Towards Large yet Imperceptible Adversarial Image Perturbations with Perceptual Color Distance

Zhengyu Zhao, Zhuoran Liu, Martha Larson
Radboud University, Nijmegen, Netherlands
{z.zhao, z.liu, m.larson}@cs.ru.nl

Abstract

The success of image perturbations that are designed to fool image classifier is assessed in terms of both adversarial effect and visual imperceptibility. The conventional assumption on imperceptibility is that perturbations should strive for tight L_p-norm bounds in RGB space. In this work, we drop this assumption by pursuing an approach that exploits human color perception, and more specifically, minimizing perturbation size with respect to perceptual color distance. Our first approach, Perceptual Color distance C&W (PerC-C&W), extends the widely-used C&W approach and produces larger RGB perturbations. PerC-C&W is able to maintain adversarial strength, while contributing to imperceptibility. Our second approach, Perceptual Color distance Alternating Loss (PerC-AL), achieves the same outcome, but does so more efficiently by alternating between the classification loss and perceptual color difference when updating perturbations. Experimental evaluation shows PerC approaches outperform conventional L_p approaches in terms of robustness and transferability, and also demonstrates that the PerC distance can provide added value on top of existing structure-based methods to creating image perturbations.

1. Introduction

Research on creating adversarial examples for deep visual classifiers has focused on perturbations that cause misclassification while being *imperceptible* to the human eye [7, 43, 49]. Larger image perturbations are known to improve adversarial strength (i.e., the ability to fool a classifier), but are also associated with visually noticeable changes in the image. A commonly agreed-upon assumption is that tight L_p-norm constraints on the size of adversarial perturbations in RGB space are a good guarantee of imperceptibility. Evaluation of adversarial examples has conventionally followed this assumption, considering perturbations with smaller L_p norms to be better (e.g., L_∞ [7, 19, 29], L_2 [7, 41, 49] and L_0 [7, 43]). Keeping with this assumption, defense approaches are designed to be effective against adversarial perturbations under a specific L_p bound [10, 39, 51, 54]. Our research is motivated by the importance of questioning the necessity of small RGB perturbations for imperceptibility.

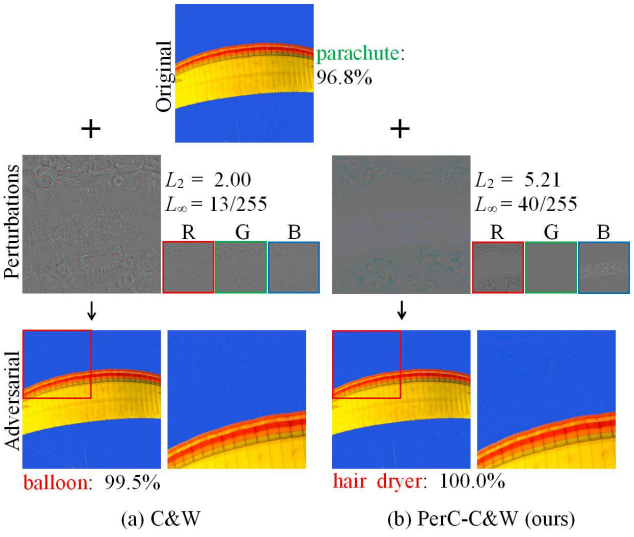

Figure 1: Comparison of (a) C&W [7] with (b) our PerC-C&W. Perceptual color (PerC) distance allows larger RGB perturbations (cf. L_2 and L_∞ norm in middle row), while also contributing to imperceptibility (bottom row). (Setting: untargeted with $\kappa = 40$; classifier Inception v3.)

In this work, we propose to create adversarial examples by perturbing images with respect to perceptual color (PerC) distance. Using PerC distance makes it possible to move away from the assumption that it is necessary to tightly constrain the L_p norm of the perturbations in RGB space. Fig. 1 illustrates the difference between C&W [7], a well-known approach that perturbs with respect to an L_p norm in RGB space, and our own extension, PerC-C&W, which perturbs with respect to a perceptual color dis-

tance. PerC perturbations are less perceptible, especially in smooth regions of saturated color (cf. Fig. 1 in bottom row). Also, they are distributed strategically over the RGB color channels (cf. downsized perturbation images in the middle row). PerC distance effectively allows us to hide large perturbations in RGB space, in a way not readily noticeable to the human eye. Our PerC-based approaches can increase the L_p norm substantially (cf. Fig. 1, L_2 and L_∞ in middle row), leading to a strong adversarial effect that maintains imperceptibility.

Fig. 2 motivates the use of perceptual color distance for creating adversarial images. Here, we have taken a solid color image (left) and added the same perturbations to the green channel (middle) and to the blue channel (right). Although both RGB channels were perturbed identically, the perturbations are only visible in the green channel. The reason is that color as it is perceived by the human eye does not change uniformly over distance in RGB space. Relatively small perturbations in RGB space may correspond to large difference in perceptual color space. Conversely, relatively large changes in RGB space may remain unnoticeable if they lead to small perceived color difference.

Our work is in line with a growing awareness in the literature on adversarial examples that the difference between two images as measured by an L_p norm in RGB space is actually quite poorly aligned with human perception [45]. Building on this observation, researchers have attempted to address imperceptibility by exploiting similarity defined with respect to semantics [15, 17, 23, 24, 46] or structural information [11, 16, 37, 55, 59] in the image. However, little work on adversarial examples has questioned the wisdom of optimizing perturbations with respect to distance in RGB space. The exceptions are a handful of approaches that have proposed allowing only luminance change when perturbing pixels [11, 16]. The approach that is closest to our own is [3], which perturbs in CIELAB color space, but carries out no investigation of the potential and limitations of the idea. Our work is distinct from this initial effort because we use a more accurate polar form (known as CIELCH) of the CIELAB color space, and more importantly, use an actual perceptual color distance. The distance is CIEDE2000 [1, 38], and will be discussed in detail in Section 2. To our knowledge, ours is the first work that proposes optimizing adversarial image perturbations directly with respect to a perceptual color distance.

In order to fully appreciate our proposal, it is necessary to understand two key aspects. First, we do not claim that PerC approaches will always yield dramatically less perceptible perturbations than conventional RGB approaches. For cases in which the perturbations are small, the difference may not be so great. However, we find that there are two cases in which PerC approaches are particularly important. First, our experimental results (see Section 5.2.2) show that as we attempt to create adversarial images that are misclassified with high confidence (i.e., high-confidence adversarial examples), it becomes important to perturb with respect to perceptual color distance. Second, we demonstrate that the effect of PerC approaches is additive and can be used in combination with existing structural approaches to improve imperceptibility.

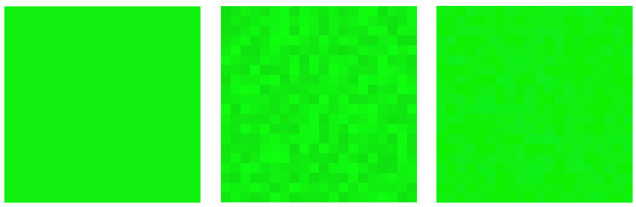

Figure 2: Left: Original image (a 20×20 8-bit RGB image patch with color (15,240,15)). Middle: Image perturbed by adding noise in the G channel, sampled from a uniform distribution in the range [-15,15]. Right: Image perturbed by adding the identical noise, but in the B channel. The B-channel perturbations are imperceptible (best viewed on screen).

The contributions of this paper are as follows:

- An in-depth study of the use of perceptual color (PerC) distance to hide large RGB perturbations in images.
- PerC-C&W: a method for creating adversarial images that introduces perceptual color distance into the joint optimization of C&W.
- PerC-AL: an efficient method that optimizes alternating loss (AL) functions, switching between classification loss and perceptual color difference.
- Experimental validation demonstrating that PerC perturbations in high-confidence settings yield more robust and transferable adversarial examples, without sacrificing imperceptibility.
- Experimental results showing that PerC perturbations can be used in combination with structural information for further improvement of imperceptibility.

The code, which also includes a differentiable solution compatible with PyTorch's autograd to efficiently implement perceptual color distance (CIEDE2000), is available at https://github.com/ZhengyuZhao/PerC-Adversarial.

2. Background on Perceptual Color Distance

Conventionally, computer vision research has intensively explored color and human perception, but has paid surprisingly little attention to distance in perceptual color spaces. Here, we mention some key points about color in computer vision history. Early on, research focused on intensity-based descriptors, which then evolved to also capture color information. Unsurprisingly, color boosted the performance

of object and scene recognition [26, 52] and semantic segmentation [8]. Researchers extracted descriptors from opponent color spaces, most notably HSV and CIELAB, which separate luminance and chrominance. Most recently, color is attracting more attention in the area of image synthesis. Notable examples, such as style transfer [18] and cross-domain image generation [50], find that color plays an important role in preserving the look of an image. In general, we observe that until now the focus has been on the color space itself, and not on color distance, which we explore here.

The perceptual color distance that we use is CIEDE2000, which is the latest ΔE standard formula developed by the CIE (International Commission on Illumination). CIEDE2000 refined the definition of previous editions by adding five corrections, and has been experimentally demonstrated to better align with human visual perception [1, 38]. Specifically, the pixel-wise perceptual color distance can be calculated as:

$$\Delta E_{00} = \sqrt{(\frac{\Delta L'}{k_L S_L})^2 + (\frac{\Delta C'}{k_C S_C})^2 + (\frac{\Delta H'}{k_H S_H})^2 + \Delta R},$$
$$\Delta R = R_T(\frac{\Delta C'}{k_C S_C})(\frac{\Delta H'}{k_H S_H}), \tag{1}$$

where $\Delta L'$, $\Delta C'$, $\Delta H'$ denotes the distance between pixel values of the three channels, L (lightness), C (chroma) and H (hue) in the CIELCH space, and ΔR is an interactive term between chroma and hue differences [38]. The weighting functions S_L, S_C, S_H and R_T are determined based on large-scale human studies and act as compensations to better simulate human color perception. The k_L, k_C and k_H are usually unity for the application of graphic arts. Detailed definitions of all the parameters and relevant explanations can be found in [38]. We note that it is also possible to use an L_p norm to measure distance in CIELAB space. However, this distance is not as close to human perceptual distance as CIEDE2000 is.

We point out that a limited amount of previous research has also adopted CIEDE2000. However, the goal has been to evaluate the color similarity of image pairs. Examples of such research include work on image quality assessment [58] and image super-resolution [34]. In contrast, in our work we use CIEDE2000 directly for optimization with back propagation and not only for evaluation.

3. Related work

In this section, we cover the existing literature, which focuses on creating L_p norm-bounded adversarial examples, and we also mention recent approaches that attempt to move beyond L_p norms. We preface our discussion with a short definition of an 'adversary', i.e., an approach that generates an adversarial image example. Given a classifier $f(\boldsymbol{x}) : \boldsymbol{x} \rightarrow y$ that predicts a label y for an image $\boldsymbol{x}$, the adversary attempts to induce a misclassification by modifying the original $\boldsymbol{x}$ to create a new $\boldsymbol{x}'$. In the untargeted setting, the adversary is successful if the image is classified into an arbitrary class other than y, i.e., meets the condition $f(\boldsymbol{x}') \neq y$. In the targeted setting, the adversary must ensure that the image is classified into a class with a predefined label t, i.e., meets the condition $f(\boldsymbol{x}') = t$. The untargeted case is generally recognized to be less challenging than the targeted case [7].

3.1. L_p norm-bounded Adversarial Examples

Typically, adversaries [7, 19, 29, 41, 43, 44, 49] create an adversarial image, x', by adding a perturbation vector $\boldsymbol{\delta} \in \mathbb{R}^n$ that is constrained by an L_p norm to the original image, $\boldsymbol{x}$. The first L_p norm-bounded approach [49] optimized an objective combining the classification loss and the L_2 norm of the perturbations, balanced by a constant λ. Formally, the solution is expressed as:

$$\underset{\boldsymbol{\delta}}{\text{minimize}}\ \lambda\|\boldsymbol{\delta}\|_2 - J(\boldsymbol{x}', y),\ \text{s.t.}\ \boldsymbol{x}' \in [0,1]^n, \tag{2}$$

where $J(\boldsymbol{x}', y)$ is the cross-entropy loss w.r.t. $\boldsymbol{x}'$. The authors of [49] solved the problem by using box-constrained L-BFGS (Limited memory Broyden-Fletcher-Goldfarb-Shanno) method [33].

The C&W method [7] improves on [49] by introducing a new variable using the tanh function to eliminate the box constraint. Additionally, it introduces a more sophisticated objective function that optimizes differences between the logits, Z, which are output before the softmax layer. This can be formulated as:

$$\begin{aligned}
&\underset{\boldsymbol{w}}{\text{minimize}}\ \ \|\boldsymbol{x}' - \boldsymbol{x}\|_2^2 + \lambda f(\boldsymbol{x}'),\\
&\text{where}\ \ f(\boldsymbol{x}') = \max(\max\{Z(\boldsymbol{x}')_i : i \neq t\} - Z(\boldsymbol{x}')_t, -\kappa),\\
&\text{and}\ \ \boldsymbol{x}' = \frac{1}{2}(\tanh(\operatorname{arctanh}(\boldsymbol{x}) + \boldsymbol{w}) + 1),
\end{aligned} \tag{3}$$

where $\boldsymbol{w}$ is the new variable and $Z(x')_i$ denotes the logit with respect to the i-th class. In an untargeted setting, the definition of f is modified to:

$$f(\boldsymbol{x}') = \max(Z(\boldsymbol{x}')_y - \max\{Z(\boldsymbol{x}')_i : i \neq y\}, -\kappa). \tag{4}$$

The parameter κ controls the confidence level of the misclassification. The first approach that we propose, PerC-C&W, is built on C&W. In our experiments, we will vary κ in order to assess the ability of an adversary to create strong adversarial images, i.e., images that are misclassified with high confidence.

Due to the need for line search in order to find the optimal constant, λ, such an optimization approach is inevitably time-consuming. For this reason, [19, 29, 44] propose a

more efficient solution that does not impose a penalty during optimization. Instead, respect of the norm constraint is ensured by projecting perturbations onto an ϵ-sphere around the original image. Specifically, the fast gradient sign method (FGSM) [19] was first proposed to achieve adversarial effect with only one step, formulated as:

$$\boldsymbol{x}' = \boldsymbol{x} + \epsilon \cdot \mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y)), \tag{5}$$

where the perturbation size is implicitly constrained by specifying a small ϵ.

Subsequently, an extension of this method referred to as I-FGSM [29] was introduced for leveraging finer gradient information by iteratively updating the perturbations with a smaller step size α:

$$\boldsymbol{x}'_0 = \boldsymbol{x}, \;\; \boldsymbol{x}'_k = \boldsymbol{x}'_{k-1} + \alpha \cdot \; \mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}'_{k-1}, y)), \tag{6}$$

where the intermediate perturbed image $\boldsymbol{x}'_k$ is projected onto a ϵ-sphere around the original $\boldsymbol{x}$, to satisfy the L_∞-norm constraint. Note that I-FGSM constrains only the maximum change of individual image coordinates without considering the image-level accumulated difference. For this reason, I-FGSM yields poor imperceptibility, especially in high-confidence settings (cf. Fig. 3).

A generalization of I-FGSM to the L_2 norm can be achieved by changing the sign operation in the updating to:

$$\frac{\nabla_{\boldsymbol{x}} J(\boldsymbol{x}'_{k-1}, y)}{\|\nabla_{\boldsymbol{x}} J(\boldsymbol{x}'_{k-1}, y)\|_2}, \tag{7}$$

where the projection is implemented by:

$$\boldsymbol{x}'_k = \boldsymbol{x} + \epsilon \frac{\boldsymbol{x}'_k - \boldsymbol{x}}{\|\boldsymbol{x}'_k - \boldsymbol{x}\|_2}. \tag{8}$$

A recent method called the Decoupled Direction and Norm (DDN) [44], which is based on the L_2 norm-based I-FGSM, yielded the best performance (smallest L_2 norm) in the untargeted track of NIPS 2018 Adversarial Vision Challenge [5], with substantially fewer iterations than the conventional C&W. In DNN, the ϵ is designed to be adjustable in each iteration based on whether the perturbed image is adversarial or not, leading to a finer search for the minimal norm. Our second approach, PerC-AL, follows a similar strategy as DDN to improve efficiency by decoupling the joint optimization.

3.2. Adversarial examples beyond L_p norms

Our work is part of the current movement away from tight L_p norms and towards conceptualization of image similarity in terms of semantics or perceptual properties. Research that defines similarity in terms of semantics, requires the adversarial image to have the same content as the original image from the point of view of the human viewer. Some of the first work in this direction has explored geometric transformation [15, 56], global color shift [2, 4, 23, 31], and image filters [9].

Such approaches are interesting, but we do not pursue them here because they tend to be limited in their adversarial strength, due to the restricted size of the search space for possible adversarial image transformations.

Research that investigates similarity with respect to texture and structure [11, 16, 37, 55, 59], has focused on hiding perturbations in image regions with visual variation. Such hiding can be achieved by either using existing structure-aware metrics [16, 55], such as structural similarity (SSIM) [53] and Wasserstein distance [25], or directly allowing more perturbations in the high-variance image regions [11, 37, 59]. All of these approaches share a common challenge: They have difficulties in dealing with smooth regions (e.g., sky, ground and artificial objects), which appear frequently in images taken in commonly occurring real-world settings (referred to as *natural images*). In contrast, our PerC perturbations are applicable in smooth regions in the case of saturated color. Our experiments show that PerC perturbations can be combined productively with a structure-based approach (see Section 5.5).

4. Proposed approaches

In this section, we present two approaches to using perceptual color (PerC) distance for adversarial image perturbations. We focus on image-level accumulated perceptual color difference, i.e., the L_2 norm of the color distance vector, in which each component represents the perceptual color distance (ΔE_{00} in Eq. (1)) calculated for the corresponding image pixel.

4.1. Perceptual color distance penalty (PerC-C&W)

Our first approach, PerC-C&W, adopts the joint optimization of the well-known C&W, but replaces the original penalty on the L_2 norm with a new one based on perceptual color difference. It can be formally expressed as:

$$\underset{\boldsymbol{w}}{\text{minimize}} \;\; \|\Delta E_{00}(\boldsymbol{x}, \boldsymbol{x}')\|_2 + \lambda f(\boldsymbol{x}'), \tag{9}$$

where $\boldsymbol{w}$ is the new introduced variable as in the Eq. (3) of C&W. Like the original C&W, the optimization problem is solved by binary search over the constant λ. By using the gradient information from perceptual color difference, perturbation updating is translated into a perceptually uniform color space. Large RGB perturbations, which have a strong adversarial effect, remain hidden from the human eye, as will be shown in Section 5.

4.2. Perceptual color distance alternating loss (PerC-AL)

Although, Eq. 9 enjoys a concise expression, the joint optimization of PerC-C&W faces difficulties in practice.

Algorithm 1 Perceptual Color Distance Alternating Loss (PerC-AL)

Input:
$\boldsymbol{x}$: original image, t: target label, K: number of iterations
α_l: step size in minimizing classification loss
α_c: step size in minimizing perceptual color difference

Output: $\boldsymbol{x}'$: adversarial image

Initialize $\boldsymbol{x}'_0 \leftarrow \boldsymbol{x}$, $\boldsymbol{\delta}_0 \leftarrow \mathbf{0}$
for $k \leftarrow 1$ to K **do**
  if $\boldsymbol{x}'_{k-1}$ is not adversarial **then**
    $\boldsymbol{g} \leftarrow -\nabla_{\boldsymbol{x}} J(\boldsymbol{x}'_{k-1}, t)$
    $\boldsymbol{g} \leftarrow \alpha_l \cdot \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$
    $\boldsymbol{\delta}_k \leftarrow \boldsymbol{\delta}_{k-1} + \boldsymbol{g}$ ▷ Update $\boldsymbol{\delta}$ in the direction of $\boldsymbol{g}$
  else
    $C_2 \leftarrow -\|\Delta E_{00}(\boldsymbol{x}, \boldsymbol{x}'_{k-1})\|_2$
    $\boldsymbol{g}_c \leftarrow \nabla_{\boldsymbol{x}} C_2$
    $\boldsymbol{g}_c \leftarrow \alpha_c \cdot \frac{\boldsymbol{g}_c}{\|\boldsymbol{g}_c\|_2}$
    $\boldsymbol{\delta}_k \leftarrow \boldsymbol{\delta}_{k-1} + \boldsymbol{g}_c$ ▷ Update $\boldsymbol{\delta}$ in the direction of $\boldsymbol{g}_c$
  end if
  $\boldsymbol{x}'_k \leftarrow \text{clip}(\boldsymbol{x} + \boldsymbol{\delta}_k, 0, 1)$
  $\boldsymbol{x}'_k \leftarrow \text{quantize}(\boldsymbol{x}'_k)$ ▷ Ensure $\boldsymbol{x}'_k$ is valid
end for
return $\boldsymbol{x}' \leftarrow \boldsymbol{x}'_k$ that is adversarial and has smallest C_2

Adversarial training [29, 39], for example, presents challenges. The reason is that PerC-C&W requires time-consuming binary search in order to find an optimal λ, which normally varies substantially among different images [44]. To address the inefficiency, we propose PerC-AL, which decouples the joint optimization by alternately updating the perturbations with respect to either classification loss or perceptual color difference. Our strategy is inspired by DDN, which is basically a projected gradient descent (PGD) method with a dynamic L_2-norm bound. However, PerC-AL goes beyond this idea to alternate two gradient descents.

The full PerC-AL method is described in Algorithm 1. We start from an original image $\boldsymbol{x}$ with the perturbation $\boldsymbol{\delta}$ initialized as $\mathbf{0}$, and iteratively update it to create an adversarial image. In each iteration, the perturbation is either enlarged to achieve stronger adversarial effect based on the gradients from the classification loss, or shrunk to minimize perceptual color differences. These two operations are alternated based on whether the intermediate perturbed image $\boldsymbol{x}'_k$ is adversarial or not, leading to a finer search of a minimal perceptual color difference by repeatedly crossing the decision boundary. To ensure the adversarial image is valid, the output is clipped into the range [0,1] and quantized into 255 levels (corresponding to 8-bit image encoding).

5. Experiments

In this section, we first provide a picture of the differences between RGB and PerC approaches (Section 5.2). Then, we carry out experiments that compare different approaches in terms of robustness (Section 5.3) and transferability (Section 5.4) by considering the case of high-confidence adversarial examples. Finally, in Section 5.5, we show that structural information can be elegantly integrated into our efficient decoupled approach, PerC-AL, for further improvement in the imperceptibility of images that contain areas with rich visual variation.

5.1. Experimental setup

Dataset and Networks. Following recent work [13, 56, 59], we conduct our experiments on the development set (1000 RGB natural images with the size of 299 × 299) of the ImageNet-Compatible dataset[1]. This dataset was introduced by the NIPS 2017 Competition on Adversarial Attacks and Defenses [30] and consists of 6000 images labeled with 1000 ImageNet classes. We choose this dataset because we would like to study imperceptibility under real-world conditions. In contrast, some other work [11, 37] on addressing imperceptibility mainly focuses on the tiny images from MNIST [32] and CIFAR-10 [28]. As in the competition, the Inception V3 [48] model pre-trained on ImageNet is used as the target classifier.

Baselines. Three well-known baselines, I-FGSM [29], C&W [7], and the state-of-the-art DDN [44], are compared with our approaches. Among them, I-FGSM targets minimum L_∞ norm, while C&W and DDN target minimum L_2 norm. Note that I-FGSM is not designed for imperceptibility, but we consider it here for completeness.

Parameters. I-FGSM is repeated multiple rounds with increased L_∞-norm bound, where in each round, a large enough iteration budget (100 in our implementation) is specified with the step size $\alpha = 1/255$.

C&W and PerC-C&W use the Adam optimizer [27] with a learning rate of 0.01 for updating the perturbations. We impose a budget on the number of search steps used to find the optimal λ. The initialization of λ is particularly important for small budgets. We perform grid search for the initialization value of λ over the range [0.01, 0.1, 1, 10, 100], and adopt the value that yields the smallest average perturbation size. The selected initialization values are given in the supplementary material.

For DDN and PerC-AL, we decrease the step size (α in DDN and α_l in PerC-AL) that is used for updating the perturbations with respect to the classification loss from 1 to 0.01 with cosine annealing. The L_2-norm constraint ϵ in DDN is initialized to 1 and adjusted iteratively by $\gamma = 0.05$,

[1] https://github.com/tensorflow/cleverhans/tree/master/examples/nips17_adversarial_competition/dataset.

Approach	Budget	Success Rate (%)	Perturbation Size $\overline{L_2}$	$\overline{L_\infty}$	$\overline{C_2}$
I-FGSM [29]	-	100.0	2.51	1.59	317.96
C&W [7]	3×100	100.0	1.32	8.84	159.85
	5×200	100.0	1.09	8.20	132.86
	9×1000	100.0	0.92	8.45	114.36
PerC-C&W (ours)	3×100	100.0	2.77	14.29	150.44
	5×200	100.0	1.48	12.06	83.93
	9×1000	100.0	1.22	15.57	67.79
DDN [44]	100	100.0	1.00	7.84	136.11
	300	100.0	0.88	7.58	120.12
	1000	100.0	0.82	7.62	111.65
PerC-AL (ours)	100	100.0	1.30	11.98	69.49
	300	100.0	1.17	13.97	61.21
	1000	100.0	1.13	17.04	57.10

Table 1: Success rates and perturbation sizes on the 1000 images from the ImageNet-Compatible dataset, with varied budgets in the targeted setting. Perturbation size is quantified in terms of L_2 and L_∞ norms of the perturbations in RGB space ($\overline{L_2}$ and $\overline{L_\infty}$) and also in terms of image-level accumulated perceptual color difference ($\overline{C_2}$). Note that C&W and PerC-C&W actually need more (here, 5×) iterations to find the optimal initialization of λ. The budget for I-FGSM varies on different images.

as in the original work DDN [44]. The α_c in PerC-AL is gradually reduced from 0.5 to 0.05 with cosine annealing.

Evaluation Protocol. We investigate a set of reasonable operating points, based on pre-defined budgets. Note that our goal is to show the relative behavior of PerC vs. RGB approaches. For this purpose, we only need to create a fair comparison, and it is not necessary to drive all approaches to an absolute optimum. For each image, an approach is considered successful if the perturbed image can achieve adversarial effect with the given budget. Specifically, I-FGSM requires varied repetitions for different images. For C&W and PerC-C&W, the budget refers to N(search steps) × N(iterations of gradient descent). We apply relatively high budget (9×1000), and are also interested in lower budgets (5×200 and 3×100), which are more directly comparable with more efficient approaches, namely, DDN and PerC-AL. We test DDN and our PerC-AL with three different iteration budgets (100, 300 and 1000), adopted from the original work [44].

Adversarial strength is evaluated by the success rate, i.e., the proportion of successful cases over the whole dataset. The averaged perturbation size over all successful images is reported. It is measured in terms of the L_2 and L_∞ norm in RGB space ($\overline{L_2}$ and $\overline{L_\infty}$) and also in terms of image-level accumulated perceptual color difference ($\overline{C_2}$).

Approach	$\kappa = 20$ Suc. (%)	$\overline{C_2}$	$\kappa = 40$ Suc. (%)	$\overline{C_2}$
I-FGSM [29]	100.0	375.74	99.9	576.06
C&W [7]	100.0	159.00	100.0	241.92
DDN [44]	100.0	150.68	98.1	238.37
PerC-C&W (ours)	100.0	90.86	100.0	136.22
PerC-AL (ours)	100.0	75.43	100.0	115.17

Table 2: Evaluation of the success rate and perceptual color difference achieved by different approaches in high-confidence settings.

5.2. Adversarial strength and imperceptibility

In this section, we investigate the adversarial strength and imperceptibility of the perturbed images generated by different approaches in a white-box scenario, where the full information of the network is accessible.

5.2.1 Sufficient-confidence adversarial examples

We first present, in Table 1, a comparison demonstrating how PerC approaches relax L_p norms. Our comparison uses adversarial examples created under a commonly used condition where the aim is to achieve a just sufficient adversarial effect. *Sufficient-confidence adversarial examples* just cross the decision boundary without pursuing a higher confidence score for the adversarial label. As expected, all approaches achieve 100% success rate and the resulting perturbation size gets smaller as the budget increases.

Table 1, which reports the targeted results, confirms that PerC approaches, PerC-C&W and PerC-AL, show the behavior they are designed for, i.e., decreasing the average accumulated perceptual color difference $\overline{C_2}$. More importantly, PerC approaches do this without tightly constraining the L_p norms in RGB space as the other approaches do, as reflected by $\overline{L_2}$ and $\overline{L_\infty}$. Moreover, PerC-AL achieves lower $\overline{C_2}$ than PerC-C&W (57.10 vs. 67.79) with notably fewer iterations. For comparison, we provide $\overline{C_2}$ for the RGB approaches. The untargeted results follow a similar pattern and can be found in the supplementary material.

5.2.2 High-confidence adversarial examples

In order to gain deeper insight into the performance of our approaches, we investigate adversarial examples that have a high confidence score for the adversarial label. High confidence was initially investigated by [7] in order to achieve more transferable adversarial examples, and also been explored in the "Unrestricted Adversarial Examples" contest [6]. In the untargeted setting, an approach is regarded as successful only if the logit with respect to the original class becomes lower than the maximum of the other logits

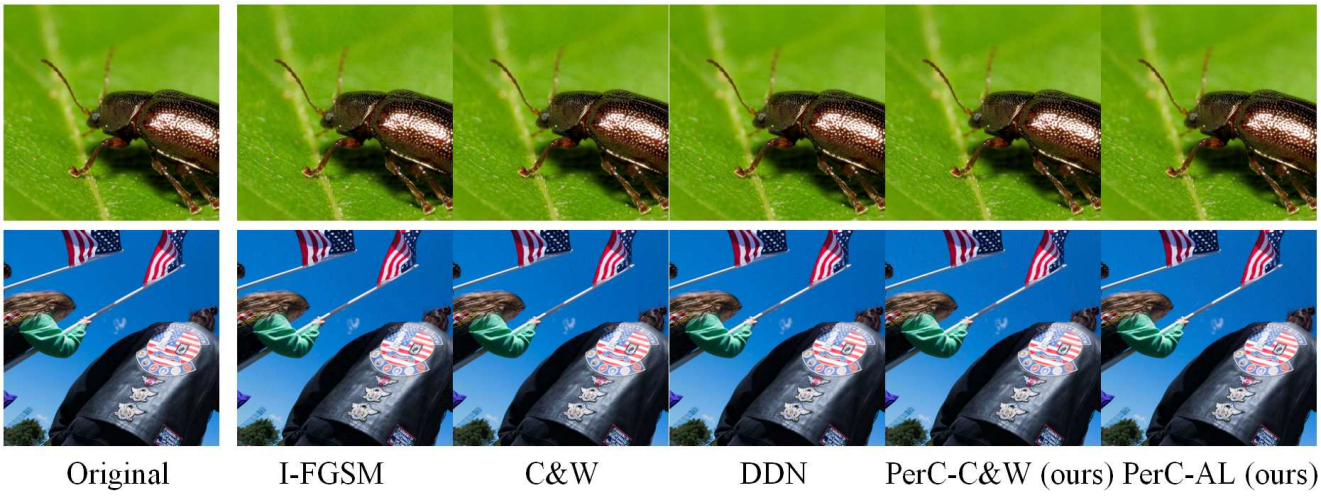

Figure 3: Examples of adversarial images generated by five different approaches with high confidence level $\kappa = 40$

by a pre-defined margin κ. For C&W and PerC-C&W, this requirement can be directly implemented by specifying the factor κ in Eq. (4). For I-FGSM, DDN and PerC-AL, this can be achieved by running the iterations until the required logit difference is satisfied. For this experiment, we adopt the settings generating the smallest perturbations for each approach in Section 5.2.1.

Fig. 3 shows some adversarial examples generated by different approaches at $\kappa = 40$. The images produced by our PerC approaches look more visually acceptable than those of the other approaches (best viewed on screen). More examples can be found in the supplementary material. The good visual appearance of the PerC examples is consistent with their low averaged aggregated perceptual color difference, $\overline{C_2}$, as seen in Table 2, which shows both $\kappa = 40$ and $\kappa = 20$ values. The challenge of the high-confidence setting is seen in the success rates, which are not longer perfect for all conditions.

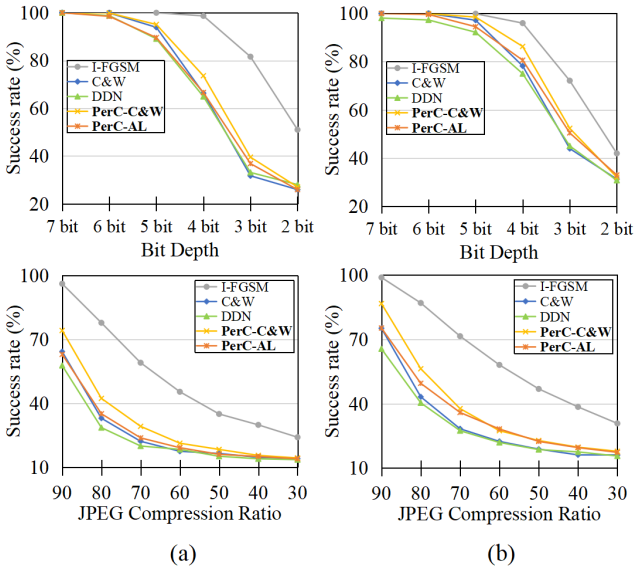

Figure 4: Evaluation of robustness of high-confidence adversarial examples at (a) $\kappa = 20$ and (b) $\kappa = 40$, against two types of image transformations: JPEG compression (top row) and bit-depth reduction (bottom row).

5.3. Robustness

In order to gain additional practical insight, we test the robustness of the adversarial examples against two commonly studied image transformation-based defense methods, i.e., JPEG compression [12, 13, 14, 20] and bit-depth reduction [20, 22, 57].

The results are shown in Fig. 4. Overall, increasing κ from 20 to 40 leads to improved robustness. For a specific κ, unsurprisingly, I-FGSM outperforms other approaches since it greedily perturbs all the pixels, but at the cost of worse image quality (see Fig. 3). Among the other four approaches that target minimal image-level accumulated image difference with sparse perturbations, the best results are consistently achieved by either our PerC-C&W or PerC-AL. Specifically, PerC-C&W outperforms the original C&W in all cases, while PerC-AL consistently outperforms DDN. Recall that our PerC approaches cause fewer visual distortions, as shown in Fig. 3, contributing to imperceptibility.

5.4. Transferability

Existing research [35, 51] has demonstrated that the adversarial effect of images optimized with respect to a specific network may transfer to another network. We test the transferability of different approaches from the original Inception V3 to other three pre-trained networks, namely,

	GoogLeNet		VGG-16		ResNet-152	
	$\kappa = 20$	$\kappa = 40$	$\kappa = 20$	$\kappa = 40$	$\kappa = 20$	$\kappa = 40$
I-FGSM [29]	*4.2*	*6.3*	*5.6*	*10.6*	*2.4*	*4.2*
C&W [7]	2.5	3.1	3.0	5.1	0.9	1.7
DDN [44]	2.1	3.1	3.3	5.7	1.2	2.4
PerC-C&W (ours)	2.9	4.7	3.9	6.9	1.2	2.5
PerC-AL (ours)	2.6	4.3	4.8	7.2	1.2	2.4

Table 3: Success rates of adversarial examples at two high confidence levels $\kappa = 20$ and $\kappa = 40$ in the transfer scenario, from the source model Inception V3 to three others.

GoogLeNet [48], ResNet-152 [21], and VGG-16 [47]. Specifically, an untargeted adversarial example generated for the original model is regarded to be transferable to a new model if it can also induce misclassification of that model.

We report results on a subset of our data containing images that all four models originally classify correctly. Table 3 reports the success rates under transferability on these images (767 in total). I-FGSM again outperforms the other approaches, but uses excessive perturbations (and for this reason is shown in italics). Among the other approaches, we can observe that the best results are always achieved by one of our two PerC approaches.

5.5. Assembling structural information

We explore the possibility of assembling structural information for further improving imperceptibility without impacting adversarial strength. Specifically, we introduce a texture complexity vector $\boldsymbol{\sigma}$, which has the same size as the image, as a weighting term into our PerC-AL framework. Following existing work [11, 37] on addressing imperceptibility with respect to image structure, this vector is obtained by calculating the standard deviation of the pixel values in each 3×3 square per channel. The components with top 5% highest values in the map are clipped for stability and the map is normalized into the range [0,1] before use. Concretely, step 8 in Algorithm 1 is adjusted to:

$$C_2 \leftarrow -\|(\mathbf{1} - \boldsymbol{\sigma}) \cdot \Delta E_{00}(\boldsymbol{x}, \boldsymbol{x}'_{k-1})\|_2, \qquad (10)$$

where C_2 also becomes sensitive to image differences in terms of local visual variation. As shown in Fig. 5, with the help of additional structural information, perturbations in the smooth regions are suppressed, while more changes, which are barely perceptible, are triggered in the area with rich visual variation. It is worthwhile for future work to investigate this combined approach in more detail.

6. Conclusion and Outlook

This paper has demonstrated the usefulness of perceptual color distance for creating large yet imperceptible adversarial image perturbations. We have proposed two approaches for creating adversarial images, PerC-C&W and PerC-AL.

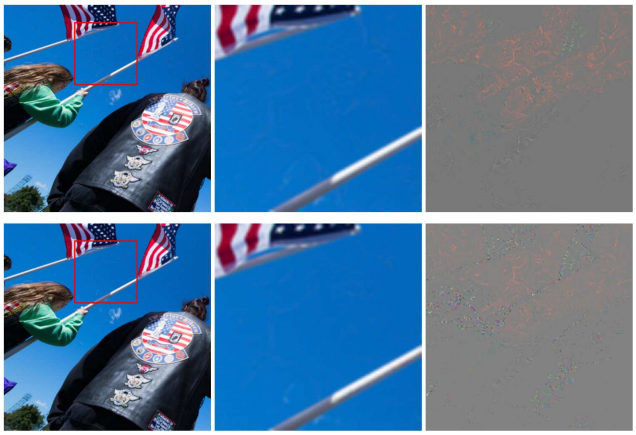

Figure 5: Adversarial examples of an image at $\kappa = 40$. Top: Generated by PerC-AL (Algorithm 1). Bottom: Generated by PerC-AL plus structure (Algorithm 1 plus Eq. (10)).

Our experimental investigation of these approaches shows that perceptual color distance is able to improve imperceptibility, especially in smooth, saturated regions. We show that these approaches have perturbations with larger RGB L_p norms than approaches that perturb directly in RGB space. This effect translates into adversarial strength, i.e., the ability of the perturbations to fool a classifier.

Our work has made a contribution to recent work that seeks to create adversarial images that are imperceptible to the eye of the human observer. This work has been carried out in the area of security [7, 17, 16, 29, 43] (defend inference of a legitimate classifier) and privacy [9, 36, 40, 42] (prevent inference of an illegitimate classifier). In the security area, imperceptible perturbations can mean that adversarial images can poison the training data without being noticed by human annotators. In the privacy area, imperceptible perturbations mean wider acceptance of the use of adversarial images to protect against classification attacks.

In the future, we will continue to consider perceptual color in adversarial images from both the privacy and the security angle. Our first direction will be related to the fact that neither conventional RGB perturbations nor PerC perturbations perform well in smooth regions with low saturation. We would like to develop techniques that can make perturbations imperceptible, or unnecessary, in such regions. Our future work will also look at model robustness specifically against our PerC adversaries. On one hand, adversarial training on images perturbed by PerC is worth exploring to complement current research on L_p robustness. On the other hand, it would be interesting to investigate ways to detect whether PerC has been applied to an image, or design countering methods that can mitigate PerC-based perturbations by, for example, applying bit-depth reduction directly in perceptual color space.

References

[1] ISO/CIE 11664-6:2014(E) Colourimetry-Part 6: CIEDE2000 Colour difference Formula. 2, 3

[2] Mahmoud Afifi and Michael S Brown. What else can fool deep learning? Addressing color constancy errors on deep neural network performance. In *ICCV*, pages 243–252, 2019. 4

[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *ICML*, pages 284–293, 2018. 2

[4] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and David A Forsyth. Unrestricted adversarial examples via semantic manipulation. In *ICLR*, 2020. 4

[5] Wieland Brendel, Jonas Rauber, Alexey Kurakin, Nicolas Papernot, Behar Veliqi, Marcel Salathé, Sharada P. Mohanty, and Matthias Bethge. Adversarial vision challenge. *arXiv preprint arXiv:1808.01976*, 2018. 4

[6] Tom B Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples. *arXiv preprint arXiv:1809.08352*, 2018. 6

[7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, pages 39–57, 2017. 1, 3, 5, 6, 8

[8] Heng-Da Cheng, Xihua Jiang, Ying Sun, and Jingli Wang. Color image segmentation: advances and prospects. *Pattern Recognition*, 34(12):2259–2281, 2001. 3

[9] Jaeyoung Choi, Martha Larson, Xinchao Li, Kevin Li, Gerald Friedland, and Alan Hanjalic. The geo-privacy bonus of popular photo enhancements. In *ICMR*, pages 84–92, 2017. 4, 8

[10] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, pages 1310–1320, 2019. 1

[11] Francesco Croce and Matthias Hein. Sparse and imperceivable adversarial attacks. In *ICCV*, pages 4724–4732, 2019. 2, 4, 5, 8

[12] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E. Kounavis, and Duen Horng Chau. Shield: Fast, practical defense and vaccination for deep learning using JPEG compression. In *SIGKDD*, pages 196–204, 2018. 7

[13] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*, pages 4312–4321, 2019. 5, 7

[14] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A study of the effect of JPG compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016. 7

[15] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. In *NIPS 2017 Workshop on Machine Learning and Computer Security*, 2017. 2, 4

[16] Gragnaniello et al. Perceptual quality-preserving black-box attack against deep learning image classifiers. *IEEE Transactions on Information Forensics and Security (TIFS)*, 2019. 2, 4, 8

[17] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning models. In *CVPR*, pages 1625–1634, 2018. 2, 8

[18] Leon A. Gatys, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Preserving color in neural artistic style transfer. *arXiv preprint arXiv:1606.05897*, 2016. 3

[19] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1, 3, 4

[20] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *ICLR*, 2018. 7

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 8

[22] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defense: Ensembles of weak defenses are not strong. In *USENIX Workshop on Offensive Technologies*, 2017. 7

[23] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *CVPR Workshops*, pages 1614–1619, 2018. 2, 4

[24] Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *ICCV*, pages 4773–4783, 2019. 2

[25] Leonid Vasilevich Kantorovich and Gennady S Rubinstein. On a space of completely additive functions. *Vestnik Leningrad. Univ*, 13(7):52–59, 1958. 4

[26] Fahad Shahbaz Khan, Rao Muhammad Anwer, Joost Van De Weijer, Andrew D. Bagdanov, Maria Vanrell, and Antonio M. Lopez. Color attributes for object detection. In *CVPR*, pages 3306–3313, 2012. 3

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 5

[28] Alex Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009. 5

[29] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *ICLR*, 2017. 1, 3, 4, 5, 6, 8

[30] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 195–231. 2018. 5

[31] Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. In *NeurIPS*, 2019. 4

[32] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[33] Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989. 3

[34] Shuaicheng Liu, Michael S Brown, Seon Joo Kim, and Yu-Wing Tai. Colorization for single image super resolution. In *ECCV*, pages 323–336, 2010. 3

[35] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017. 7

[36] Zhuoran Liu, Zhengyu Zhao, and Martha Larson. Who's afraid of adversarial queries? The impact of image modifications on content-based image retrieval. In *ICMR*, pages 306–314, 2019. 8

[37] Bo Luo, Yannan Liu, Lingxiao Wei, and Qiang Xu. Towards imperceptible and robust adversarial example attacks against neural networks. In *AAAI*, 2018. 2, 4, 5, 8

[38] Ming Ronnier Luo, Guihua Cui, and B. Rigg. The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research and Application*, 26(5):340–350, 2001. 2, 3

[39] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 1, 5

[40] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. Gender privacy: An ensemble of semi adversarial networks for confounding arbitrary gender classifiers. In *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10, 2018. 8

[41] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: a simple and accurate method to fool deep neural networks. In *CVPR*, pages 2574–2582, 2016. 1, 3

[42] Seong Joon Oh, Mario Fritz, and Bernt Schiele. Adversarial image perturbation for privacy protection a game theory perspective. In *ICCV*, pages 1491–1500, 2017. 8

[43] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387, 2016. 1, 3, 8

[44] Jérôme Rony, Luiz G. Hafemann, Luiz S. Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based L_2 adversarial attacks and defenses. In *CVPR*, pages 4322–4330, 2019. 3, 4, 5, 6, 8

[45] Mahmood Sharif, Lujo Bauer, and Michael K. Reiter. On the suitability of L_p-norms for creating and preventing adversarial examples. In *CVPR Workshops*, pages 1605–1613, 2018. 2

[46] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)*, 2019. 2

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 8

[48] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 5, 8

[49] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 1, 3

[50] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *ICLR*, 2017. 3

[51] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *ICLR*, 2018. 1, 7

[52] Koen Van De Sande, Theo Gevers, and Cees Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32(9):1582–1596, 2009. 3

[53] 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 (TIP)*, 13(4):600–612, 2004. 4

[54] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *ICML*, pages 5283–5292, 2018. 1

[55] Eric Wong, Frank Schmidt, and Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In *ICML*, pages 6808–6817, 2019. 2, 4

[56] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *ICLR*, 2018. 4, 5

[57] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Network and Distributed Systems Security Symposium (NDSS)*, 2018. 7

[58] Yang Yang, Jun Ming, and Nenghai Yu. Color image quality assessment based on CIEDE2000. *Advances in Multimedia*, 2012:11, 2012. 3

[59] Hanwei Zhang, Yannis Avrithis, Teddy Furon, and Laurent Amsaleg. Smooth adversarial examples. *EURASIP Journal on Information Security (JIS)*, 2020. 2, 4, 5