TAIGen: Training-Free Adversarial Image Generation via Diffusion Models

Supplementary Material

7. Experimental Setting

In addition to the details presented in the main paper, the additional experimental settings are detailed below.

7.1. Hyperparameter Details:

The values of Ω and Φ have been set to 0.9. Additionally, we used linear beta time scheduling (β_t) with the standard DDPM [22] configurations in all the cases.

ImageNet: We used the Guided Diffusion model [15] from OpenAI on this dataset. The batch size was set to 4. We used 512 images, each of size 224×224. The values of t_{start} and t_{end} here are 80 and 60 respectively. We also kept $\epsilon = 4/255$ when comparing against PGD, AutoAttack, and AdaMSI-FGM. While testing against ACA, we used the MobileNet-V2 as the source and target model and 1000 images were chosen at random. The reported time is in seconds/image. The momentum factor was set to 1.2, and I was fixed at 20 iterations in both of these experiments. We perform a 10-class classification and randomly choose a class from the top-5 classes next in line to the true label class for creating the Grad-CAM.

CelebA-HQ: Whilst testing against BPDA+EOT, we kept $\epsilon = 8/255$, $t_{start} = 20$ and $t_{end} = 18$ and batch size as 4. Since there are only two classes in this case, we choose the class that does not correspond to the true label in this case for creating the Grad-CAM.

CIFAR-10: We set batch size as 8, $I = 50$, 512 as the number of images, $t_{start} = 80$ and $t_{end} = 55$. We also fixed $\epsilon = 4/255$ when comparing the robust accuracy against SPSA, Square Attack, Joint Attack (Full), Diff-BPDA and Auto-Attack. The momentum factor is 1.8. We perform a 1000-class classification and choose a class randomly from the top-500 classes, which are next in line to the true label class for creating the Grad-CAM.

7.2. Additional Experimental Results:

To compare the image quality against the two variations (that is, with and without early stopping) of our algorithm, we randomly choose 512 images from the validation set of CIFAR-10 and report the PSNR and SSIM values in Table 7. We also showcase our algorithm with early stopping in Algorithm 2. We observe that image quality is retained more with early stopping. However, one limitation of using this is that the images become more susceptible to purification algorithms like Diffpure [40]. This is because, with early stopping, the images are not as robust as those that have undergone more iterations of adding adversarial noise. Thus, they are easier to purify. Since we present a black-box attack, we use the WideResNet-50-2 classifier [43] as the

	PSNR	SSIM
Clean	24.16	81.55
w/o early stopping	22.42	75.39
w early stopping	**23.28**	**77.12**

Table 7. PSNR and SSIM values on the CIFAR-10 dataset. The model used is the standard WideResNet-28-10 with $\epsilon = 4/255$.

Algorithm 2 Adversarial Image Generation with Early Stopping

function TAIGEN(Input Image x_0, Noise Schedule $\beta_{1:T}$, Target Classifier f, Grad-CAM g_{CAM}, Diffusion U-Net Model ϵ_θ, Ground Truth y, Adversarial Iterations I, Momentum Factor μ, Step Size α, Cross-Entropy loss J)
  Ensure: Adversarial Image x_{adv}
  Initialize: $x_T \leftarrow x_0 \cdot \sqrt{\bar{\alpha}_T} + \sqrt{1-\bar{\alpha}_T} \cdot \epsilon, \epsilon \sim \mathcal{N}(0, I)$
  $\hat{x}_t^{adv} \leftarrow \hat{x}_{t_{start}}$, $\sigma_t^2 \leftarrow \beta_t$, $z \sim \mathcal{N}(0, I)$, $N = t_{start} - t_{end}$
  for $t \in [T, T-1, \ldots, 1]$ **do**
    $\epsilon_t, w_{t-1} \leftarrow \epsilon_\theta(\hat{x}_t, t)$
    $\hat{x}_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_t \right) + \sigma_t z$
    $x_0 \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}} (x_t - \sqrt{1-\bar{\alpha}_t} \epsilon_t)$
    if $\arg\max f(x_0) \neq y$ **then**
      $\hat{x}_0 = x_0$
      break
    else if $(t \leq t_{start})$ & $(t \geq t_{end})$ **then**
      $\epsilon_t \leftarrow \epsilon_\theta(\hat{x}_t^{adv}, t)$
      $z_0 \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_t \right) + \sigma_t z$
      $G_{t-1} \leftarrow 1(g_{CAM} > \Omega)$
      $W_{t-1} \leftarrow 1(w_{t-1} > \Phi)$
      $C_{t-1} \leftarrow (W_{t-1} \oplus G_{t-1} \oplus G_{t-1})$
      for $i \in [0, 1, 2, \ldots I-1]$ **do**
        $g_{i+1} = \mu g_i + \frac{\nabla_z J(z_i, y)}{\|\nabla_z J(z_i, y)\|_1}$
        $z_{i+1} = z_i + \alpha \cdot \text{Sign}(g_{i+1})$
      end for
      return z_I
    end if
    $\hat{x}_{t-1}^{adv} \leftarrow C_{t-1} \odot \hat{x}_{t-1}^{adv} + (1 - C_{t-1}) \odot z_I$
    if $t == t_{end}$ **then**
      $\hat{x}_t = \hat{x}_t^{adv}$
    end if
  end for
  return $\hat{x}_0$
end function

Model	Clean Accuracy (%)	ASR (%)
R50 [20]	76.52	96.10
R50 [43]	64.02	57.23
WRN-50-2 [43]	68.46	54.11
R50 [17]	62.56	57.43
ViT-B [37]	68.38	49.22

Table 8. Accuracy and ASR on the ImageNet dataset with $\epsilon = 4/255$ using the WideResNet-50-2 as the source model.

source model and test the efficacy on various adversarially trained target models, ResNet-50 [17], ResNet-50[43] and WideResNet-50-2 [43], ViT-B [37] and a standard ResNet-50 model [20]. The values for this experiment are reported in Table 8.

8. Finding Mixing Step

[60] has explained in detail about the mixing step. We give a brief explanation of the same. We first define the radius of a high-dimensional Gaussian space. Mathematically, it can defined as $r = \sigma\sqrt{d}$. Now let us take a point in this vector space, $x = (x_1, x_2, \ldots, x_d)$, chosen at random from Gaussian, the square root of the expected square length of x is formulated in Equation 8.

$$\sqrt{\mathrm{E}(x_1^2 + x_2^2 + \cdots + x_d^2)} = \sqrt{d\mathrm{E}(x_1^2)} = \sqrt{d}\sigma \quad (8)$$

Equation 8 is used to find the radius of our sampled latent variables at each time step. We also define the total variation distance which will be used in the further proofs in Equation 9.

$$||\mu - \tau||_{TV} = \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu(x) - \tau(x)| \quad (9)$$

Here μ and τ are two probability distributions on $\mathcal{X}$. Next, we define the quantity $\sigma_t(x, y)$ for an irreducible transition matrix P with stationary distribution π in Equation 10 and $d^{(p)}$ distance in Equation 11.

$$\sigma_t(x, y) = \frac{P^t(x, y)}{\pi(y)} \quad (10)$$

$$d^p(t) := \max_{x\in\mathrm{X}} ||\sigma_t(x, .) - 1||_p \quad (11)$$

Replacing the above notations with the ones from a standard DDPM model, we get Equations 12 and 13.

$$d^1(t) := \max_{x\in\mathrm{X}} ||\sigma_t(x, .) - 1||_1 \quad (12)$$

and,

$$\sigma_t(x, y) = \frac{P^t(x, y)}{\pi(y)} = \frac{x \sim \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathrm{I})}{y \sim \mathcal{N}(0, \mathrm{I}_d)} \quad (13)$$

Mathematically, we can define the mixing time via Equation 14

$$t^{(1)}_{mix}(\epsilon) := \inf\{t \geq 0; d^1(t) \leq \epsilon\} \quad (14)$$

Taking $\epsilon = 1/2$, we get Equation 15

$$t^{(1)}_{mix}(\epsilon) := \inf\{t \geq 0; d^1(t) \leq \frac{1}{2}\} \quad (15)$$

Replacing Equation 15 with Equation 13, we get Equation 16.

$$\max_{x\in\mathcal{X}} || \frac{x \sim \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathrm{I})}{y \sim \mathcal{N}(0, \mathrm{I}_d)} - 1|| \leq \frac{1}{2} \quad (16)$$

Using Equation 9, we can substitute in Equation 16 which gives us the approximation in Equation 17.

$$||x \sim \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathrm{I})|| \leq 4 \quad (17)$$

Equation 17 searches for the mixing step where the Gaussian radius changes by an amount of **4** units.

9. Solid Angle using Quaternions

To find the similarity between the latent variable x_t during the forward process and $\hat{x}_t$ during the backward process, we use the cosine similarity between these two variables. Specifically, we consider the vector or quaternion q_1 and q_2 and the cosine similarity is found using Equation 18.

$$\Omega = \cos^{-1}\left(\frac{q_1}{||q_1||_2} \cdot \frac{q_2}{||q_2||_2}\right) \quad (18)$$

As observed from the graph (Figure 3 of the main paper), it forms a concave curve. Intuitively, we can say that while the Gaussian distribution doesn't converge with the stationary distribution, the solid angle or dissimilarity increases. At a certain point, in this case the maxima, they converge and then the similarity increases between the two vectors. Thus, via this analysis, it is evident that the inversion and the sampling processes are **not** symmetric. Now our intuition lies in utilizing this fact and perturbing the $\hat{x}_t$ around the time step corresponding to the maxima. If instead, we perturbed at any other time step, the dissimilarity would again increase till it reached a maximum (which would not be the mixing step) and then decrease. In the latter case, we empirically found the artifacts are visible in the reconstructed image.