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. To the best of our knowledge, no study to date has systematically examined 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 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 FlowAdv, our flow-based spatial attack, in a white-box setting to craft adversarial examples that can effectively transfer to other similarity metrics in a black-box setting. In addition, we combine the spatial attack FlowAdv 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?"*

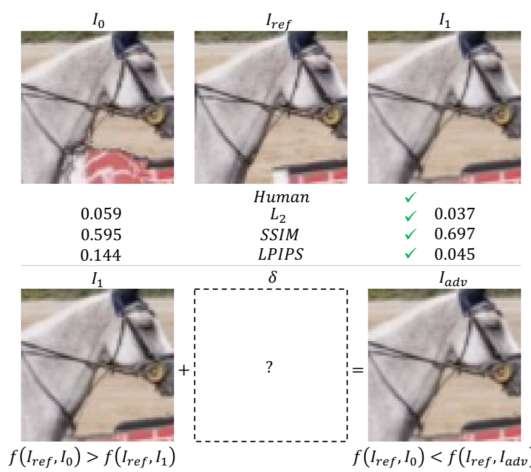


Figure 1: I_1 is more similar to I_{ref} than I_0 according to all perceptual similarity metrics and humans. We attack I_1 by adding imperceptible adversarial perturbations (δ) such that the metric (f) flips its earlier assigned rank, i.e., in the above sample, I_0 becomes more similar to I_{ref}.

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 similarity metrics often overturn their judgment.

There are two approaches to examining the robustness of perceptual similarity metrics: (1) addition of small amounts of hand-crafted geometric distortions, 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), and black-box attacks such as 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. 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 develop an additional spatial adversarial attack, which geometrically deforms the image. We call it FlowAdv as 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.

Previous studies have shown that adversarial examples generated using the parameters of a source model are transferable to a target model (Liu et al., 2017; Xie et al., 2018; 2019). In our work, we use LPIPS(AlexNet) as the source model and attack it via FlowAdv. 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. Many approaches have been studied to improve the transferability of attacks (Szegedy et al., 2014; Papernot et al., 2016; Liu et al., 2017; Wu & Zhu, 2020). In our work, we combine FlowAdv (spatial attack) with PGD (l_∞-bounded attack) that strengthens the severity of the adversarial examples.

Our paper is organized as follows. In Section 2, we review past literature and highlight recent developments. In Section 3, we describe the adversarial attacks used in this paper and explain how we extend them for tricking perceptual similarity metrics into overturning their judgment. While classical adversarial attacks like FGSM, PGD, and the One-pixel attack are effective, they do not geometrically distort the image. Therefore, in addition, we propose our spatial attack FlowAdv to create transferable adversarial examples and describe it in Section 3. We further combine FlowAdv (spatial attack) with PGD (l_∞-bounded attack) to craft stronger, transferable adversarial perturbations. In Section 4, we explain our experimental setup and report our results on (1) validating that similarity judgments by perceptual similarity metrics can flip on the addition of imperceptible perturbations, and (2) comparing the robustness of various metrics to adversarial perturbations.

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 (Bhardwaj et al., 2020; Ding et al., 2020; Kettunen et al., 2019b; Ma et al., 2018; Prashnani et al., 2018; Zhang et al., 2018b) 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 the robustness of these metrics 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 towards 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 estimation is performed using symmetrized Kullback–Leibler 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). So far, the majority of prior research has focused on geometric distortions, while no study has been reported with more advanced adversarial perturbations. 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; Carlini & Wagner, 2017; Hosseini & Poovendran, 2018; Madry et al., 2018; Xiao et al., 2018; Brendel et al., 2018; Song et al., 2018; Zhang et al., 2018a; Laidlaw & Feizi, 2019; Su et al., 2019; Wong et al., 2019; Bhattad et al., 2019; Zeng et al., 2019; Dolatabadi et al., 2020; Tramèr et al., 2020; Laidlaw et al., 2020; Croce et al., 2020), but none of the previous investigations have ever considered attacking perceptual similarity metrics. This paper focuses on investigating the adversarial robustness of similarity metrics.

Recent work underlines the importance of perceptual distance as a bound for adversarial attacks (Laidlaw et al., 2020; Wang et al., 2021). 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. Laidlaw et al. (2020) provided evidence that l_p-bounded and spatial attacks are near subsets of the NPTM. Further, in one of their studies, they found that LPIPS correlates 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.

Optical flow can be used for crafting adversarial samples that utilize the structure of the image being attacked. AdvFlow by Dolatabadi et al. (2020) is one such attack which uses normalizing flows (Rezende & Mohamed, 2015) and natural evolution strategies (Wierstra et al., 2008). Spatially transformed adversarial example optimization method, commonly known as stAdv attack (Xiao et al., 2018) is more closely related to our spatial attack method, FlowAdv. The stAdv attack optimizes a flow vector, increasing the probability of misclassification using Carlini & Wagner loss, while simultaneously minimizing displacement in pixels (Carlini & Wagner, 2017). We propose a variation of the stAdv attack, that generates a displacement vector via a CNN using the image being attacked. Then, with backward warping, we create an adversarial image. Engstrom et al. (2019) create small translations or rotations using a spatial transformer (Jaderberg et al., 2015) to evaluate the spatial robustness of image classifiers. They further combine their spatial attack with an l_∞-bounded attack to increase misclassification rates. Many approaches have been studied to improve the transferability of attacks (Liu et al., 2017; Papernot et al., 2016; Szegedy et al., 2014; Wu & Zhu, 2020). Liu et al. (2017) apply the attacks simultaneously to create an ensemble of attacks. Xie et al. (2019) used random transformations to increase the diversity of the adversarial samples that aids in the transferability of the attack. Our approach to improve transferability is more similar to Engstrom et al. (2019), where we combine FlowAdv, our spatial attack, with PGD, an l_∞-bounded attack.

3 METHOD

Dataset. LPIPS is trained on the Berkeley-Adobe perceptual patch similarity (BAPPS) dataset. 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 image quality assessment, the similarity scores were generated using a two-alternative forced choice test, where, the participants were asked "Which of the two I_0 or I_1 is more similar to I_{ref}?". For the validation set, 5 responses per sample were collected. The final human judgment was based on the average of the responses. The types of distortions in this dataset are traditional, CNN-based, and distortions by real algorithms such as superresolution, frame interpolation, deblurring, and colorization. Human opinions are divided in some of the samples in the validation set, i.e., all responses in a sample may not have voted for the same distorted image. For our experiment, 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 (see example in Figure 1). In total, there are 12,227 such samples that we used for our experiments.

Attack Methods. 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), and One-pixel attack (Su et al., 2019), and our adversarial flow attack (FlowAdv) to generate such adversarial samples. The existing attack methods we use were originally devised to dupe 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}. Hence, considering s_i as the similarity score [1] between I_i and I_{ref}, we decide I_{prey} and I_{other} as follows:

$$(I_{prey}, I_{other}) = \begin{cases} (I_0, I_1), & \text{if } (s_0 < s_1). \\ (I_1, I_0), & \text{otherwise.} \end{cases} \tag{1}$$

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:

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

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 2, 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. We define Algorithm 3 (refer Appendix A.2) 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.

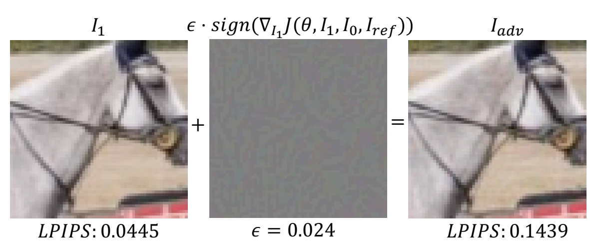


Figure 2: FGSM attack.

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

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 (P_c), restricts the pixel values to a predefined range $[-\epsilon, +\epsilon]$.

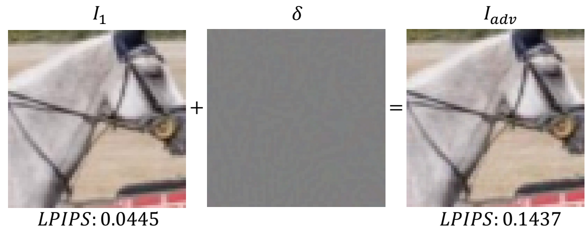


Figure 3: PGD attack.

Algorithm 1: PGD attack on Similarity Metrics

Input: I_0, I_1, I_{ref}, metric f, step size α = 0.001, max iterations k = 40, perturbation limit ϵ = 0.1
Output: $attack\_success$ True on rank flip

1 $s_0 = f(I_{ref}, I_0)$
2 $s_1 = f(I_{ref}, I_1)$
3 // If I_0 is more similar to I_{ref} then $rank$ is 0 else 1
4 $rank = int(s_0 > s_1)$ // smaller $s_i \equiv$ more similar
5 **if** $rank = 1$ **then** $I_{prey} = I_1$; $s_{other} = s_0$;
6 **else** $I_{prey} = I_0$; $s_{other} = s_1$;
7 $k = 0$
8 δ = zeros_like(I_{prey}) // perturbation
9 **while** $k \leq 40$ **do**
10 $\quad I_{adv} = clip(I_{prey} + \delta, min = -1, max = 1)$
11 $\quad s_{adv} = f(I_{ref}, I_{adv})$
12 $\quad$ **if** $s_{adv} > s_{other}$ **then**
13 $\quad\quad$ **return** *True* // Attack successful
14 $\quad J = \left((s_{other}/(s_{other} + s_{adv})) - 1\right)^2$ // Loss
15 $\quad signed\_grad = sign\left(\nabla_{I_{adv}} J\right)$
16 $\quad I'_{adv} = I_{adv} + \alpha * signed\_grad$
17 $\quad \delta = clip(I'_{adv} - I_{prey}, min = -\epsilon, max = +\epsilon)$
18 $\quad k = k + 1$
19 **return** *False* // Attack unsuccessful

We describe the algorithm for PGD attack in Algorithm 1. Using the same loss as Equation 2, this multistep attack is defined as:

$$I_{adv}^{t+1} = P_c\big(I_{adv}^t + \alpha \cdot sign(\nabla_{I_{adv}^t} J(\theta, I_{adv}^t, I_{other}, I_{ref})\big) \tag{3}$$

Alternatively, the attack can be stated as:

$$I_{adv}^{t+1} = P_c\big(FGSM(I_{adv}^t)\big) \tag{4}$$

As expressed in equation 3, 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} (Line 18 Algorithm 1), and in our method, δ is bounded by l_∞ norm. Hence, this attack is an $\boldsymbol{l_\infty}$**-bounded attack**.

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 differential evolution optimization starts with an initial population X for a subset of pixels. Each vector x in X contains a pixel's index and its 3 perturbation values for the channels r, g, and b. In each iteration, mutation and recombination evolve the population towards an optimal x^* that flips rank.

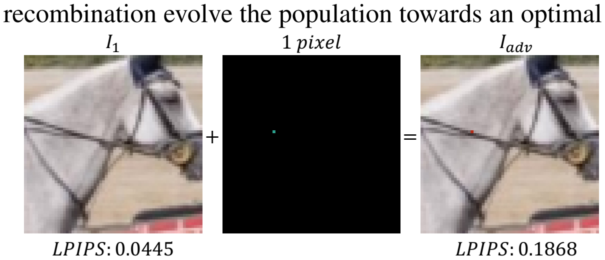


Figure 4: One-pixel attack.

The objective of the One-pixel attack is to find the optimal adversarial perturbation vector x^* as summarized below:

$$\begin{aligned} \underset{x^*}{\text{maximize}} \quad & f(I_{prey} + x, I_{ref}) \\ \text{subject to} \quad & x \subset X \text{ and } ||x||_0 \leq d \end{aligned} \tag{5}$$

where d is 1 for the One-pixel attack. The attack terminates when the condition for rank flip is satisfied, i.e., $s_{adv} > s_{other}$. We refer the reader to Appendix A.3 for more details on the steps involved in finding x^* via differential evolution, and the algorithm used for the One-pixel attack on similarity metrics.

3.1 Spatial Attack: FlowAdv

We introduce a new spatial attack FlowAdv. The goal of this attack is to deform the image geometrically by displacing pixels. FlowAdv generates adversarial perturbations in the spatial domain rather than directly manipulating pixel intensity values. We consider FlowAdv as a variation of the stAdv attack (Xiao et al., 2018) which optimizes a flow map directly, whereas FlowAdv uses the input image to predict the flow. Previous works have studied the problem of optical flow estimation from a single image for motion prediction (Pintea et al., 2014; Walker et al., 2015; Yang & Soatto, 2018; Gao et al., 2018; Holynski et al., 2021). We estimate the flow from a static image for a different purpose, i.e., we create an invisible spatial distortion using the flow biased by the input image.

In our attack method, we predict the flow vector ((u, v)) for x and y direction) by passing I_{prey} through a CNN (f_θ) having 5 layers with 256 channels each, followed by 1x1 convolutional layers. The flow is then applied to I_{prey} via backward-warping to generate the adversarial image I_{adv}.

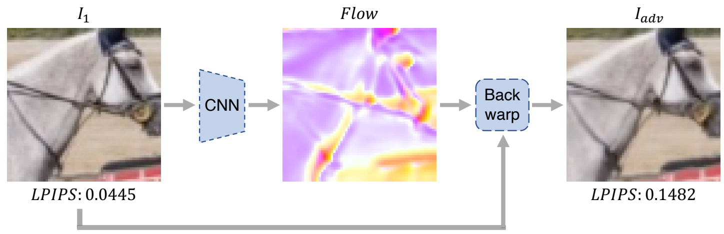


Figure 5: FlowAdv attack.

For each sample, we start with random weights and then optimize using Adam (Kingma & Ba, 2015) for the following loss function:

$$\mathcal{L} = \mathcal{L}_{perturb} + \mathcal{L}_{rank} \quad (6)$$

$$\mathcal{L}_{perturb} = \alpha * \mathcal{L}_1 + \beta * \mathcal{L}_{Charbonnier} \quad (7)$$

$$\mathcal{L}_1 = ||I_{prey} - I_{adv}||_1 \quad (8)$$

$$\mathcal{L}_{Charbonnier} = \rho(I_{prey} - I_{adv}) \quad (9)$$

$$\mathcal{L}_{rank} = \left(\frac{s_{other}}{s_{other} + s_{adv}} \right)^2 \quad (10)$$

where, $\rho(x) = \sqrt{x^2 + 1e\text{-}6}$ (Lai et al., 2017), $\alpha = 0.0001$, and $\beta = 0.5$. As we minimize $\mathcal{L}_{rank}$, s_{adv} will increase, and thus rank will be flipped. Simultaneously, we also minimize the l_1 distance between I_{prey} and I_{adv}, enforcing the perturbations to be constrained within an l_1 ball. For a successful attack, we apply an additional constraint that $\mathcal{L}_1 < 0.05$, thus ensuring that the optimal I_{adv} that satisfies the rank flip condition makes as little change to the attacked image I_{prey} as possible.

Algorithm 2: FlowAdv attack on LPIPS

```
Input: I_0, I_1, I_ref, LPIPS f, max_itr(50),
  max_restarts (2), attack model f_A
Output: attack_success True on rank flip
s_0 = f(I_ref, I_0)
s_1 = f(I_ref, I_1)
// If I_0 is more similar to I_ref then rank is 0 else 1
rank = int(s_0 > s_1) // smaller s_i ≡ more similar
if rank = 1 then I_prey = I_1; s_other = s_0;
else I_prey = I_0; s_other = s_1;
while k ≤ max_restarts do
    f_A = init() // initialize the attack model
    i = 1
    while i ≤ max_itr do
        f_A.optimize_parameters(L)
        (u, v) = f_A(I_prey) // adversarial flow
        I_adv = backwarp((u, v), I_prey)
        L_rank, L_perturb =
          calc_loss(I_ref, I_other, I_prey, I_adv, f)
        L = L_rank + L_perturb
        s_adv = f(I_ref, I_adv)
        if s_adv > s_other and L_perturb < 0.05 then
            return True // Attack successful
        i = i + 1
    k = k + 1
return False // Attack unsuccessful
```

4 Experiments and Results

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 2AFC scores on the complete BAPPS validation dataset (refer Appendix A.1 for 2AFC calculation). As shown in Table 1, all these 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 difficult cases.

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

Network	2AFC (%) complete BAPPS (36344 samples)	Accuracy (%) 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

We organize our experiments into two sections: (1) demonstrating that perceptual similarity metrics are sensitive to imperceptible adversarial perturbations (Section 4.1), and (2) measuring the robustness of various similarity metrics against our transferable attack (Section 4.2). In Section 4.1 we primarily show that similarity metrics are susceptible to both white-box and black-box attacks. Based

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 / 38.5%	0.023	2.9	1.7	2348 / 24.1%	84.4	56.1	0.0	1.9	1.0	4225 / 43.4%
	✗	2477	1550 / 62.6%	0.017	2.2	1.6	1202 / 48.5%	82.0	42.7	0.0	1.5	1.0	1412 / 57.0%
SSIM (Wang et al., 2004)	✓	9883	6922 / 70.0%	0.018	2.5	1.7	5297 / 53.6%	94.6	53.6	0.0	1.8	1.0	1787 / 18.1%
	✗	2344	2013 / 85.9%	0.011	1.6	1.3	1843 / 78.6%	87.3	32.0	0.0	1.3	0.8	1005 / 42.9%
WadIQaM-FR (Bosse et al., 2018)	✓	10191	8841 / 86.8%	0.006	1.0	1.0	10176 / 99.8%	69.2	4.3	0.0	0.7	0.3	3130 / 30.7%
	✗	2036	2012 / 98.8%	0.001	0.6	0.3	2035 / 99.9%	41.2	0.1	0.0	0.5	0.1	1598 / 78.5%
LPIPS(Alex) (Zhang et al., 2018b)	✓	11303	7247 / 64.1%	0.018	2.4	1.7	8806 / 77.9%	86.8	28.7	0.0	1.3	0.6	9255 / 81.9%
	✗	924	912 / 98.7%	0.004	0.9	0.7	917 / 99.2%	59.5	3.2	0.0	0.8	0.3	921 / 99.7%
LPIPS(VGG) (Zhang et al., 2018b)	✓	10976	8434 / 76.8%	0.012	1.7	1.5	9689 / 88.3%	81.6	15.6	0.0	1.0	0.5	7212 / 65.7%
	✗	1251	1244 / 99.4%	0.003	0.8	0.5	1246 / 99.6%	52.3	1.6	0.0	0.7	0.2	1219 / 97.4%
DISTS (Ding et al., 2020)	✓	11158	3043 / 27.3%	0.025	3.3	1.8	2306 / 20.7%	97.0	75.4	0.0	2.6	1.3	7416 / 66.5%
	✗	1069	795 / 74.4%	0.016	2.2	1.7	723 / 67.6%	91.9	50.0	0.0	2.0	1.3	1033 / 96.6%

on this premise, we hypothesize that all metrics are vulnerable to transferable attacks. To prove this, we attack the widely adopted LPIPS using our spatial attack FlowAdv, to create adversarial examples. We use the generated adversarial examples to benchmark the adversarial robustness of various traditional and recently proposed perceptual similarity metrics in Section 4.2. 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 Perturbations can Overturn Perceptual Similarity Judgment

Attack evaluation. Through the following study, we gather evidence that 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 three attacks FGSM, PGD, and One-pixel, were successful in flipping the rank assigned by both traditional and learned metrics in several samples. We observed for the PGD attack that none of the samples needed a perturbation[2] of more than 0.05 at the pixel-level. Therefore, for reporting the results of the FGSM attack, we use the threshold $\epsilon < 0.05$. We present the results separately for samples where the 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 30% of the samples flipped approximately. While LPIPS and WadIQaM-FR were the least robust, with approximately 80% of the samples flipped. The same conclusion can also be reached by observing ϵ (or perturbations) required to attack these metrics. Next, despite being a black-box attack, the One-pixel attack is successful in conveniently flipping rank. 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.

We present the results separately for samples where the predicted rank by the metric corresponds with the rank provided by humans. 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 rank. We posit that the early rank flipping for these samples is attributed to the fact that the distorted images in the sample, i.e., I_{other} and I_{prey} are closer to the decision boundary for the rank flip. We confirm this by calculating the absolute difference between the distances of I_{other} and I_{prey} from I_{ref} (see Appendix A.4 Table 4).

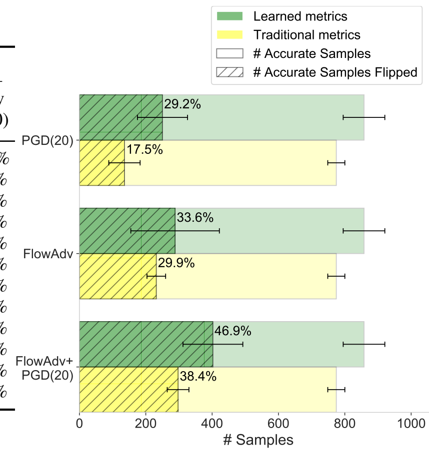
Imperceptibility. We discuss the imperceptibility of the adversarial perturbations by comparing the root mean square error (RMSE[3]) between the original and the perturbed image. As expected, the PGD attack is stronger than FGSM as it is capable of flipping a significant number of samples with lesser adversarial perturbations. As reported in Table 2, for the PGD attack, a good portion of the

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

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

Table 3: Transferable adversarial attacks on perceptual similarity metrics. The adversarial examples were generated by attacking LPIPS(AlexNet) via FlowAdv. In total, there are 1061 samples. Next, we attacked LPIPS(AlexNet) using PGD(20). Then, we combined FlowAdv+PGD(20) by perturbing the FlowAdv generated images with PGD(20). 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. On the right, the visualization compares traditional metrics (L2, SSIM, and FSIMc) versus traditional metrics (WaDIQaM-FR, GTI-CNN, LPIPS, DISTS, E-LPIPS, Watson-DFT, and PIM).

Network	# Accurate Samples	# Accurate Samples Flipped		
		PGD(20)	FlowAdv	FlowAdv +PGD(20)
L2	790 / 74%	82 / 10%	207 / 26%	253 / 32 %
SSIM (Wang et al., 2004)	795 / 75%	197 / 25%	217 / 27%	331 / 42%
FSIMc (Zhang et al., 2011)	738 / 70%	128 / 17%	271 / 37%	308 / 42%
WaDIQaM-FR (Bosse et al., 2018)	791 / 75%	84 / 11%	134 / 17%	207 / 26%
GTI-CNN (Ma et al., 2018)	730 / 69%	162 / 22%	310 / 42%	323 / 44%
LPIPS(Squz.) (Zhang et al., 2018b)	939 / 89%	287 / 31%	201 / 21%	452 / 48%
LPIPS(VGG) (Zhang et al., 2018b)	851 / 80%	319 / 37%	164 / 19%	428 / 50%
DISTS (Ding et al., 2020)	884 / 83%	244 / 28%	189 / 21%	379 / 43%
E-LPIPS (Kettunen et al., 2019b)	890 / 84%	275 / 31%	382 / 43%	475 / 53%
Watson-DFT (Czolbe et al., 2020)	821 / 77%	247 / 30%	216 / 26%	363 / 44%
PIM-1 Bhardwaj et al. (2020)	909 / 86%	310 / 34%	501 / 55%	483 / 53%
PIM-5 Bhardwaj et al. (2020)	906 / 85%	325 / 36%	498 / 55%	511 / 56%



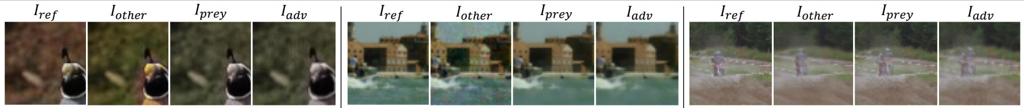
adversarial image (I_{adv}) has $\epsilon < 0.01$, while for FGSM, the amount of pixel perturbation all over the image is a constant ϵ value which moreover is higher for a successful attack. Consequently, on average, the I_{adv} generated via PGD has lower RMSE and a higher PSNR (see Appendix A.5 Table 5) 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 in the I_{adv} generated via the One-pixel attack, which we consider suitably imperceptible.

4.2 Transferable adversarial attacks on perceptual similarity metrics

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 compare against such transferable adversarial examples. Specifically, we transfer the FlowAdv attack in Section 3.1 on LPIPS(AlexNet) to other metrics. We chose LPIPS(AlexNet) as it is widely adopted. Furthermore, we combine the FlowAdv attack with PGD to increase the transferability of the adversarial examples to other metrics. In this experiment, we only consider samples for which the metrics and the human opinions agree on their rankings.

FlowAdv. As shown in Figure 5, our spatial attack, FlowAdv, has the capability of attacking high-level image features. As a white-box attack on LPIPS(AlexNet), out of the total 12,227 samples, FlowAdv was able to flip judgment on 5703 samples with a mean RMSE of 0.064. Because we need high imperceptibility, we remove samples with RMSE > 0.05 and are left with 1924 samples. We then perform a visual sanity check and remove some more with ambiguity, keeping only strictly imperceptible samples. In the end, we have 1061 samples, with a mean RMSE of 0.034, which we transfer to other metrics as a black-box attack. As reported in Table 3, all metrics are prone to the attack. Surprisingly, WaDIQaM-FR (Bosse et al., 2018) has the most robustness, while the recently proposed PIM (Bhardwaj et al., 2020) metric that was found robust to small imperceptible shifts is highly susceptible to this attack. Although, PIM is 10% more accurate than WaDIQaM-FR. Finally, we saw that, on average, the learned metrics are more correlated with human opinions, but traditional metrics exhibit more robustness to the imperceptible transferable FlowAdv adversarial perturbations.

PGD(20). We now attack the original 1061 selected samples with the l_∞-bounded attack, PGD. As shown in Section 4.1, perturbations generated via PGD have low perceptibility; hence, we cre-



	I_{other}	I_{prey}	I_{adv}	I_{other}	I_{prey}	I_{adv}	I_{other}	I_{prey}	I_{adv}
LPIPS(AlexNet)	0.137	0.109	0.351	0.039	0.024	0.258	0.118	0.045	0.584
l_2	0.008	0.005	0.009	0.003	0.001	0.003	0.003	0.002	0.003
SSIM	0.033	0.024	0.058	0.107	0.054	0.099	0.116	0.074	0.135
FSIMc	0.000	0.000	0.001	0.001	0.002	0.003	0.003	0.002	0.004
WaDIQaM-FR	1.335	1.188	1.223	1.163	1.116	1.085	1.166	1.163	1.166
LPIPS(Squeeze)	0.157	0.135	0.159	0.055	0.056	0.092	0.109	0.032	0.118
LPIPS(VGG)	0.205	0.179	0.242	0.137	0.130	0.161	0.176	0.087	0.221
DISTS	0.085	0.086	0.133	0.123	0.097	0.134	0.141	0.082	0.142
E-LPIPS	0.011	0.009	0.015	0.009	0.007	0.013	0.008	0.003	0.010
Watson-DFT	1218	1190	1913	2037	971	1543	1525	1154	1578
PIM-1	1.139	2.130	2.744	1.926	1.183	2.982	2.409	0.417	4.096
PIM-5	13.170	25.230	31.419	14.906	13.517	29.668	23.225	4.804	39.530

Figure 6: Adversarial examples (I_{adv}) generated via FlowAdv+PGD(20) to attack LPIPS(AlexNet) transfer successfully to most perceptual similarity metrics. A successful attack is marked in red. For the above samples, the RMSE between I_{prey} and I_{adv} is 0.050, 0.037, and 0.038 (left to right).

ate adversarial samples using the PGD attack. In FlowAdv, 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 20 for each sample to guarantee the transferability of perturbations. We call this transferable attack PGD(20), and the mean RMSE of the adversarial images generated is only 0.014. The metrics SSIM and WaDIQaM-FR are the most robust to the transferable PGD(20) attack, as reported in Table 3.

Combining FlowAdv and PGD(20). FlowAdv and PGD are orthogonal approaches as PGD (l_∞-bounded attack) manipulates the intensity of individual pixels while FlowAdv (spatial attack) manipulates the location of the pixels. We now combine the two by attacking the samples generated via FlowAdv with PGD(20). The mean RMSE of the generated adversarial images is 0.038, just 0.004 higher than images generated via FlowAdv. As reported in Table 3, the increase in severity of the adversarial perturbations in FlowAdv+PGD(20) 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 FlowAdv, can be transferred to other metrics. Furthermore, combining FlowAdv (spatial attack) with PGD (l_∞-bounded attack) increases their transferability. We showcase a few examples in Figure 6. Our investigations also show that although more accurate, learned metrics may not be more robust than traditional ones. 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 introduced a spatial attack, FlowAdv, that was 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.

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.

Ruohan Gao, Bo Xiong, and Kristen Grauman. Im2flow: Motion hallucination from static images for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5937–5947, 2018.

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.

Aleksander Holynski, Brian L Curless, Steven M Seitz, and Richard Szeliski. Animating pictures with eulerian motion fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5810–5819, 2021.

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, and koray kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, volume 28, 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.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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

Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *IEEE Conferene on Computer Vision and Pattern Recognition*, 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.

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.

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.

Silvia L. Pintea, Jan C. van Gemert, and Arnold W. M. Smeulders. Déjà vu: Motion prediction in static images. In *European Conference on Computer Vision*, pp. 172–187, 2014.

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.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pp. 1530–1538, 2015.

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.

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.

Jacob Walker, Abhinav Gupta, and Martial Hebert. Dense optical flow prediction from a static image. In *IEEE International Conference on Computer Vision*, pp. 2443–2451, 2015.

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.

Daan Wierstra, Tom Schaul, Jan Peters, and Juergen Schmidhuber. Natural evolution strategies. In *IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, pp. 3381–3387. IEEE, 2008.

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.

Yanchao Yang and Stefano Soatto. Conditional prior networks for optical flow. In *European Conference on Computer Vision*, pp. 271–287, 2018.

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.

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 APPENDIX

A.1 TWO-ALTERNATIVE FORCED CHOICE (2AFC) SCORE

Here we explain how the 2AFC score is calculated. Zhang et al. (2018b) used the 2AFC score to decide which metric is more correlated with human judgment on image similarity. We follow Zhang et al. (2018b) for the 2AFC score calculation in Table 1. Considering I_0 and I_1 as the two images being compared with each other with respect to a reference I_{ref}, the authors collected 5 human responses for each such sample in the BAPPS validation dataset. Now, if p humans voted for I_0, and $1-p$ human voted for I_1, a metric's 2AFC score for that sample would be computed as follows:

$$(s_0 < s_1) \times (1-p) + (s_1 < s_0) \times p + (s_1 == s_0) \times 0.5 \tag{11}$$

where the similarity score $s_i = f(I_i, I_{ref})$, and a smaller value for s_i indicates more similarity. Hence, if 4 humans voted for I_0 and 1 human voted for I_1, and the metric predicts that I_0 is more similar to I_{ref}, then the metric would get a score of 80%. The final 2AFC score is an average over all samples.

A.2 FGSM ATTACK ON SIMILARITY METRICS

We explain the FGSM in Algorithm 3.

Algorithm 3: FGSM attack on Similarity Metrics

Input: I_1, I_2, I_{ref}, metric f, max_ϵ (0.05)
Output: Least ϵ value which led to rank flip

$s_0 = f(I_{ref}, I_0)$
$s_1 = f(I_{ref}, I_1)$
// If I_0 is more similar to I_{ref} then $rank$ is 0 else 1
$rank = int(s_0 > s_1)$ // smaller $s_i \equiv$ more similar
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 ϵ

A.3 ONE-PIXEL ATTACK ON SIMILARITY METRIC

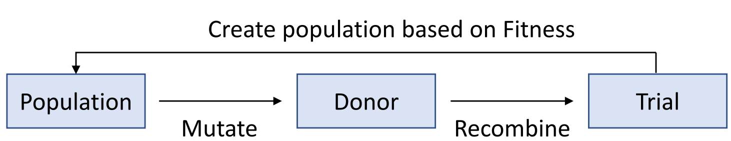


Figure 7: Stages in Differential Evolution.

The steps involved in the differential evolution algorithm are shown in Figure 7 and described as follows:

1. The initial population X contains vectors X_i (for simplicity we refer it as "x" in the main text) having pixel's index ($x\_position$, $y\_position$), and perturbation values for the 3 channels r, g, and b.
2. For mutation the donor vector (D_i) is generated using three random vectors X_{r_1}, X_{r_2}, and X_{r_3} as follows:
$$D_i = X_{r_1} + factor * (X_{r_2} - X_{r_3}) \tag{12}$$
where $factor$ is a scaling-factor and r_1, r_2, and r_3 are random indices such that $r_1 \neq r_2 \neq r_3 \neq i$. Therefore, the minimum population size for differential evolution is 4.
3. For the recombination step, we apply a crossover by updating index j of the vector X_i to create the trial vector T_i. It is described as follows:
$$T_{ij} = \begin{cases} D_{ij}, & \text{if } r < p_c \text{ or } j = \delta \\ X_{ij}, & \text{if } r > p_c \text{ and } j \neq \delta. \end{cases} \tag{13}$$
where D_{ij} is the index j of donor vector D_i, r is a random value from [0,1], p_c is cross-over probability, and δ is a randomly selected index ensuring that at least one index is from the donor vector.
4. The fitness of the trial vectors T is decided by computing the scores of the sample as mentioned in ***PerturbImage*** function in Algorithm 4. The trial vector T_i replaces the original vector X_i if its score is better. This way, the population is re-generated, and the process starts all over again.

The attack terminates when one of the trial vectors T_i (or I_{adv}) satisfies the condition for rank flip, i.e. $s_{adv} > s_{other}$.

Algorithm 4: One-pixel attack on LPIPS

Input: I_0, I_1, I_{ref}, trained LPIPS model f
Output: x_position, y_position, r, g, b of the perturbation

Function PerturbImage(I_{prey}, I_{ref}, s_{other}, T) **:**
    population_size = len(T) // Trial vector T
    I_{adv} = repeat(I_{prey}, population_size) // repeat I_{prey} to create a batch
    factor = 0.1
    for *$i \leftarrow$ 1 to population_size* **do**
        // Apply perturbation to each I^i_{adv}
        x_position, y_position, r, g, b = T_i
        I_{adv}[i, 0, x_position, y_position] = (r/255 - 0.5)/factor
        I_{adv}[i, 1, x_position, y_position] = (g/255 - 0.5)/factor
        I_{adv}[i, 2, x_position, y_position] = (b/255 - 0.5)/factor
        $s^i_{adv} = f(I_{ref}, I_{adv})$ // compute scores of the perturbed images
        $s^i = 1 - (s^i_{adv}/(s^i_{adv} + s^i_{other}))$ // Trial vector fitness score
        // If score s^i of T_i is better than the score of X_i
        // then T_i replaces X_i during differential evolution
    return s *// scores*

$s_0 = f(I_{ref}, I_0)$
$s_1 = f(I_{ref}, I_1)$
// If I_0 is more similar to I_{ref} then $rank$ is 0 else 1
$rank = int(s_0 > s_1)$ // smaller $s_i \equiv$ more similar
if $rank = 1$ **then**
    I_{prey} = I_1
    s_{other} = s_0
else
    I_{prey} = I_0
    s_{other} = s_1
successfull_vector_X_i = differential_evolution(func=PerturbImage,args=(I_{prey}, I_{ref}, s_{other}))
// The differential_evolution algorithm optimizes population X
// to find optimal X^*_i (see Figure 7 and steps in Appendix A.3)
x_position, y_position, r, g, b = successfull_vector_X_i
return *x_position, y_position, r, g, b*

A.4 SAMPLES WHERE RANK PREDICTED BY METRIC $\neq$ RANK ASSIGNED BY HUMANS

In Table 2 we observed that it was easier to flip rank when the rank predicted by metric is not the same as rank assigned by humans. We believe that such samples lie closer to the decision boundary. 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 4, the $abs(s_0 - s_1)$ for these samples is much lesser than samples where rank predicted by metric = rank assigned by humans. This results indicates that samples where rank predicted by metric $\neq$ rank assigned by humans, lie closer to the decision boundary causing them to flip earlier.

Table 4: 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	$abs(s_0 - s_1)$
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

A.5 IMPERCEPTIBLITY OF ADVERSARIAL PERTURBATIONS: FGSM VERSUS PGD

Table 5: Comparing PSNR of adversarial images generated via FGSM versus PGD. For adversarial images generated via FGSM, ϵ is < 0.05. Higher PSNR of PGD examples shows that adversarial perturbations are less perceptible. Furthermore, we also confirmed this through visual comparison.

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