Attacking Perceptual Similarity Metrics

Anonymous authors
Paper under double-blind review

Abstract

Perceptual similarity metrics have progressively become more correlated with human judgments on perceptual similarity; however, despite recent advances, the addition of an imperceptible distortion can still compromise these metrics. In our study, we systematically examine the robustness of these metrics to imperceptible adversarial perturbations. Following the two-alternative forced-choice experimental design with two distorted images and one reference image, we perturb the distorted image closer to the reference via an adversarial attack until the metric flips its judgment. We first show that all metrics in our study are susceptible to perturbations generated via common adversarial attacks such as FGSM, PGD, and the One-pixel attack. Next, we attack the widely adopted LPIPS metric using spatial-transformation-based adversarial perturbations (stAdv) in a white-box setting to craft adversarial examples that can effectively transfer to other similarity metrics in a black-box setting. We also combine the spatial attack stAdv with PGD (ℓ_∞-bounded) attack to increase transferability and use these adversarial examples to benchmark the robustness of both traditional and recently developed metrics. Our benchmark provides a good starting point for discussion and further research on the robustness of metrics to imperceptible adversarial perturbations.

1 Introduction

Comparison of images using a similarity measure is crucial for defining the quality of an image for many applications in image and video processing. Recently, perceptual similarity metrics have become vital for optimizing and evaluating deep neural networks used in low-level computer vision tasks (Dosovitskiy & Brox, 2016; Zhu et al., 2016; Johnson et al., 2016; Ledig et al., 2016; Sajjadi et al., 2017; Kettunen et al., 2019a; Zhang et al., 2020; Son et al., 2020; Niklaus & Liu, 2020; Karras et al., 2020). Learned perceptual image patch similarity (LPIPS) metric by Zhang et al. (2018b) is one such widely adopted perceptual similarity metric. Apart from these image enhancement and generation tasks, similarity metrics are also used in optimizing, constraining, and evaluating adversarial attacks (Szegedy et al., 2014; Goodfellow et al., 2015; Carlini & Wagner, 2017; Kurakin et al., 2017; Hosseini & Poovendran, 2018; Dong et al., 2018; Shamsabadi et al., 2020; Laidlaw & Feizi, 2019). More recently, Laidlaw et al. (2020) employed LPIPS to optimize adversarial examples, introducing adversarial attacks based on a neural perceptual threat model, and subsequently a defense method that could generalize well against unforeseen adversarial attacks. However, it remains unanswered whether LPIPS itself is robust towards imperceptible adversarial perturbations. The question then arises, *"How robust are perceptual similarity metrics against imperceptible adversarial perturbations?"*

We begin by examining whether it is possible to find imperceptible adversarial perturbations that can overturn perceptual similarity judgments. It is well known that machine learning models are easy to fool with adversarial perturbations imperceptible to the human eye (Szegedy et al., 2014). Interestingly, similar imperceptible perturbations can bring about a sizeable change in the measured distance of a distorted image from its reference. As shown in Figure 1, we examine this change in measured distances using a two-alternative forced choice (2AFC) test example, where the participants were asked, "which of the two distorted images (I_0 and I_1) is more similar to the reference image (I_{ref})". Then, we apply an imperceptible perturbation to the distorted image that has the lower perceptual distance (i.e., more similar to I_{ref}) to see if the similarity judgment for the sample overturns. In such a scenario, human opinion remains the same, while perceptual

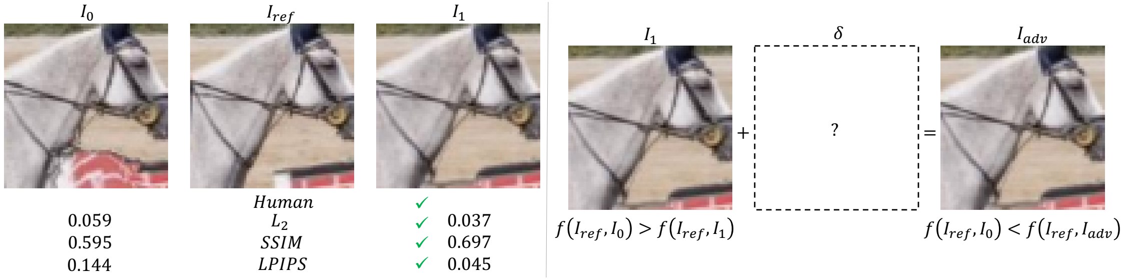


Figure 1: I_1 is more similar to I_{ref} than I_0 according to all perceptual similarity metrics and humans. We attack I_1 by adding imperceptible adversarial perturbations (δ) such that the metric (f) flips its earlier assigned rank, i.e., in the above sample, I_0 becomes more similar to I_{ref}.

similarity metrics often overturn their judgment. Interestingly, recent work began to investigate the perceptual robustness of image quality assessment methods via adversarial perturbations Zhang et al. (2022) and Lu et al. (2022). However, these studies focus on no-reference image quality assessment methods. The robustness of perceptual similarity metrics, often used as full-reference image quality assessment methods, has not been studied.

There are two approaches to examining the robustness of perceptual similarity metrics: (1) addition of small amounts of hand-crafted geometric distortions such as translation, rotation, dilation, JPEG compression, and gaussian blur, and (2) analysis of more advanced adversarial perturbations. For the former, seminal contributions have been made (Ma et al., 2018; Ding et al., 2020; Bhardwaj et al., 2020; Gu et al., 2020). However, in contrast to previous work, we focus on performing the latter as it has not received considerable attention. In our work, we demonstrate that threats to similarity metrics can be easily created using common gradient-based iterative white-box attacks, such as fast gradient sign method (FGSM) (Goodfellow et al., 2015) and projected gradient descent (PGD) (Madry et al., 2018). These attacks do not deform the structure but rather manipulate pixel values in the image. However, in recent research, questions regarding the robustness of perceptual similarity metrics towards geometric distortions are of central interest (as discussed above). Hence, we also use the spatial adversarial attack stAdv (Xiao et al., 2018), which geometrically deforms the image. It utilizes optical flow for crafting perturbations in the spatial domain. We use this attack to generate adversarial samples for comparing the robustness of various metrics.

We also examine whether perceptual metrics can be attacked in black box settings. To this end, we first use the One-pixel attack (Su et al., 2019) that uses differential evolution (Storn & Price, 1997) to optimize a single-pixel perturbation on the adversarial image. While compared to white box attacks such as FGSM and PGD, this One-pixel attack does not need the model parameters of a similarity metric, it needs to access its output. Therefore, we furthermore explore transferable attacks (Liu et al., 2017; Xie et al., 2018; 2019) which requires no information about the model. Specifically, we generate adversarial examples using the parameters of a source model and use them to attack a target model. In our study, we use LPIPS(AlexNet) as the source model and attack it via stAdv. We extend the successfully attacked examples onto a target perceptual similarity metric. It is a black-box setting as it does not require access to the target perceptual metric's parameters. In our work, we combine stAdv (spatial attack) with PGD (ℓ_∞-bounded attack) that strengthens the severity of the adversarial examples.

The main contribution of this paper is the first systematical investigation on whether existing perceptual similarity metrics are susceptible to state-of-the-art adversarial attacks. Our study includes a set of carefully selected attacking methods and a wide variety of perceptual similarity metrics. Our study shows that all these similarity metrics, including both traditional quality metrics and recently proposed deep learning-based metrics, can be successfully attacked by both white-box and black-box attacks.

2 Related Work

Earlier metrics such as SSIM (Wang et al., 2004) and FSIMc (Zhang et al., 2011) were designed to approximate the human visual systems' ability to perceive and distinguish images, specifically using statistical features of local regions in the images. Whereas, recent metrics (Zhang et al., 2018b; Prashnani et al., 2018; Ma et al., 2018; Kettunen et al., 2019b; Ding et al., 2020; Bhardwaj et al., 2020; Ghildyal & Liu, 2022) are deep neural network based approaches that learn from human judgments on perceptual similarity. LPIPS (Zhang et al., 2018b) is one such widely used metric. It leverages the activations of a feature extraction network at each convolutional layer to compute differences between two images which are then passed on to linear layers to finally predict the perceptual similarity score.

In recent years, apart from making the perceptual similarity metrics correlate well with human opinion, there has been growing interest in examining their robustness towards geometric distortions. Ma et al. (2018) benchmarked the sensitivity of various metrics against misalignment, scaling artifacts, blurring, and JPEG compression. They then trained a CNN with augmented images to create the geometric transformation invariant metric (GTI-CNN). In a similar study, Ding et al. (2020) suggested computing global measures instead of pixel-wise differences and then blurred the feature embeddings by replacing the max pooling layers with l_2-pooling layers. It made their metric, deep image structure and texture similarity (DISTS), robust to local and global distortions. Bhardwaj et al. (2020) developed the perceptual information metric (PIM). PIM has a pyramid architecture with convolutional layers that generate multi-scale representations, which get processed by dense layers to predict mean vectors for each spatial location and scale. The final score is estimated using symmetrized KL divergence using Monte Carlo sampling. PIM is well correlated with human opinions and is robust against small image shifts, even though it is just trained on consecutive frames of a video, without any human judgments on perceptual similarity. Czolbe et al. (2020) used Watson's perceptual model (Watson, 1993) and replaced discrete cosine transform with discrete fourier transform (DFT) to develop a perceptual similarity loss function robust against small shifts. Kettunen et al. (2019b) compute the average LPIPS score over an ensemble of randomly transformed images. Their self-ensembling metric E-LPIPS is robust to the Expectations over Transformations attacks (Athalye et al., 2018; Carlini & Wagner, 2017). Our attack approach is similar to an attack investigated by Kettunen et al. (2019b), where the adversarial images look similar but have a large LPIPS distance (smaller distance means more similarity). However, they only investigate the LPIPS metric. So far, the majority of prior research has focused on geometric distortions, while no study has systematically investigated the robustness of various similarity metrics to more advanced adversarial perturbations that are more perceptually indistinguishable. We seek to address this critical open question, *whether perceptual similarity metrics are robust against imperceptible adversarial perturbations*. In our paper, we show that the metrics often overturn their similarity judgment after the addition of adversarial perturbations, unlike humans, to whom the perturbations are unnoticeable.

There exists a considerable body of literature on adversarial attacks (Szegedy et al., 2014; Goodfellow et al., 2015; Liu et al., 2017; Papernot et al., 2016; Carlini & Wagner, 2017; Xie et al., 2018; Hosseini & Poovendran, 2018; Madry et al., 2018; Xiao et al., 2018; Brendel et al., 2018; Song et al., 2018; Zhang et al., 2018a; Engstrom et al., 2019; Laidlaw & Feizi, 2019; Su et al., 2019; Wong et al., 2019; Bhattad et al., 2019; Xie et al., 2019; Zeng et al., 2019; Dolatabadi et al., 2020; Tramèr et al., 2020; Laidlaw et al., 2020; Croce et al., 2020; Wu & Zhu, 2020), but none of the previous investigations have ever considered attacking perceptual similarity metrics. This paper builds upon this line of research and carefully selects a set of representative attacking algorithms to investigate the adversarial robustness of similarity metrics. We briefly describe these methods and how we employ them to attack similarity metrics in Section 3. In parallel, Lu et al. (2022) developed an adversarial attack for neural image assessment (NIMA) (Talebi & Milanfar, 2018) to prevent misuse of high-quality images on the internet. NIMA is NR-IQA with scores $\in [1, 10]$, while we systematically investigate several FR-IQA methods, with scores $\in [0, \infty)$, against various attacks.

Recent work underlines the importance of perceptual distance as a bound for adversarial attacks (Laidlaw et al., 2020; Wang et al., 2021; Zhang et al., 2022). Laidlaw et al. (2020) developed a neural perceptual threat model (NPTM) that employs the perceptual similarity metric LPIPS(AlexNet) as a bound for generating adversarial examples and provided evidence that l_p-bounded and spatial attacks are near subsets of the NPTM. Similarly, Zhang et al. (2022) developed a perceptual threat model to attack no-reference image quality assessment (IQA) methods by constraining the perturbations via full-reference IQA, i.e., perceptual similarity metrics such as SSIM, LPIPS, and DISTS. They posit that the metrics are "approximations to human perception of just-noticeable differences" (Zhang et al., 2022), therefore, can keep perturbations imperceptible. Moreover, Laidlaw et al. (2020) found LPIPS to correlate well with human opinion when evaluating adversarial examples. *However, it has not yet been established whether LPIPS and other perceptual similarity metrics are adversarially robust.* We investigate this in our work, and the findings in our study indicate that all metrics, including LPIPS, are not robust to various kinds of adversarial perturbations.

3 Method

Dataset. Our study uses the Berkeley-Adobe perceptual patch similarity (BAPPS) dataset, originally used to train a perceptual similarity metric (Zhang et al., 2018b). Each sample in this dataset contains a set of 3 images: 2 distorted (I_0 and I_1) and 1 reference (I_{ref}). For perceptual similarity assessment, the scores were generated using a two-alternative forced choice test where the participants were asked, "which of two distortions is more similar to a reference" (Zhang et al., 2018b). For the validation set, 5 responses per sample were collected. The final human judgment is the average of the responses. The types of distortions in this dataset are traditional, CNN-based, and distortions by real algorithms such as super resolution, frame interpolation, deblurring, and colorization. Human opinions could be divided, i.e., all responses in a sample may not have voted for the same distorted image. In our study, to ensure that the two distorted images in the sample have enough disparity between them, we only select those samples where humans unanimously voted for one of the distorted images. In total, there are 12,227 such samples.

Attack Models. As observed in Figure 1, the addition of adversarial perturbations can lead to a rank flip. We make use of existing attack methods such as FGSM (Goodfellow et al., 2015), PGD (Madry et al., 2018), One-pixel attack (Su et al., 2019), and spatial attack stAdv (Xiao et al., 2018) to generate such adversarial samples. These attack methods were originally devised to fool image classification models, therefore, we introduce minor modifications in their procedures to attack perceptual similarity metrics.

We select one of the distorted images, I_0 or I_1, that is more similar to I_{ref} to attack. The distorted image being attacked is I_{prey}, and the other image is I_{other}; accordingly, for the sample in Figure 1, I_1 is I_{prey} and I_0 is I_{other}. Consider s_i as the similarity score between I_i and I_{ref}[1]. Before the attack, the original rank is $s_{other} > s_{prey}$, but after the attack I_{prey} turns into I_{adv}, and when the rank flips, $s_{adv} > s_{other}$. In image classification, a misclassification is used to measure the attack's success, while for perceptual similarity metrics, an attack is successful when the rank flips.

Fast Gradient Sign Method. FGSM is a popular white-box attack introduced by Goodfellow et al. (2015). This attack method projects the input image I onto the boundary of an ϵ sized ℓ_∞-ball, and therefore, restricts the perturbations to the locality of I. We follow this method to generate imperceptible perturbations by constraining ϵ to be small for our experiments. This attack starts by first computing the gradient with respect to the loss function of the image classifier being attacked. The signed value of this gradient multiplied by ϵ generates the perturbation, and thus, $I_{adv} := I + \epsilon \cdot sign(\nabla_I J(\theta, I, target))$, where θ are the model parameters. We adopt this method to attack perceptual similarity metrics. We formulate a new loss function for an untargeted attack as:

$$J(\theta, I_{prey}, I_{other}, I_{ref}) = \left(\frac{s_{other}}{s_{other} + s_{prey}} - 1\right)^2 \quad (1)$$

We maximize this loss, i.e., move in the opposite direction of the optimization by adding the perturbation to the image. The human score of all the samples in our selected dataset is either 0 or 1, unanimous vote. Hence, we can easily employ the loss function in Equation 1, because if the metric predicts the rank correctly then $(s_{other}/(s_{other} + s_{prey}))$ would be ≈ 1. Afterwards, if the attack is successful then $(s_{other}/(s_{other} + s_{adv}))$ becomes less than 0.5, causing the rank to flip.

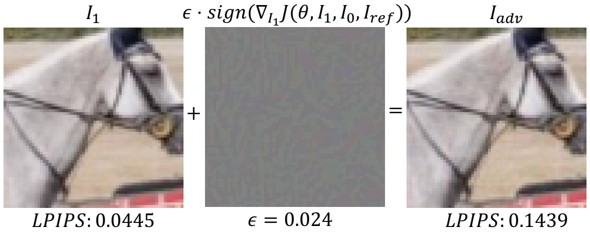

I_1
$\epsilon \cdot sign(\nabla_{I_1} J(\theta, I_1, I_0, I_{ref}))$
I_{adv}
+
=
$LPIPS$: 0.0445
$\epsilon = 0.024$
$LPIPS$: 0.1439

Figure 2: FGSM attack on LPIPS(AlexNet). In this white-box attack, we use the LPIPS network parameters to compute the signed gradient. With increase in ϵ, the severity of the attack increases. In this example, the adversarial perturbations are hardly visible. The RMSE between the prey image I_1 and the adversarial image I_{adv} is 3.53.

---

[1] smaller s_i means I_i is more similar to I_{ref}

Algorithm 3 (refer Appendix A) provides the details for the FGSM attack. First, I_{prey} is selected based on the original rank. The model parameters remain constant, and we compute the gradients with respect to the input image I_{prey}. To increase perturbations in normalized images, we increase the ϵ in steps of 0.0001 starting from 0.0001. When ϵ is large enough, the rank flips. It would mean that the attack was successful (see Figure 2 for example). If the final value of ϵ is small then the perturbation is imperceptible, making it hard to discern any difference between the original input image and its adversarial sample.

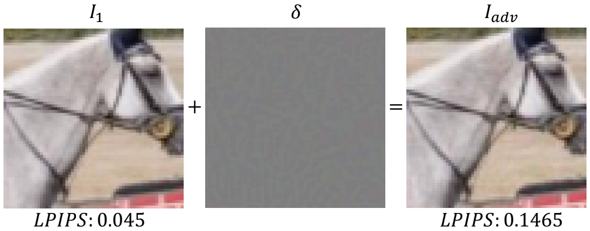


Figure 3: PGD attack on LPIPS(AlexNet). In this white-box attack, we use the LPIPS network parameters to compute the signed gradient. Increasing the attack iterations, increases the severity of the attack. In this example, perturbations in I_{adv} are not visible. The RMSE between the prey image I_1 and the adversarial image I_{adv} is 2.10.

Projected Gradient Descent. PGD attack by Madry et al. (2018) takes a similar approach to FGSM, but instead of a single large step like in FGSM, it takes multiple small steps for generating perturbation δ. Hence, the projection of I stays either inside or on the boundary of the ϵ-ball. Each time $\delta > \epsilon$, the projection operator under ℓ_∞ constraint, restricts the pixel values to a predefined range $[-\epsilon, +\epsilon]$. This multistep attack is defined as:

$$I_{adv}^{t+1} = P_c\big(I_{adv}^{t} + \alpha \cdot sign(\nabla_{I_{adv}^{t}} J(\theta, I_{adv}^{t}, I_{other}, I_{ref})\big) \tag{2}$$

where J is the loss defined in Equation 1. The perturbation on each pixel is bounded to a predefined range using the projection constraint P_c. We implement P_c using a clip operation on the final perturbation δ (Line 15 Algorithm 1). As shown in Algorithm 1, the signed gradient is multiplied with step size α, and this adversarial perturbation is added to I_{adv}^t. The final perturbation δ is the difference between I_{adv}^t and I_{prey}, and in our method, δ is bounded by l_∞ norm. Hence, the PGD attack is an **ℓ_∞-bounded attack**.

Algorithm 1: PGD attack on Similarity Metrics

```
Input: $I_0$, $I_1$, $I_{ref}$, metric $f$, $\epsilon$ (perturbation limit 0.1),
       $max\_iterations$ (40), $\alpha$ (step size 0.001)
Output: $attack\_success$ True on rank flip
1  $s_0 = f(I_{ref}, I_0)$; $s_1 = f(I_{ref}, I_1)$; $rank = int(s_0 > s_1)$;
2  // If $I_0$ is more similar to $I_{ref}$ then $rank$ is 0 else 1
3  if $rank = 1$ then $I_{prey} = I_1$; $s_{other} = s_0$;
4  else $I_{prey} = I_0$; $s_{other} = s_1$;
5  $\delta = zeros\_like(I_{prey})$ // perturbation
6  $k = 0$
7  while $k \le max\_iterations$ do
8      $I_{adv} = clip(I_{prey} + \delta, min = -1, max = 1)$
9      $s_{adv} = f(I_{ref}, I_{adv})$
10     if $s_{adv} > s_{other}$ then return $True$ // Attack successful
11     $J = \big((s_{other}/(s_{other} + s_{adv})) - 1\big)^2$ // Loss
12     $signed\_grad = sign\big(\nabla_{I_{adv}} J\big)$
13     $I'_{adv} = I_{adv} + \alpha * signed\_grad$
14     $\delta = clip(I'_{adv} - I_{prey}, min = -\epsilon, max = +\epsilon)$
15     $k = k + 1$
16 return $False$ // Attack unsuccessful
```

One-pixel Attack. The previous two approaches are white-box attacks. We now use a black-box attack, the One-pixel attack by Su et al. (2019) that perturbs only a single pixel using differential evolution (Storn & Price, 1997). The objective of the One-pixel attack is defined as:

$$\begin{aligned} \underset{e(I_{prey})^*}{\text{maximize}} \quad & f(I_{prey} + e(I_{prey}), I_{ref}) \\ \text{subject to} \quad & ||e(I_{prey})||_0 \le d \end{aligned} \tag{3}$$

where f is the similarity metric, and the vector $e(I_{prey})$ is the additive adversarial perturbation, and d is 1 for the One-pixel attack. This algorithm aims to find a mutation to one particular pixel such that a similarity metric f, such as LPIPS, will consider I_{prey} is less similar to I_{ref} than it is originally, and thus, the rank is flipped. Note, for LPIPS, a larger score indicates the two images being less similar. Please refer to Su et al. (2019) for more details of this attack algorithm.

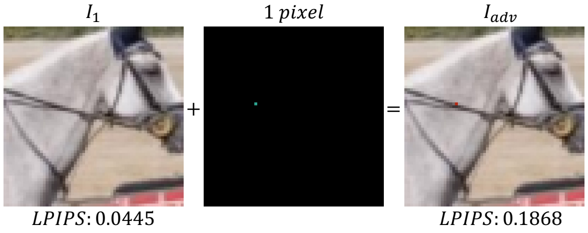


Figure 4: One-pixel attack on LPIPS(AlexNet). This is a black-box attack as it does not require LPIPS network parameters to generate the adversarial perturbation. The one pixel perturbation is hardly visible. The RMSE between the prey image I_1 and the adversarial image I_{adv} is 1.38.

Spatial Attack (stAdv). The goal of the stAdv attack is to deform the image geometrically by displacing pixels (Xiao et al., 2018). It generates adversarial perturbations in the spatial domain rather than directly manipulating pixel intensity values. This attack synthesizes the spatially distorted adversarial image (I_{adv}) via optimizing a flow vector and backward warping with the input image (I_{prey}) using differentiable bilinear interpolation (Jaderberg et al., 2015). For each sample, we start with a flow initialized with zeros and then optimize it using L-BFGS (Liu & Nocedal, 1989) for the following loss.

$$\mathcal{L} = \alpha\mathcal{L}_{rank} + \beta\mathcal{L}_{flow} \tag{4}$$

$$\mathcal{L}_{flow} = \sum_{p}^{pixels} \sum_{q}^{neighbors(p)} \sqrt{(u_p - u_q)^2 + (v_p - v_q)^2} \tag{5}$$

where (u, v) is the displacement vector at pixel location p and its 4 neighbors q.

$$\mathcal{L}_{rank} = \left(\frac{s_{other}}{s_{other} + s_{adv}}\right)^2 \tag{6}$$

where α is 50 and β is 0.05.

As we minimize $\mathcal{L}_{rank}$, the perturbations in s_{adv} will increase, and thus rank will flip. Simultaneously, we also minimize $\mathcal{L}_{flow}$ which defines the amount of perturbations generated by flow to distort the image. It enforces the perturbations to be constrained to make as little change to the attacked image I_{prey} as possible. Xiao et al. (2018) performed a user study to test the perceptual quality of the images having perturbations generated by the stAdv attack and found them to be indistinguishable by humans. By visual inspection, we found the adversarial perturbations on the images imperceptible in our studies as well.

Algorithm 2: stAdv attack on LPIPS

```
Input: $I_0$, $I_1$, $I_{ref}$, LPIPS $f$, $max\_iterations$ (250)
Output: $attack\_success$ True on rank flip
Function stAdv_attack($flow$, $f$, $I_{prey}$, $I_{ref}$, $s_{other}$):
    $I_{adv}$ = warp($flow$,$I_{prey}$) // Backwarp via bilinear interpolation
    $s_{adv} = f(I_{ref}, I_{adv})$
    $\mathcal{L}_{rank}, \mathcal{L}_{perturb} = calc\_loss(I_{ref}, I_{prey}, I_{adv}, s_{other}, f)$
    $\mathcal{L} = \mathcal{L}_{rank} + \mathcal{L}_{perturb}$
    $gradient = \nabla_{flow}\mathcal{L}$
    if $s_{adv} > s_{other}$ then return $0$, $gradient$, $flow$ // Attack successful
    else return $\mathcal{L}$, $gradient$, $flow$ // Attack unsuccessful
$s_0 = f(I_{ref}, I_0)$; $s_1 = f(I_{ref}, I_1)$;
$rank = int(s_0 > s_1)$ // If $I_0$ is more similar to $I_{ref}$ then $rank$ is 0 else 1
if $rank = 1$ then $I_{prey} = I_1$; $s_{other} = s_0$;
else $I_{prey} = I_0$; $s_{other} = s_1$;
// Initialize a flow vector with zeros
$flow = zeros\_like(2 * I_{prey}$ height $* I_{prey}$ width)
$converge, grad, flow = L\text{-}BFGS$(func=$stAdv\_attack$, args=($flow$, $f$, $I_{prey}$, $I_{ref}$, $s_{other}$), iterations=$max\_iterations$) // Optimize flow vector
if $converge = 0$ then $attack\_success$ = True
else $attack\_success$ = False
return $attack\_success$
```

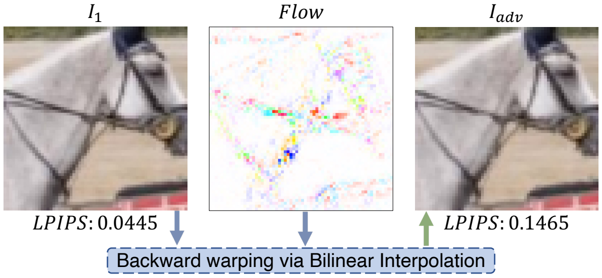


Figure 5: Spatial attack stAdv on LPIPS(AlexNet). We attack LPIPS(AlexNet) to create adversarial images. This attack optimizes a flow vector to create perturbations in the spatial domain. In this example, flow distorts the structure of the horse to generate the adversarial image. The RMSE between the prey image I_1 and the adversarial image I_{adv} is 2.50.

4 Experiments and Results

We experiment with a wide variety of similarity metrics including both traditional ones, such as L2, SSIM (Wang et al., 2004), and FSIMc (Zhang et al., 2011), and the recent deep learning based ones, such as WaDIQaM-FR (Bosse et al., 2018), GTI-CNN (Ma et al., 2018), LPIPS (Zhang et al., 2018b), E-LPIPS (Kettunen et al., 2019b), DISTS (Ding et al., 2020), Watson-DFT (Czolbe et al., 2020), and PIM (Bhardwaj et al., 2020). We adopt the BAPPS validation dataset (Zhang et al., 2018b) for our experiments. Following Zhang et al. (2018b) we scale the image patches from size 256×256 to 64×64. As mentioned in Section 3, we believe that the predicted rank by a metric will be easy to flip on samples close to the decision boundary; therefore, we take a subset of the samples in the dataset which have a clear winner, i.e., all human responses indicated that one was distinctly better than the other. Now, in our dataset, we have 12,227 samples. We report the accuracy of metrics on the subset of selected samples and compare it with their Two-alternative forced choice (2AFC) scores on the complete BAPPS validation dataset. As shown in Table 1, all these

Table 1: Accuracy on the subset selected for our experiments correlates with the 2AFC score computed on the complete BAPPS validation dataset.

Network	2AFC (%) on complete BAPPS (36344 samples)	Accuracy (%) on a subset of BAPPS (12227 samples)
L2	63.2	79.7
SSIM (Wang et al., 2004)	63.1	80.8
WaDIQaM-FR (Bosse et al., 2018)	66.5	83.3
LPIPS(Alex) (Zhang et al., 2018b)	69.8	92.4
LPIPS(VGG) (Zhang et al., 2018b)	68.1	89.8
DISTS (Ding et al., 2020)	68.9	91.3

Table 2: FGSM, PGD, and One-pixel attack results. Larger ϵ allows more perturbations, and lower RMSE relates to higher imperceptibility.

Network	Same Rank by Human & Metric	Total Samples	FGSM ($\epsilon < 0.05$)				PGD						One-pixel
			#Samples Flipped	Mean ϵ	RMSE		#Samples Flipped	% pixels with ϵ			RMSE		#Samples Flipped
					μ	σ		>0.001	>0.01	>0.05	μ	σ	
L2	✓	9750	3759/39%	0.023	2.9	1.7	2348/24%	84.4	56.1	0.0	1.9	1.0	4225/43%
	✗	2477	1550/63%	0.017	2.2	1.6	1202/49%	82.0	42.7	0.0	1.5	1.0	1412/57%
SSIM	✓	9883	6922/70%	0.018	2.5	1.7	5297/54%	94.6	53.6	0.0	1.8	1.0	1787/18%
(Wang et al., 2004)	✗	2344	2013/86%	0.011	1.6	1.3	1843/79%	87.3	32.0	0.0	1.3	0.8	1005/43%
WadIQaM-FR	✓	10191	8841/87%	0.006	1.0	1.0	10176/100%	69.2	4.3	0.0	0.7	0.3	3130/31%
(Bosse et al., 2018)	✗	2036	2012/100%	0.001	0.6	0.3	2035/100%	41.2	0.1	0.0	0.5	0.1	1598/79%
LPIPS(Alex)	✓	11303	7247/64%	0.018	2.4	1.7	8806/78%	86.8	28.7	0.0	1.3	0.6	9255/82%
(Zhang et al., 2018b)	✗	924	912/99%	0.004	0.9	0.7	917/99%	59.5	3.2	0.0	0.8	0.3	921/100%
LPIPS(VGG)	✓	10976	8434/77%	0.012	1.7	1.5	9689/88%	81.6	15.6	0.0	1.0	0.5	7212/66%
(Zhang et al., 2018b)	✗	1251	1244/100%	0.003	0.8	0.5	1246/100%	52.3	1.6	0.0	0.7	0.2	1219/98%
DISTS	✓	11158	3043/27%	0.025	3.3	1.8	2306/21%	97.0	75.4	0.0	2.6	1.3	7416/67%
(Ding et al., 2020)	✗	1069	795/74%	0.016	2.2	1.7	723/68%	91.9	50.0	0.0	2.0	1.3	1033/97%

metrics consistently correlated better with the human opinions on the subset of BAPPS than on the full dataset, which is expected as we removed the ambiguous cases.

In this section, we first show that similarity metrics are susceptible to both white-box and black-box attacks. Based on this premise, we hypothesize that these similarity metrics are vulnerable to transferable attacks. To prove this, we attack the widely adopted LPIPS using the spatial attack stAdv to create adversarial examples and use them to benchmark the adversarial robustness of these similarity metrics. Furthermore, we add a few iterations of the PGD attack, hence combining our spatial attack with ℓ_∞-bounded perturbations, to enhance transferability to other perceptual similarity metrics.

4.1 Adversarial Attack on Perceptual Similarity Metrics

Through the following study, we test our hypothesis that similarity metrics are susceptible to adversarial attacks. We first determine whether it is possible to create imperceptible adversarial perturbations that can overturn the perceptual similarity judgment, i.e., flip the rank of the images in the sample. We try to achieve this by simply attacking with widely used white-box attacks like FGSM, and PGD, and a black-box attack like the One-pixel attack. As reported in Table 2, all these attacks can successfully flip the rank assigned by both traditional metrics such as L2, and SSIM (Wang et al., 2004), and learned metrics such as WaDIQaM-FR (Bosse et al., 2018), LPIPS (Zhang et al., 2018b), and DISTS (Ding et al., 2020), in a significant amount of samples.

For the PGD attack, none of the samples needed a perturbation of more than 0.05[2]. Therefore, to report the results of the FGSM attack, we use the threshold $\epsilon < 0.05$. We present the results separately for samples where the originally predicted rank by the metric matches the rank provided by humans. Now, focusing only on the samples where the metric matches with the ranking by humans, we found L2 and DISTS to be the most robust against FGSM and PGD with only about 30% of the samples flipped, while LPIPS and WadIQaM-FR were the least robust, with about 80% of the samples flipped. The same conclusion can also be reached by observing ϵ (or perturbations) required to attack them. Next, despite being a black-box

---

[2] All ϵ (or perturbation) values in this paper were computed from normalized images in the range [-1,1].

Table 3: Comparing samples where the rank by metric was the same as assigned by humans versus samples where it was not.

Network	Same Rank by Human & Metric	Similarity Difference $abs(s_0 - s_1)$
L2	✓	0.036
	✗	0.025
SSIM (Wang et al., 2004)	✓	0.114
	✗	0.054
WadIQaM-FR (Bosse et al., 2018)	✓	0.231
	✗	0.064
LPIPS(Alex) (Zhang et al., 2018b)	✓	0.169
	✗	0.024
LPIPS(VGG) (Zhang et al., 2018b)	✓	0.174
	✗	0.037
DISTS (Ding et al., 2020)	✓	0.103
	✗	0.022

attack, the One-pixel attack can also successful flip ranks. LPIPS(AlexNet) has the least robustness to the One-pixel attack with 82% of the samples flipped, and this lack of adversarial robustness is consistent across all three attacks. SSIM and WadIQaM-FR are more robust to this attack, with only 18% and 31% samples flipped. It is interesting to note that similar results are achievable by using just the score of the adversarial image, i.e., s_{adv} as loss for optimization.

Not surprisingly, it is easier to flip rank for the samples where the metric does not match with human opinion. As reported in Table 2, a much higher number of those samples flip where the rank by metric and humans did not match. These samples have a lower ϵ, which means that lesser perturbations were required to flip the rank. We attribute the easy rank-flipping for these samples to the fact that the distorted images in each sample, i.e., I_{other} and I_{prey}, are much closer to the decision boundary for the rank flip. To test this, we calculate the absolute difference between s_{other} and s_{prey}, i.e., the perceptual distances of I_{other} and I_{prey} from I_{ref}. As reported in Table 3, the similarity difference for these samples is much lesser than samples where the rank predicted by metric is the same as the rank assigned by humans. This result indicates that the samples where rank predicted by metric is not the same as the rank assigned by humans lie closer to the decision boundary, causing them to flip easier.

Imperceptibility. We discuss the imperceptibility of the adversarial perturbations by comparing the root mean square error (RMSE[3]) between the original and the perturbed image. As expected, the PGD attack is stronger than FGSM as it is capable of flipping a significant number of samples with lesser adversarial perturbations. As reported in Table 2, for the PGD attack, a good portion of the adversarial image (I_{adv}) has $\epsilon < 0.01$, while for FGSM, the amount of pixel perturbation all over the image is a constant ϵ value which moreover is higher for a successful attack. Thus, on average, the I_{adv} generated via PGD has lower RMSE and a higher PSNR (see Table 4) with the original image I_{prey}, compared to the I_{adv} generated via FGSM. We also perform a visual sanity check and find the perturbations satisfactorily imperceptible. Only a single pixel is perturbed for I_{adv} generated via the One-pixel attack, which we consider suitably imperceptible.

4.2 Transferable Adversarial Attack

In a real-world scenario, the attacker may not have access to the metric's architecture, hyper-parameters, data, or outputs. In such a scenario, a practical solution for the attacker is to transfer adversarial examples crafted on a source metric to a target perceptual similarity metric. Previous studies have suggested reliable approaches for creating such black-box transferable adversarial examples for image classifiers (Tramèr et al., 2017; Zhou et al., 2018; Inkawhich et al., 2019; Huang et al., 2019; Li et al., 2020; Hong et al., 2021). This paper focuses on perceptual similarity metrics and how they perform against such transferable adversarial examples. Specifically, we transfer the stAdv attack on LPIPS(AlexNet) to other metrics. We chose LPIPS(AlexNet) as it is widely adopted in many computer vision, graphics, and image / video processing applications. Furthermore, we combine the stAdv attack with PGD to increase the transferability of the

---

[3] Throughout this paper, RMSE was calculated on images with pixel values ranging [0,255].

Table 4: Comparing PSNR of adversarial images generated via FGSM vs. PGD. The ϵ for the adversarial images generated via FGSM is < 0.05. A higher mean PSNR of the PGD examples shows that the adversarial perturbations are less perceptible.

Network	Same Rank by Human & Metric	FGSM		PGD	
		PSNR		PSNR	
		μ	σ	μ	σ
L2	✓	40.81	6.49	44.15	5.49
	✗	43.75	7.00	46.08	5.70
SSIM	✓	42.51	6.55	44.60	5.31
(Wang et al., 2004)	✗	46.39	6.09	47.19	5.16
WadIQaM-FR	✓	50.81	5.60	52.19	3.47
(Bosse et al., 2018)	✗	53.92	3.25	54.35	2.73
LPIPS(Alex)	✓	42.80	6.70	46.82	4.09
(Zhang et al., 2018b)	✗	49.98	4.19	50.80	3.14
LPIPS(VGG)	✓	45.96	6.38	48.68	3.72
(Zhang et al., 2018b)	✗	50.56	3.27	51.09	2.46
DISTS	✓	39.50	6.22	41.19	5.75
(Ding et al., 2020)	✗	43.64	6.95	44.41	6.39

Table 5: Transferable adversarial attacks on perceptual similarity metrics. The adversarial examples were generated by attacking LPIPS(AlexNet) via stAdv. In total, there are 2726 samples. Next, we attacked LPIPS(AlexNet) using PGD(10). Then, we combined stAdv+PGD(10) by perturbing the stAdv generated images with PGD(10). Accurate samples are the ones for which the predicted rank by metric is equal to the rank assigned by humans. The transferability increases when the attacks are combined.

Network	#Accurate Samples	# Accurate Samples Flipped						
		PGD(10)	PGD(20)	stAdv	stAdv + PGD(5)	stAdv + PGD(10)	stAdv + PGD(15)	stAdv + PGD(20)
L2	2099/77%	101/5%	174/8%	77/4%	134/6%	189/9%	200/10%	257/12%
SSIM (Wang et al., 2004)	2093/77%	237/11%	442/21%	78/4%	180/9%	339/16%	370/18%	540/26%
FSIMc (Zhang et al., 2011)	2025/74%	222/11%	325/16%	202/10%	233/12%	302/15%	310/15%	393/19%
WaDIQaM-FR (Bosse et al., 2018)	2083/76%	95/5%	186/9%	59/3%	85/4%	146/7%	156/7%	238/11%
GTI-CNN (Ma et al., 2018)	1946/71%	448/23%	480/25%	494/25%	488/25%	504/26%	510/26%	543/28%
DISTS (Ding et al., 2020)	2413/89%	311/13%	576/24%	146/6%	257/11%	510/21%	546/23%	801/33%
LPIPS(Squz.) (Zhang et al., 2018b)	2503/92%	298/12%	656/26%	114/5%	221/9%	519/21%	555/22%	886/35%
LPIPS(VGG) (Zhang et al., 2018b)	2317/85%	435/19%	814/35%	131/6%	288/12%	643/28%	685/30%	992/43%
E-LPIPS (Kettunen et al., 2019b)	2442/90%	503/21%	643/26%	517/21%	552/23%	641/26%	655/27%	817/33%
Watson-DFT (Czolbe et al., 2020)	2179/80%	387/18%	614/28%	216/10%	324/15%	532/24%	562/26%	750/34%
PIM-1 (Bhardwaj et al., 2020)	2468/91%	696/28%	814/33%	756/31%	772/31%	826/33%	852/35%	958/39%
PIM-5 (Bhardwaj et al., 2020)	2457/90%	751/31%	844/34%	765/31%	791/32%	864/35%	893/36%	963/39%

adversarial examples to other metrics. In this study, we only consider samples for which the metrics and the human opinions agree on their rankings.

stAdv. As shown in Figure 5, stAdv has the capability of attacking high-level image features. As a white-box attack on LPIPS(AlexNet), out of the 11,303 accurate samples from total 12,227 samples, stAdv was able to flip judgment on 4658 samples with a mean RMSE of 2.37 with standard deviation 1.42. Because we need high imperceptibility, we remove samples with RMSE > 3 and are left with 3327 samples. We then perform a visual sanity check and remove some more with ambiguity, keeping only strictly imperceptible samples. In the end, we have 2726 samples, with a mean RMSE of 1.58 with standard deviation 0.63, which we transfer to other metrics as a black-box attack. As reported in Table 5, all metrics are prone to the attack. WaDIQaM-FR (Bosse et al., 2018) is most robust, while PIM (Bhardwaj et al., 2020) that was found robust to small imperceptible shifts is highly susceptible to this attack, although PIM is 15% more accurate than WaDIQaM-FR. Finally, we saw that, on average, learned metrics are more correlated with human opinions, but traditional metrics exhibit more robustness to the imperceptible transferable stAdv adversarial perturbations.

PGD(10). We now attack the original 2726 selected samples with the PGD attack. As shown in Section 4.1, perturbations generated via PGD have low perceptibility; hence, we create adversarial samples using PGD. In stAdv, we stopped the attack when the rank predicted by LPIPS(AlexNet) flipped. While in PGD, for comparison's sake, we fix the number of attack iterations to 10 for each sample to guarantee the transferability of perturbations. We call this transferable attack PGD(10), and the mean RMSE of the adversarial images

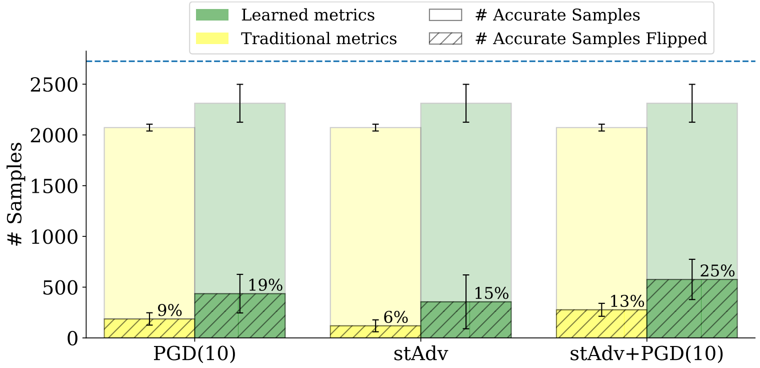


Figure 6: Comparing traditional metrics (L2, SSIM (Wang et al., 2004), and FSIMc (Zhang et al., 2011)) versus learned metrics (WaDIQaM-FR (Bosse et al., 2018), GTI-CNN (Ma et al., 2018), LPIPS (Zhang et al., 2018b), DISTS (Ding et al., 2020), E-LPIPS (Kettunen et al., 2019b), Watson-DFT (Czolbe et al., 2020), and PIM (Bhardwaj et al., 2020)).

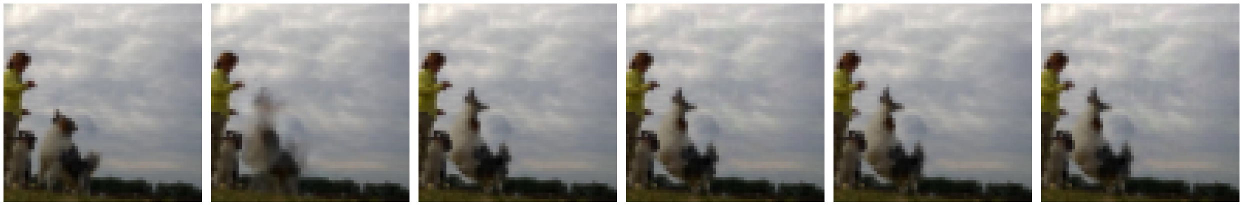
generated is 1.28 with a standard deviation of 0.11. The metrics SSIM and WaDIQaM-FR are most robust to the transferable PGD(10) attack, as reported in Table 5.

Combining stAdv and PGD(10). The attacks stAdv and PGD are orthogonal approaches as PGD (ℓ_∞-bounded attack) manipulates the intensity of individual pixels while stAdv (spatial attack) manipulates the location of the pixels. We now combine the two by attacking the samples generated via stAdv with PGD(10). The mean RMSE of the generated adversarial images is 2.19 with a standard deviation of 0.41, just 0.61 higher than images generated via stAdv. As reported in Table 5, the increase in severity of the adversarial perturbations in FlowAdv+PGD(10) leads to increased transferability. This result also is consistent with previous findings by Engstrom et al. (2019) where they combined PGD on top of their spatial attack and found that it leads to an additive increment in the misclassification rate.

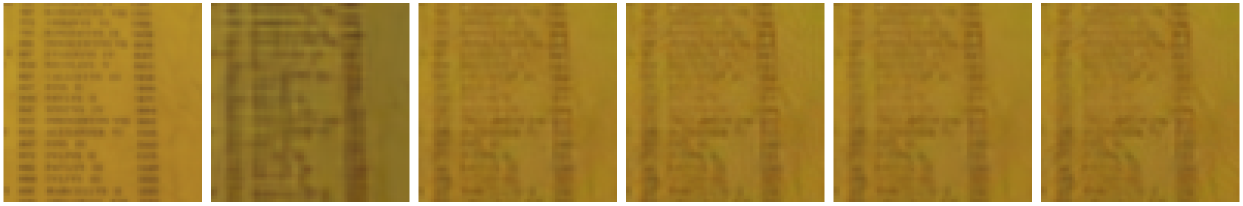
Summary. In this paper, we successfully demonstrate that a wide variety of perceptual similarity metrics are susceptible to adversarial attacks. We show that adversarial perturbations crafted for LPIPS(AlexNet) generated via stAdv, can be transferred to other metrics. Furthermore, combining stAdv (spatial attack) with PGD (ℓ_∞-bounded attack) increases their transferability. We showcase a few examples in Figures 7, and 8. In addition, the severity of the attack increases with the increasing number of PGD iterations (see Table 5). Our investigations also show that although more accurate, learned metrics may not be more robust than traditional ones (see Figure 6). In summary, our findings point towards the need to develop robust perceptual similarity metrics.

5 Conclusion

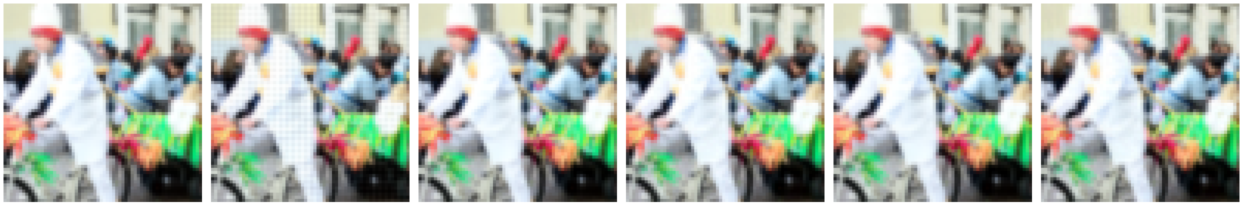
In this paper, we studied the robustness of various traditional and learned perceptual similarity metrics to imperceptible perturbations. We devised a methodology to craft such perturbations via adversarial attacks. Our findings suggest that, when comparing two images with respect to a reference, the addition of imperceptible distortions can overturn a metric's similarity judgment. The results of our study indicate that even learned perceptual metrics that match with human similarity judgments are susceptible to such imperceptible adversarial perturbations. We crafted adversarial examples using the spatial attack, stAdv, that were transferable to other metrics. We show that when combined with the PGD attack, the transferability of the adversarial examples can be further increased. We will make our code and data publicly available to encourage further studies on the current topic with more comprehensive benchmarks. Perceptual similarity metrics are designed to simulate the human visual system, and for this reason, these metrics are increasingly used in the assessment of image and video quality in real-world scenarios. Since invisible distortions can negatively impact the performance of similarity metrics, future studies for the design and development of newer metrics should also focus on validating robustness.



I_{ref}	I_{other}	I_{prey}	I_{adv} PGD(10)	I_{adv} stAdv	I_{adv} stAdv+PGD(10)
L2 ↓	0.0091	0.0127	0.0128	0.0128	0.0128
SSIM ↑	0.8754	0.8823	0.8721	0.8770	0.8635
FSIMc ↑	0.99069	0.99058	0.99061	0.99061	0.99064
WaDIQaM-FR ↓	1.2747	1.3567	1.3730	1.3622	1.3572
GTI-CNN ↓	135.61	255.97	220.48	217.10	217.65
DISTS ↓	0.0996	0.0729	0.0952	0.0873	0.1152
LPIPS(Squeeze) ↓	0.0736	0.0393	0.0421	0.0490	0.0517
LPIPS(VGG) ↓	0.0916	0.0669	0.0802	0.0783	0.1011
E-LPIPS ↓	0.0057	0.0041	0.0069	0.0068	0.0075
Watson-DFT ↓	908.63	922.66	1112.21	1071.77	1136.02
PIM-1 ↓	0.6141	0.4485	1.1852	1.2937	1.2917
PIM-5 ↓	6.2894	5.0282	11.3717	12.0675	12.2006

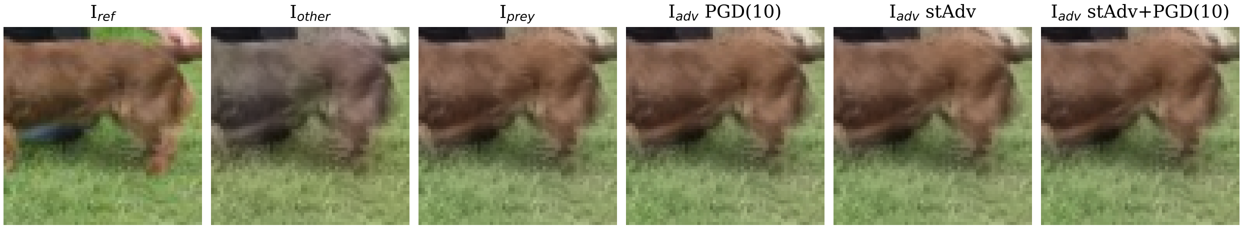


I_{ref}	I_{other}	I_{prey}	I_{adv} PGD(10)	I_{adv} stAdv	I_{adv} stAdv+PGD(10)
L2 ↓	0.0361	0.0050	0.0057	0.0056	0.0063
SSIM ↑	0.3163	0.5807	0.5528	0.5646	0.5357
FSIMc ↑	0.98102	0.98274	0.98079	0.98016	0.97770
WaDIQaM-FR ↓	1.3614	1.2760	1.2575	1.2983	1.2943
GTI-CNN ↓	133.18	59.11	77.51	78.95	85.07
DISTS ↓	0.2772	0.2324	0.2739	0.2678	0.3021
LPIPS(Squeeze) ↓	0.0986	0.0761	0.1231	0.1058	0.1762
LPIPS(VGG) ↓	0.2167	0.1601	0.2451	0.2028	0.3269
E-LPIPS ↓	0.0115	0.0103	0.0169	0.0170	0.0178
Watson-DFT ↓	2433.66	1344.98	1415.91	1392.29	1410.53
PIM-1 ↓	2.9635	2.5469	3.2072	3.2161	3.5531
PIM-5 ↓	33.8370	27.0413	35.6628	37.6837	39.1791

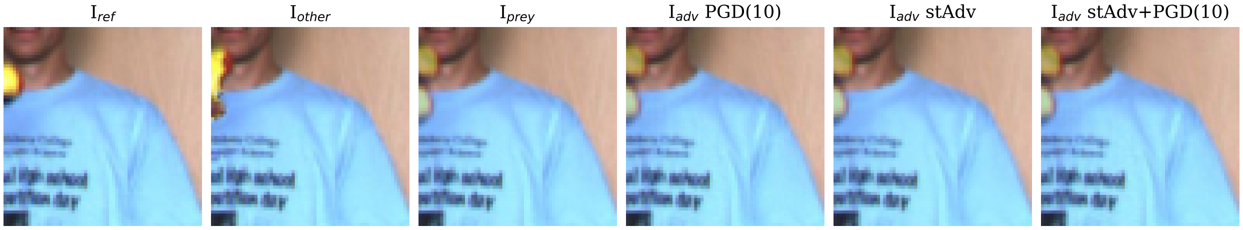


I_{ref}	I_{other}	I_{prey}	I_{adv} PGD(10)	I_{adv} stAdv	I_{adv} stAdv+PGD(10)
L2 ↓	0.0010	0.0010	0.0012	0.0012	0.0015
SSIM ↑	0.9739	0.9779	0.9730	0.9743	0.9681
FSIMc ↑	0.99992	0.99985	0.99983	0.99983	0.99980
WaDIQaM-FR ↓	1.1214	1.1190	1.1177	1.1165	1.1184
GTI-CNN ↓	47.72	11.53	79.21	85.79	84.42
DISTS ↓	0.1180	0.0065	0.0200	0.0129	0.0283
LPIPS(Squeeze) ↓	0.0023	0.0013	0.0025	0.0017	0.0033
LPIPS(VGG) ↓	0.0791	0.0027	0.0069	0.0038	0.0103
E-LPIPS ↓	0.0139	0.0002	0.0045	0.0047	0.0052
Watson-DFT ↓	924.09	541.48	783.71	693.21	861.64
PIM-1 ↓	0.7539	0.0110	1.0787	1.1750	1.1291
PIM-5 ↓	7.0737	0.1121	11.2964	12.0483	11.7169

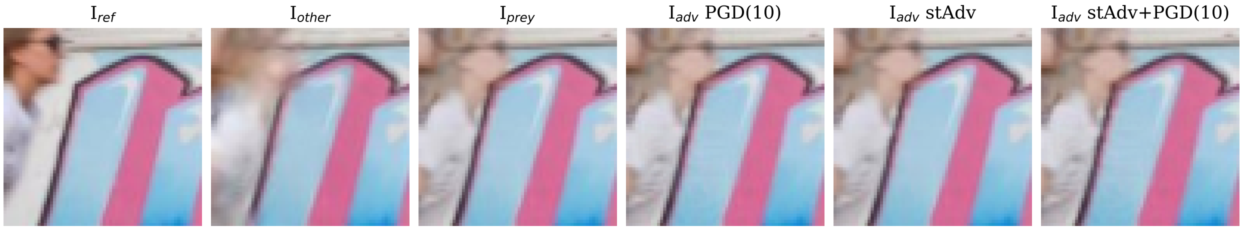
Figure 7: Transferable attack on perceptual similarity metrics. In example 1 (Top), the RMSE between I_{prey} and I_{adv} images (left to right) is 1.26, 2.89, and 2.47. In example 2 (Mid.), the RMSE between I_{prey} and I_{adv} images (left to right) is 1.29, 1.02, and 1.91. In example 3 (Bot.), RMSE between I_{prey} and I_{adv} images (left to right) is 1.43, 1.2, and 2.15.



	I_{other}	I_{prey}	I_{adv} PGD(10)	I_{adv} stAdv	I_{adv} stAdv+PGD(10)
L2 ↓	0.0079	0.0047	0.0048	0.0047	0.0048
SSIM ↑	0.7717	0.8554	0.8501	0.8526	0.8436
FSIMc ↑	0.99940	0.99937	0.99926	0.99922	0.99903
WaDIQaM-FR ↓	1.3602	1.2796	1.2949	1.2962	1.3113
GTI-CNN ↓	95.99	81.11	139.53	133.61	165.44
DISTS ↓	0.1303	0.1030	0.1111	0.1070	0.1139
LPIPS(Squeeze) ↓	0.1149	0.0794	0.0880	0.0855	0.0940
LPIPS(VGG) ↓	0.1893	0.1188	0.1346	0.1244	0.1409
E-LPIPS ↓	0.0115	0.0077	0.0124	0.0131	0.0140
Watson-DFT ↓	1501.56	1025.79	1278.13	1305.37	1422.43
PIM-1 ↓	2.1654	1.0225	2.8536	3.2559	3.1688
PIM-5 ↓	21.3579	10.0332	26.9732	29.6540	29.2036



	I_{other}	I_{prey}	I_{adv} PGD(10)	I_{adv} stAdv	I_{adv} stAdv+PGD(10)
L2 ↓	0.0121	0.0133	0.0133	0.0133	0.0133
SSIM ↑	0.9068	0.9112	0.9006	0.9103	0.8958
FSIMc ↑	0.99392	0.99181	0.99185	0.99187	0.99183
WaDIQaM-FR ↓	1.1942	1.2634	1.2653	1.2699	1.2813
GTI-CNN ↓	53.66	28.88	62.75	61.31	69.90
DISTS ↓	0.1341	0.1034	0.1121	0.1056	0.1132
LPIPS(Squeeze) ↓	0.0264	0.0371	0.0395	0.0375	0.0411
LPIPS(VGG) ↓	0.0545	0.0462	0.0520	0.0472	0.0571
E-LPIPS ↓	0.0039	0.0033	0.0055	0.0054	0.0065
Watson-DFT ↓	1097.13	901.26	1147.84	1078.05	1157.19
PIM-1 ↓	0.2170	0.2429	1.0924	1.2546	1.2119
PIM-5 ↓	3.4366	2.9138	12.0777	13.0601	13.2696



	I_{other}	I_{prey}	I_{adv} PGD(10)	I_{adv} stAdv	I_{adv} stAdv+PGD(10)
L2 ↓	0.0368	0.0336	0.0339	0.0338	0.0342
SSIM ↑	0.6526	0.6797	0.6693	0.6778	0.6667
FSIMc ↑	0.96532	0.97225	0.97213	0.97228	0.97208
WaDIQaM-FR ↓	1.1196	1.0508	1.0513	1.0550	1.0580
GTI-CNN ↓	473.92	325.45	379.77	382.28	386.48
DISTS ↓	0.2266	0.2105	0.2278	0.2132	0.2297
LPIPS(Squeeze) ↓	0.0694	0.0467	0.0516	0.0483	0.0545
LPIPS(VGG) ↓	0.2057	0.1614	0.1823	0.1640	0.1897
E-LPIPS ↓	0.0138	0.0119	0.0131	0.0139	0.0132
Watson-DFT ↓	2314.26	2072.06	2137.63	2006.34	2305.92
PIM-1 ↓	1.6024	0.8015	1.4746	1.5086	1.5491
PIM-5 ↓	14.5537	10.2235	14.8395	15.0432	15.2823

Figure 8: Transferable attack on perceptual similarity metrics. In example 1 (Top), the RMSE between I_{prey} and I_{adv} images (left to right) is 1.35, 1.43, and 2.25. In example 2 (Mid.), the RMSE between I_{prey} and I_{adv} images (left to right) is 1.25, 0.95, and 1.77. In example 3 (Bot.), RMSE between I_{prey} and I_{adv} images (left to right) is 1.37, 0.99, and 2.0.

References

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, volume 80, pp. 274–283, 2018.

Sangnie Bhardwaj, Ian Fischer, Johannes Ballé, and Troy Chinen. An unsupervised information-theoretic perceptual quality metric. In *Advances in Neural Information Processing Systems 33*, 2020.

Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and DA Forsyth. Unrestricted adversarial examples via semantic manipulation. In *International Conference on Learning Representations*, 2019.

Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, 2018.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations*, 2018.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv/2010.09670*, 2020.

Steffen Czolbe, Oswin Krause, Ingemar Cox, and Christian Igel. A loss function for generative neural networks based on watson's perceptual model. In *Advances in Neural Information Processing Systems*, pp. 2051–2061, 2020.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

Hadi Mohaghegh Dolatabadi, Sarah Erfani, and Christopher Leckie. Advflow: Inconspicuous black-box adversarial attacks using normalizing flows. In *Advances in Neural Information Processing Systems*, 2020.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018.

Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems*, pp. 658–666, 2016.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, pp. 1802–1811, 2019.

Abhijay Ghildyal and Feng Liu. Shift-tolerant perceptual similarity metric. In *European Conference on Computer Vision*, 2022.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy S. Ren, and Chao Dong. Pipal: A large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, volume 12356, pp. 633–651, 2020.

Sanghyun Hong, Yigitcan Kaya, Ionuț-Vlad Modoranu, and Tudor Dumitras. A panda? no, it's a sloth: Slowdown attacks on adaptive multi-exit neural network inference. In *International Conference on Learning Representations*, 2021.

Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1614–1619, 2018.

Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *IEEE International Conference on Computer Vision*, pp. 4733–4742, 2019.

Nathan Inkawhich, Wei Wen, Hai Helen Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7066–7074, 2019.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pp. 2017–2025, 2015.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, volume 9906, pp. 694–711, 2016.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

Markus Kettunen, Erik Härkönen, and Jaakko Lehtinen. Deep convolutional reconstruction for gradient-domain rendering. *ACM Transactions on Graphics*, 38, 2019a.

Markus Kettunen, Erik Härkönen, and Jaakko Lehtinen. E-lpips: Robust perceptual image similarity via random transformation ensembles. *arXiv/1906.03973*, 2019b.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *International Conference on Learning Representations - Workshop*, 2017.

Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. In *Advances in Neural Information Processing Systems*, 2019.

Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *International Conference on Learning Representations*, 2020.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv/1609.04802*, 2016.

Qizhang Li, Yiwen Guo, and Hao Chen. Yet another intermediate-level attack. In *European Conference on Computer Vision*, pp. 241–257, 2020.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.

Ning Lu, Li Dong, Diqun Yan, and Xianliang Jiang. On attacking deep image quality evaluator via spatial transform. In *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2876–2881, 2022.

Kede Ma, Zhengfang Duanmu, and Zhou Wang. Geometric transformation invariant image quality assessment using convolutional neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6732–6736, 2018.

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*, 2018.

Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5437–5446, 2020.

Nicolas Papernot, P. Mcdaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv/1602.02697*, 2016.

Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1808–1817, 2018.

Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *IEEE International Conference on Computer Vision*, pp. 4491–4500, 2017.

Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1151–1160, 2020.

Sanghyun Son, Jaerin Lee, Seungjun Nah, Radu Timofte, Kyoung Mu Lee, Yihao Liu, Liangbin Xie, Li Siyao, Wenxiu Sun, Yu Qiao, Chao Dong, Woonsung Park, Wonyong Seo, Munchurl Kim, Wenhao Zhang, Pablo Navarrete Michelini, Kazutoshi Akita, and Norimichi Ukita. AIM 2020 challenge on video temporal super-resolution. In *European Conference on Computer Vision - Workshops*, pp. 23–40, 2020.

Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. *Advances in Neural Information Processing Systems*, 31, 2018.

Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv/1704.03453*, 2017.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33, 2020.

Yajie Wang, Shangbo Wu, Wenyi Jiang, Shengang Hao, Yu-an Tan, and Quanxin Zhang. Demiguise attack: Crafting invisible semantic adversarial perturbations with perceptual similarity. In *International Joint Conference on Artificial Intelligence*, 2021.

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.

Andrew B Watson. DCT quantization matrices visually optimized for individual images. In *Human vision, visual processing, and digital display IV*, volume 1913, pp. 202–216. International Society for Optics and Photonics, 1993.

Eric Wong, Frank Schmidt, and Zico Kolter. Wasserstein adversarial examples via projected Sinkhorn iterations. In *International Conference on Machine Learning*, volume 97, pp. 6808–6817, 2019.

Lei Wu and Zhanxing Zhu. Towards understanding and improving the transferability of adversarial examples in deep neural networks. In *Asian Conference on Machine Learning*, volume 129 of *PMLR*, pp. 837–850, 18–20 Nov 2020.

Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan Yuille. Improving transferability of adversarial examples with input diversity. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L Yuille. Adversarial attacks beyond the image space. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Huan Zhang, Hongge Chen, Zhao Song, Duane Boning, Inderjit S Dhillon, and Cho-Jui Hsieh. The limitations of adversarial training and the blind-spot attack. In *International Conference on Learning Representations*, 2018a.

Kai Zhang, Shuhang Gu, and Radu Timofte. NTIRE 2020 challenge on perceptual extreme super-resolution: Methods and results. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 492–493, 2020.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: a feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018b.

Weixia Zhang, Dingquan Li, Xiongkuo Min, Guangtao Zhai, Guodong Guo, Xiaokang Yang, and Kede Ma. Perceptual attacks of no-reference image quality models with human-in-the-loop. In *Advances in Neural Information Processing Systems*, 2022.

Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *European Conference on Computer Vision*, pp. 452–467, 2018.

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, volume 9909, pp. 597–613, 2016.

A FGSM Attack on Similarity Metrics

We explain the FGSM attack on perceptual similarity metrics in Algorithm 3.

Algorithm 3: FGSM attack on Similarity Metrics

```
Input: $I_1$, $I_2$, $I_{ref}$, metric $f$, $max\_\epsilon$ (0.05)
Output: Least $\epsilon$ value which led to rank flip
$s_0 = f(I_{ref}, I_0)$; $s_1 = f(I_{ref}, I_1)$;
$rank = int(s_0 > s_1)$ // If $I_0$ is more similar to $I_{ref}$ then $rank$ is 0 else 1
if $rank = 1$ then $I_{prey} = I_1$; $s_{other} = s_0$;
else $I_{prey} = I_0$; $s_{other} = s_1$;
$s_{prey} = f(I_{ref}, I_{prey})$
$J = \Big((s_{other}/(s_{other} + s_{prey})) - 1\Big)^2$ // Loss
$signed\_grad = sign\Big(\nabla_{I_{prey}} J\Big)$
$\epsilon = 0.0001$
while $\epsilon \leq max\_\epsilon$ do
    $I_{adv} = I_{prey} + \epsilon \cdot signed\_grad$
    $I_{adv} = clip(I_{adv}, min = -1, max = 1)$ // range [-1,1]
    $s_{adv} = f(I_{ref}, I_{adv})$
    if $s_{adv} > s_{other}$ then
        return True // Attack successful
    $\epsilon = \epsilon + 0.0001$
return 1 // Largest value of $\epsilon$
```