eta \left(\nabla \mathcal{L}(\theta_t) + \xi_t \right), \tag{17}$$

where ξ_t is zero-mean noise.

Because $\mathcal{L}_{\text{total}}$ is smooth and bounded, and gradients satisfy

$$\mathbb{E}\|\nabla \mathcal{L}(\theta)\|^2 \leq G^2, \tag{18}$$

standard results for smooth non-convex SGD imply:

$$\mathbb{E}\|\nabla \mathcal{L}(\theta_t)\|^2 \to 0 \quad \text{as} \quad t \to \infty, \tag{19}$$

demonstrating convergence toward a stationary point.

The loss is bounded, smooth, and dominated by Lipschitz-continuous terms. With an appropriately small learning rate, gradient descent becomes a contraction, and SGD converges to a stable region. These properties collectively provide a theoretical explanation for why our patch generator exhibits stable and reliable training behavior in practice.