Perception Improvement for Free: Exploring Imperceptible Black-box Adversarial Attacks on Image Classification

Yongwei Wang[a], Mingquan Feng, Rabab Ward, Z. Jane Wang, Lanjun Wang

[a]*Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada*

Abstract

Deep neural networks are vulnerable to adversarial attacks. White-box adversarial attacks can fool neural networks with small adversarial perturbations, especially for large size images. However, keeping successful adversarial perturbations imperceptible is especially challenging for transfer-based black-box adversarial attacks. Often such adversarial examples can be easily spotted due to their unpleasantly poor visual qualities, which compromises the threat of adversarial attacks in practice. In this study, to improve the image quality of black-box adversarial examples perceptually, we propose structure-aware adversarial attacks by generating adversarial images based on psychological perceptual models. Specifically, we allow higher perturbations on perceptually insignificant regions, while assigning lower or no perturbation on visually sensitive regions. In addition to the proposed spatial-constrained adversarial perturbations, we also propose a novel structure-aware frequency adversarial attack method in the discrete cosine transform (DCT) domain. Since the proposed attacks are independent of the gradient estimation, they can be directly incorporated with existing gradient-based attacks. Experimental results show that, with the comparable attack success rate (ASR), the proposed methods can produce adversarial examples with considerably improved visual quality for free. With the comparable perceptual quality, the proposed approaches achieve higher attack success rates: particularly for the frequency structure-aware attacks, the average ASR improves more than 10% over the baseline attacks.

Keywords: adversarial attacks, spatial perceptual attacks, frequency perceptual attacks, perceptual quality,

1. Introduction

Deep neural networks (DNNs) have achieved significant progress in a wide range of machine learning tasks [1–6]. However, their robustness has been greatly challenged by the existence of adversarial examples, where carefully perturbed images (as the inputs) can easily fool deep neural networks. Since Szegedy et al. [7] first reported adversarial examples, there have been intensive studies on the effectiveness of adversarial examples [8–15].

In practice, a valid adversarial example satisfies two constraints: a) *high attack success rate*, i.e., adversarial examples can fool the target models with a high attack success rate; and b) *high perceptual quality*, i.e., adversarial examples are semantically preserving and meaningful, which indicates the image content is preserved and the image perceptual quality is as naturally-looking as possible.

White-box adversarial attack methods [10–12] can easily generate valid adversarial examples satisfying the above two constraints, because the adversary has full knowledge of the deployed model. However, to meet these constraints is much more challenging for black-box attacks [9, 15]. For example, [15], one of the latest attacks with high attack success rates, requires relatively large perturbations, which can generally degrade the perceptual quality of the generated adversarial examples. For example in Fig. 1, we depict an adversarial example with perturbation generated by [15], which displays unpleasant or unnatural visual artifacts. Despite those adversarial examples with poor visual qualities remain fooling the model, their threats to certain practical deployed systems (e.g., deepfake forensics [16]) can be largely compromised, because indeed they break the 'imperceptibility' property of adversarial attacks and can be easily spotted and filtered out by sanity checks. As a result, a key problem we need to solve for black-box adversarial attacks is *whether it is possible to keep a high attack success rate while preserving a naturally-looking visual quality?*

According to studies on the human cognition system, we realize that the essence of visual degradation issues is the identical and independent perturbation bound for each pixel. More specifically, such identical and independent perturbation bound is incompatible with the human visual system, which is highly sensitive to the structural information in scene perception [17, 18]. In detail, structural representations can be described by edge, texture and luminance contrast by extracting oriented gradients and relative intensity from neighboring pixels in the spatial domain [19]. Moreover, visual frequency sensitivity can be integrated into constructing visual descriptors in the frequency domain [20, 21]. Therefore, uniform distortions in previous adversarial attack studies are not aligned well with the human visual system. The visual quality issue is not obvious for white-box attacks because the perturbation bound can be very small due to the fully known information. However, we have to solve the visual

Email address: yongweiw@ece.ubc.ca (Yongwei Wang)

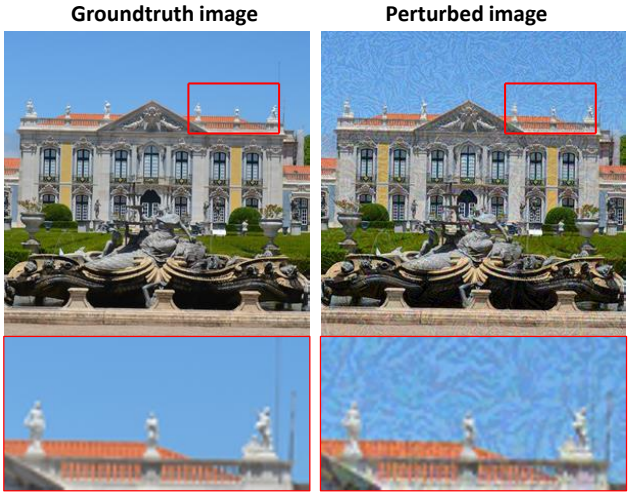

Fig. 1: Illustration of a visually degraded adversarial example. Left: the groundtruth image; right: the perturbed image with adversarial perturbations generated by [15]. By zooming into the image patch (i.e, the red box), we can clearly notice the visual artifacts introduced by adversarial perturbations.

quality issue in black-box attacks.

To generate black-box adversarial examples by considering the human perception behavior is very challenging. Firstly, to replace the uniform perturbation, we need a new type of distortion metric to represent the structural properties of images. In this study, we incorporate the results from psychological studies. The structure-aware image-dependent perceptual models [17] can identify which regions the visual systems pay more attention and which regions is more likely to be ignored. These models have been applied in the fields of image compression [19] and video coding [18], where higher compression rates are applied on the unnoticeable regions, but lower compression rates or even not to compress on the noticeable regions. We propose to leverage these perceptual models on setting a structure-aware adversarial attack. More specifically, we allow higher perturbations on perceptually insignificant regions, while assigning lower or no perturbation on significant regions.

Secondly, only considering perturbations in the spatial domain is not enough, because perceptual systems are closely related to frequency selectivity [21]. Although there exist frequency perceptual models to quantitatively measure frequency sensitivity (e.g., [19, 20]), it is nontrivial to incorporate frequency visual models to the adversarial attack setting. To leverage the frequency perceptual models, we propose to directly add adversarial perturbations in the frequency domain, i.e., we formulate a novel adversarial attack objective function in the frequency domain with the frequency sensitivity constraint, then frequency perturbation is conducted with gradients derived for each frequency sub-band.

Thirdly, there is always a trade-off between the attack success rate and perceptual quality [22]. Simply achieving imperceptible perturbations alone is not enough, while we still need to keep a high success rate in the black-box attacks. In this study, we carefully select the structure-aware perceptual incorporation strategy to make them independent of the existing gradient-based attack algorithms. As a result, we can leverage the state-of-art gradient estimation methods, while constrain the perturbation setting based on the perceptual models.

We summarize our major contributions as follows:

- We design a framework to generate structure-aware distortions in adversarial attacks, and apply it on black-box adversarial attacks to preserve a naturally-looking visual quality while keeping a high attack success rate. Since the structure-aware strategy is independent of the gradient estimation, this framework can be generally extended to any gradient-based adversarial attack regardless of the white-box or black-box setting.

- Besides the spatial structure-aware perturbations, we propose to incorporate the frequency perceptual models in the adversarial perturbation generation and we develop a novel structure-aware attack approach by adding adversarial perturbations in the frequency domain.

- Experiments demonstrate that, with the comparable attack success rate, the proposed methods have significant perceptual improvements when compared with the baseline attacks. Meanwhile, with the comparable perceptual quality, we also observe the improved attack success rate over the baseline attacks .

2. Background

The existence of adversarial examples poses severe threats to deep learning models. A wide range of studies have been investigated to generate adversarial examples to fool neural networks with a high probability [8–15, 23]. However, all of these studies neglected the perceptual quality evaluation on adversarial examples.

Meanwhile, limited attention has been paid to generating adversarial examples with high perceptual quality. Luo et al. introduced an overall noise sensitivity measure based on noise variance estimation for white-box attacks [24]. Croce et al. introduced a sparse ℓ_0 ball constraint to the query-limited attack and optimized the perturbation with local search [25]. The sparse perturbations are assigned to sparse regions with high variances to reduce visual distortions. However, neither of them is aligned with human visual systems, because some complicated structures (e.g., textures, edges, luminance contrast) or frequency response of an image are not explicitly modeled.

Furthermore, the idea of incorporating psychological studies [17, 18] is inspired by related works from image compression [19] and video coding [18]. For instance, in video coding [18], the perceptual-model incorporated codecs achieve both high perceptual fidelity and a high compression rate.

2.1. Adversarial Attack Models

Given a clean image $\boldsymbol{x}$, an image classifier f_θ predicts its label as y, i.e., $f_\theta(\boldsymbol{x}) = y$. Conventionally, a non-targeted adversarial example $\boldsymbol{x}^*$ can be formally defined as

$$f_\theta(\boldsymbol{x}^*) \neq y, \quad \text{s.t. } \|\boldsymbol{x}^* - \boldsymbol{x}\|_p \leq \epsilon \tag{1}$$

By definition, an adversarial example $\boldsymbol{x}^*$ is bounded within the ϵ ball of $\boldsymbol{x}$, with distance measured by the ℓ_p norm.

With a higher ϵ factor, the attack methods produce relatively high attack success rate at the expense of possibly severely degraded perceptual quality. The key issue lies in the fact that the distortion criterion treats each pixel independently and assigns a uniform bound with each pixel. However, human eyes mainly perceive images using local and regional statistics. Therefore, the tolerable distortion level should be different from pixel to pixel due to their different structure information defined by neighboring regions [18, 19, 21].

3. Perceptual Models

The understanding of the human visual system is essential to generate high quality adversarial examples. We employ perceptual models to guide adversarial example generation. Perceptual models are developed over the years based on the property of human visual system over scene perception. Psychovisual study reveals that visual sensitivity relies on structural information rather than value changes at a single pixel [17, 19]. A common paradigm of perceptual models in image processing is the just-noticeable difference (JND) model, which was originally derived for image compression [19].

In JND models, the structure and local statistics are generally described by luminance sensitivity, contrast masking and frequency masking effects [21]. There are various types of JND models in the literature, e.g., the spatial domain JND [26] and the frequency domain JND [17, 20, 27]. Based on JND models, we can estimate the maximal perturbation bounds within the imperceptibility constraint. It also indicates the perceptual importance on each pixel, and so we leverage it to design non-uniform distortions. We are motivated to incorporate the perceptual-model based constraint to generate adversarial examples with high visual quality. We briefly describe two JND models we adopt in the following sections.

3.1. Spatial JND Model

In the spatial domain, we apply a basic JND model, which considers the image structure that consists of textures and local luminance distribution [18]. The JND profile is obtained by calculating the dominant values of its two structural components. Specifically, the spatial JND for a grayscale image, denoted by JND_s, is defined as follows:

$$\mathcal{JND}_s = \mathcal{TM} + \mathcal{LA} - C \cdot \min\{\mathcal{TM}, \mathcal{LA}\} \tag{2}$$

where $\mathcal{TM}$ represents the texture masking, $\mathcal{LA}$ represents the luminance adaptation, and $C \in (0, 1)$ measures the overlapping effect between the texture masking and luminance adaptation effects. Empirically, we set C as 0.3 according to [18].

Texture masking refers to the ability of hiding or obscuring a superimposed stimulus with textures. [17, 18] show that the visual sensitivity to distortion is low in the texture-rich regions. The visual importance defined by texture masking is estimated as:

$$\mathcal{TM} = \max_{k=1,2,3,4} |\boldsymbol{x} * \boldsymbol{h}_k| \cdot (\boldsymbol{m}_{\boldsymbol{x}} * \boldsymbol{l}_g) \tag{3}$$

where $\boldsymbol{h}_k$ ($k = 1, 2, 3, 4$) are four directional high-pass filters for texture detection, $\boldsymbol{m}_{\boldsymbol{x}}$ denotes the edge map of image $\boldsymbol{x}$ given by the Canny edge detector [28], and $\boldsymbol{l}_g$ represents a Gaussian low-pass filter. The filter parameters are provided in Appendix A.

Compared with the absolute luminance of a single pixel, human perceptions are more sensitive to the relative luminance among its neighboring pixels. The luminance adaptation threshold is calculated based on Weber's law and deduced from psychological experiments under uniform background [29]. The luminance adaptation effect $\mathcal{LA}$ is modeled as,

$$\mathcal{LA}_{i,j} = \begin{cases} 17 \times (1 - \sqrt{\frac{\tilde{\boldsymbol{x}}_{i,j}}{127}}) & \text{if } \tilde{\boldsymbol{x}}_{i,j} \leq 127 \\ \frac{3 \times (\tilde{\boldsymbol{x}}_{i,j} - 127)}{128} + 3 & \text{otherwise} \end{cases} \tag{4}$$

where (i, j) denotes pixel position of a grayscale image, $\mathcal{LA}_{i,j}$ denotes the $(i, j)-$th component of the luminance adaptation map, $\tilde{\boldsymbol{x}} = \boldsymbol{x} * \boldsymbol{l}$, and $\boldsymbol{l}$ is a low-pass filter.

3.2. Frequency JND Model

In addition to spatial luminance adaptation and texture masking effects, the sensitivity of the human visual system is closely related to frequency sensitivity [19]. We adopt a frequency perceptual model proposed in [20]. To describe the frequency perceptual model in short, images are firstly decomposed into sub-band domains. Then, local contrast masking and spatial contrast sensitivity factors can be modeled based on frequency coefficients in each block. The final frequency JND is obtained as the multiplication of these two factors.

4. Methodology

4.1. Imperceptible Spatial-Domain Attack

By departing from the identical ℓ_p bound for each pixel value, we consider the perceptual importance of pixels which directly depend on image local structures [11, 15]. Specifically, we allow larger perturbations to perceptually insignificant regions while smaller or no perturbations to perceptually significant regions. The perceptual importance is estimated from the neighborhood structures using the spatial JND model. To explicitly consider the imperceptibility property, we propose to incorporate and rectify existing adversarial example generation methods utilizing the spatial JND constraint.

In the spatial domain, the optimization function of our JND-constraint spatial adversarial attack model is formulated as,

$$\begin{aligned} & f_\theta(\boldsymbol{x}^*) \neq y \\ & \text{s.t.} \quad |\boldsymbol{x} - \boldsymbol{x}^*| \leq \mathcal{JND}_s \end{aligned} \tag{5}$$

where $|\cdot|$ is the absolute value operator, $\mathcal{JND}_s$ denotes the spatial importance matrix estimated from the JND model computed from $\boldsymbol{x}$. An intuitive explanation of our objective function is the following: We distinguish pixel-wise importance inherent in images extracted from local structures. Consequently the perturbation budgets vary from region to region. The gradient of the output with respect to the clean input $\boldsymbol{x}$ is a key value

in the adversarial attack generation. In black-box attacks, as there is no internal knowledge on either the model architecture or the loss function, it is impossible to calculate gradient directly. However, different types of black-box attacks leverage different methods to estimate such gradient information. In the substitute model based attack which is our main focus, we can estimate the gradient with an substitute model [9, 11, 15], then generate adversarial examples regarding to the new constraint as in Eq.(5), and finally transfer examples to the black-box model. In this study, we denote the gradient estimation method as $\boldsymbol{g}^{est}(\boldsymbol{x}, y)$.

The perceptual-constraint model can be solved using the gradient-based method iteratively,

$$\boldsymbol{x}^*_{t+1} = \boldsymbol{x}^*_t + \alpha \cdot \mathcal{JND}_s \odot \text{sign}(\boldsymbol{g}^{est}_t(\boldsymbol{x}^*_t, y)) \tag{6}$$

where $\odot$ denotes the elementwise product, $\boldsymbol{g}^{est}_t(\boldsymbol{x}^*_t, y)$ is the estimated gradient w.r.t. $\boldsymbol{x}_t$ at the $t-$th iteration. Informal studies show that exceeding JND thresholds occasionally does not yield severely degraded visibility. Therefore, we can exploit the additional tolerance by multiplying the importance map by a scalar factor $\alpha(\alpha \geq 1)$ in Eq.(6). Then, we can better balance the trade-off between the attack success rate and image quality.

Compared with the commonly used ϵ ball uniform bound, the proposed perturbation bound is image dependent and region dependent, which directly incorporates spatial perceptual models. In Eq.(6), our image-dependent and stepsize-variant expression is a more general solution. Moreover, our method reduces to existing methods (e.g.,[11]) when we choose a uniform perturbation bound ϵ as $\alpha \cdot \max\{\mathcal{JND}_s\}$, then we have the same constraint optimization problem as in Eq.(1) with an ℓ_∞ norm.

The overall structure-aware adversarial spatial attack framework is illustrated in Algorithm 1. In this study, the JND threshold is calculated based on the grayscale version of a natural image. The final JND profile of a color image is formed by replicating the grayscale JND for each color channel. Although there exists color JND models, here we adopt a simple JND model in order to show the perceptual improvement by explicit utilization of structural information.

Algorithm 1: The proposed spatial structure-aware (SSA) adversarial attack algorithm.

Data: A black-box model $f(\boldsymbol{x})$, clean image $\boldsymbol{x}$, correct label y, gradient estimation method $\boldsymbol{g}^{est}(\boldsymbol{x}, y)$, scalar factor α_0, and iteration number T.
Result: An adversarial example $\boldsymbol{x}^*$.
Calculate $\mathcal{JND}_s$ from the grayscale version of $\boldsymbol{x}$.
$\mathcal{JND} \leftarrow [\mathcal{JND}_s; \mathcal{JND}_s; \mathcal{JND}_s]$.
Initialize $\boldsymbol{x}^*_0 \leftarrow \boldsymbol{x}$, $t \leftarrow 0$
while $t < T$ *and* $f(\boldsymbol{x}^*_t) == y$ **do**
  estimate the gradient $\boldsymbol{g}^{est}_t(\boldsymbol{x}^*_t, y)$.
  $\boldsymbol{x}^*_{t+1} \leftarrow \boldsymbol{x}^*_t + \frac{\alpha_0}{T} \cdot \mathcal{JND} \odot \text{sign}(\boldsymbol{g}^{est}_t(\boldsymbol{x}^*_t, y))$, $t \leftarrow t + 1$
$\boldsymbol{x}^* \leftarrow \boldsymbol{x}^*_t$.

4.2. Imperceptible Frequency-Domain Attack

Apart from the spatial domain perturbation, recently there were several pioneering works on perturbing images in the frequency domain. Tsuzuku el al. investigated the sensitivity of neural networks to certain Fourier basis functions based on the linearity hypothesis of neural networks [30]. The adversarial examples can be crafted by making queries to the target model to find suitable Fourier basis. Adversarial examples from single Fourier attack method display repeated patterns in the pixel domain. Guo et al. restricted the adversarial perturbation space to the low frequency domain, and proposed a query-efficient attack method [31]. Despite the effectiveness of low frequency perturbations, the visual quality of adversarial examples is significantly degraded [32].

In previous frequency attack methods, adversarial perturbations are added in the spatial domain with the uniform ℓ_p-norm bound with frequency correction. However, our proposed frequency domain attack is directly conducted in the frequency domain iteratively without any spatial domain constraint. Instead we explicitly consider the perceptual distortion bounds in the frequency domain. This makes the proposed perceptual-constraint frequency attack different from existing adversarial attack methods. We observe that the proposed perceptual frequency-constraint adversarial attack generally yields higher perceptual quality than the spatial domain attack.

In this study, we use the discrete cosine transform (DCT) domain for the frequency domain attack. For expression convenience, we consider the single channel case, since it is straightforward to extend operations to color images by performing transformations for each channel. Assume the clean image $\boldsymbol{x} \in \mathbb{R}^{N \times N}$, then we can obtain $\boldsymbol{X}$ by dividing the spatial image into square blocks of size $\mathcal{B} \times \mathcal{B}$. The DCT transform is conducted for each block $\boldsymbol{x}^b (b = 0, 1, \cdots, \lceil \frac{N}{\mathcal{B}} \rceil - 1)$ as,

$$\boldsymbol{X}^b = \boldsymbol{D}\boldsymbol{x}^b\boldsymbol{D}^T \tag{7}$$

where $\boldsymbol{D}$ is an orthogonal matrix, $\boldsymbol{D}\boldsymbol{D}^T = I_{\mathcal{B}\times\mathcal{B}}$, with entries $\boldsymbol{D}_{m,n}(m, n = 0, 1, \cdots, \mathcal{B} - 1)$ as,

$$\boldsymbol{D}_{m,n} = \begin{cases} \sqrt{\frac{1}{\mathcal{B}}} & \text{if } m = 0 \\ \sqrt{\frac{2}{\mathcal{B}}} cos(\frac{(2m+1)n\pi}{2\mathcal{B}}) & \text{otherwise} \end{cases} \tag{8}$$

Similarly to the perceptual model-based spatial attack, we formulate the objective function for the frequency attack as,

$$\begin{aligned} & f_\theta(\boldsymbol{x}^*) \neq y \\ & \text{s.t.} \quad |\boldsymbol{X} - \boldsymbol{X}^*| \leq \mathcal{JND}_f \end{aligned} \tag{9}$$

where $\boldsymbol{X}, \boldsymbol{X}^*$ denote the clean and adversarial example in the frequency domain, respectively; $\mathcal{JND}_f$ refers to the JND matrix estimated by the frequency JND model.

To solve Eq.(9), we need to calculate the gradient of the loss function w.r.t. $\boldsymbol{X}$. At each block, the gradient w.r.t. each frequency coefficient $\boldsymbol{X}^b_{u,v}(u, v = 0, 1, \cdots, \mathcal{B} - 1)$ can be calculated by propagating spatial gradient to the DCT domain,

$$\boldsymbol{G}^{est}(\boldsymbol{X}^b_{u,v}, y) = \sum_{i=1}^{\mathcal{B}} \sum_{j=1}^{\mathcal{B}} \boldsymbol{g}^{est}(\boldsymbol{x}^b_{i,j}, y) \cdot \frac{\partial \boldsymbol{x}^b_{i,j}}{\partial \boldsymbol{X}^b_{u,v}} \tag{10}$$

Algorithm 2: The proposed frequency structure-aware (FSA) adversarial attack algorithm.

Data: A black-box model $f(\boldsymbol{x})$, clean image $\boldsymbol{x}$, correct label y, gradient estimation method $\boldsymbol{g}^{est}(\boldsymbol{x}, y)$, scalar factor β_0, and iteration number T.
Result: An adversarial example $\boldsymbol{x}^*$.
Calculate $\mathcal{JND}_f$ from DCT coefficients of grayscale version of $\boldsymbol{x}$.
$\mathcal{JND} \leftarrow [\mathcal{JND}_f; \mathcal{JND}_f; \mathcal{JND}_f]$.
Initialize: $\boldsymbol{x}_0^* \leftarrow \boldsymbol{x}$, $\boldsymbol{X}_0^* \leftarrow DCT(\boldsymbol{x})$, $t \leftarrow 0$.
while $t < T$ *and* $f(\boldsymbol{x}_t^*) == y$ **do**
  Estimate the spatial gradient as $\boldsymbol{g}_t^{est}(f_\theta, \boldsymbol{x}_t^*, y)$.
  Calculate gradient w.r.t. DCT coefficient using Eq.(11) as $\boldsymbol{G}_t^{est}$
  $\boldsymbol{X}_{t+1}^* \leftarrow \boldsymbol{X}_t^* + \frac{\beta_0}{T} \cdot \mathcal{JND} \odot \boldsymbol{G}_t^{est}$
  $\boldsymbol{x}_{t+1}^* \leftarrow iDCT(\boldsymbol{X}_{t+1}^*)$, $t \leftarrow t + 1$
$\boldsymbol{x}^* \leftarrow \boldsymbol{x}_t^*$

And we derive the gradient propagation in the matrix form,

$$\boldsymbol{G}^{est}(\text{Vec } \boldsymbol{X}^b) = \boldsymbol{g}^{est}(\text{Vec } \boldsymbol{x}^b)^T \cdot \left(\boldsymbol{D}^T \otimes \boldsymbol{D}^T\right) \tag{11}$$

where Vec denotes the matrix vectorization operation, $\cdot$ and $\otimes$ denote inner product and matrix Kronecker product, respectively. Finally, we obtain the frequency gradient estimation $\boldsymbol{G}^{est}$.

With the frequency coefficient gradient computed from Eq.(11) as $\boldsymbol{G}_t^{est}$ at the $t-$th iteration, we can readily perform frequency attack with frequency JND in the DCT domain,

$$\boldsymbol{X}_{t+1}^* = \boldsymbol{X}_t^* + \beta \cdot \mathcal{JND}_f \odot \boldsymbol{G}_t^{est} \tag{12}$$

where $\boldsymbol{X}_t^*$ denotes adversarial example in the DCT domain at iteration t ($t = 1, 2, \cdots, T$), $\beta = \beta_0/T$, β_0 is a scalar factor of frequency JND to balance the compromise between perceptual quality and attack success rates.

Finally, the structure-aware frequency perturbation method is described in detail in Algorithm 2.

5. Experimental results

In this section, we evaluate the proposed structure-aware algorithms on three baseline attacks: Fast Sign Gradient Method(FGSM) [11], Momentum Iterative Method(MIM) [9], and Diverse Inputs Method(DIM) [15]. Firstly, we describe the experimental setup and introduce the quantitative visual metrics that we adopted in the comparison. We then experimentally demonstrate the superiority of the proposed methods over baselines on the perceptual quality and the attack success rate. The perturbation residues are illustrated to show the structure-aware property. Finally, we discuss the sensitivity of the parameters in the methods.

5.1. Experimental Setup

For substitute model based attacks, the substitute model is a cleanly trained Inc-v3 model [33] provided by the PyTorch pretrained model zoo [34]. We evaluate the effectiveness of the adversarial examples on six models, three of which are cleanly trained, i.e,. Inc-v4 [35], ResNet-101, ResNet-152 [2], and the rest three models are adversarially trained, i.e., Inc-v3$_{adv}$, Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ [36]. These models are from the NeurIPS 2017 competition track on adversarial attacks [37].

For the dataset, we randomly choose 1000 images from the ImageNet validation dataset [38], which can be correctly classified by the six evaluation models in the substitute model attack setting. The uniform perturbation bound ϵ is set as 14 to have a good attack success rate. The maximum iteration number T is set as 10, which is a default parameter in existing studies [9, 15]. For DCT transformation, we set block size as 8×8, as commonly used in JPEG compression and video coding [20].

5.2. Evaluation Metrics

To reliably evaluate the perceptual improvement of the proposed structure-aware attacks, we adopt four image quality assessment (IQA) metrics: multiSim3 [39], Feature Similarity for color images (FSIMc) [40], Natural Image Quality Evaluator (NIQE) [41] and Mean Opinion Score (MOS) [42]. MultiSim3 and FSIMc are full-reference IQA metrics, with scores within [0,1] where a higher score indicates better visual quality. NIQE is a no-reference IQA metric to measure the naturalness of tested images. NIQE produces a non-negative value, where lower values suggest better naturalness. MOS is a popular human subjective test where we adopt the absolute category rating principle, with image quality score ranging from 1 to 5. The higher the MOS, the better images appears visually similar to clean images. The detailed setting of our MOS test is in Appendix B.

To evaluate the attack effectiveness, we employ the averaged attack success rates (ASR) on six victim models [37]. In the following sections, for simplicity, we term structure-aware approaches as SSA (Spatial-Structure-Aware) and FSA (Frequency-Structure-Aware). Since our proposed methods are independent of gradient estimation methods, we individually incorporate the structure-aware strategies to different gradient-based baseline attacks in the following sections.

5.3. Perception Improvement Assessment

In this section, we compare the perceptual quality between three baseline attacks [9, 11, 15] and our proposed ones, respectively. In each comparison, we firstly keep the average ASRs comparable between the baseline and the proposed ones, i.e., the proposed methods produce equal or slightly higher ASR than the baselines. Then, we provide both quantitative and qualitative comparison results on generated adversarial examples.

Comparison with FGSM Attack [11]: The Fast Sign Gradient Method (FGSM) is a one-step gradient-based attack method, which is a fundamental and widely adopted attack method. The perturbation is generated by maximizing the loss (e.g. cross-entropy) function $J(f_\theta, \boldsymbol{x}, y)$ w.r.t. the input image. The FGSM method meets $\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \epsilon$, and it has an expression as,

$$\boldsymbol{x}^* = \boldsymbol{x} + \epsilon \cdot \text{sign}(\boldsymbol{g}) \tag{13}$$

where $\boldsymbol{g} = \nabla_{\boldsymbol{x}} J(f_\theta, \boldsymbol{x}, y)$ denotes the gradient of the loss function w.r.t the clean sample.

Table 1 shows the attack success rates of FGSM and our proposed variants, i.e., SSA-FGSM and FSA-FGSM. To have a comparable attack success rate, we choose $\alpha_0 = 2.2$, $\beta_0 = 50$. This table shows that SSA-FGSM has similar attack success rate with FGSM for both cleanly trained models and adversarially trained models, while FSA-FGSM approach gives superior attack success rate for adversarially trained models than cleanly trained ones. For a fair comparison, we keep their averaged ASR comparable as: 27.8% (FGSM), 27.8% (SSA-FGSM) and 29.8% (FSA-FGSM), respectively.

Table 1: Attack success rate comparisons between FGSM and the proposed SSA-FGSM and FSA-FGSM methods. The attack success rate is in percent (%).

Attack	ResNet-101	ResNet-152	Inc-v4	Inc-v3$_{adv}$	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	Avg ASR
FGSM	33.2	32.0	**35.3**	22.7	25.3	18.1	27.8
SSA-FGSM	**34.9**	**34.4**	34.2	21.8	25.4	16.1	27.8
FSA-FGSM	28.4	27.2	29.7	**25.5**	**31.7**	**36.4**	**29.8**

Then we quantitatively assess the *visual superiority* of the proposed methods in Table 2. For IQA metrics, i.e., multiSim3, NIQE, FSIMc and MOS, the proposed SSA-FGSM achieves improvement by 3.4%, 5.2%, 0.45 and 1.09; and the proposed FSA-FGSM improves four IQA metrics by: 7.5%, 15.3%, 1.44, and 2.0, respectively. The quantitative comparison results confirm the significant perceptual improvement of the proposed methods over the vanilla FGSM attack. It is worthy to note that such visual improvement is obtained for free since we directly incorporate our strategies into vanilla FGSM. More importantly, compared with vanilla FGSM, the proposed methods require no sacrifice of the attack performance (i.e., average attack success rates).

Table 2: Visual quality comparisons between FGSM, SSA-FGSM and FSA-FGSM methods. The symbol "↑" ("↓") indicates that a higher (lower) value is better in perceptual quality.

Attack	multiSim3 (↑)	FSIMc (↑)	NIQE (↓)	MOS (↑)
FGSM	0.862	0.762	3.002	1.79
SSA-FGSM	0.896	0.814	2.664	2.88
FSA-FGSM	**0.937**	**0.915**	**1.560**	**3.79**

Comparison with MIM Attack [9]: To improve the adversarial transferability, momentum is introduced to obtain the iterative version of FGSM as Momentum Iterative Method (MIM):

$$\boldsymbol{x}^*_{t+1} = \boldsymbol{x}^*_t + \frac{\epsilon}{T} \cdot \text{sign}(\boldsymbol{g}_{t+1}) \tag{14}$$

where T denotes the number of iterations, and the accumulated gradient is updated as, $\boldsymbol{g}_{t+1} = \mu \cdot \boldsymbol{g}_t + \frac{\nabla_{\boldsymbol{x}} J(f_{\theta,\boldsymbol{x}^*_t,y})}{\left\|\nabla_{\boldsymbol{x}} J(f_{\theta,\boldsymbol{x}^*_t,y})\right\|_1}$. After getting the updated gradient, we incorporate structural-aware strategies into MIM, and obtain the proposed SSA-MIM and FSA-MIM methods. We use $\mu = 1.0$ as suggested in [9].

In Table 3, we compare the attack success rates of the MIM method and the proposed variants, e.g., SSA-MIM and FSA-MIM methods. The parameters for the two methods are $\alpha_0 = 2.3$, $\beta_0 = 6.0$ for a comparable attack success rate with respect to the vanilla MIM method, i.e., the averaged attack success rates are 38.6% for MIM, 38.8% for SSA-MIM, and 39.5% for FSA-MIM, respectively.

Table 3: Attack success rate comparisons between MIM and the proposed SSA-MIM and FSA-MIM methods. The attack success rate is in percent (%).

Attack	ResNet-101	ResNet-152	Inc-v4	Inc-v3$_{adv}$	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	Avg ASR
MIM	**46.8**	44.8	56.0	24.8	29.3	30.0	38.6
SSA-MIM	44.2	**46.0**	**56.1**	26.1	30.3	29.8	38.8
FSA-MIM	40.7	39.2	47.0	**34.3**	**34.5**	**41.2**	**39.5**

The quantitative IQA results are computed and reported in Table 4. Overall, SSA-MIM improves four metrics individually and FSA-MIM achieves even more improved perceptual qualities. Specifically, the quantitative IQA improvements are: 4.3% on multiSim3, 8.3% on FSIMc, 0.83 on NIQE and 1.68 on MOS, respectively. Therefore, the perceptual qualities of vanilla MIM can be largely improved by the utilization of the proposed structural-aware approaches.

Table 4: Visual quality comparisons between MIM, SSA-MIM and FSA-MIM methods. The symbol "↑" ("↓") indicates that a higher (lower) value is better in perceptual quality.

Attack	multiSim3 (↑)	FSIMc (↑)	NIQE (↓)	MOS (↑)
MIM	0.905	0.815	2.398	2.12
SSA-MIM	0.927	0.855	2.058	3.17
FSA-MIM	**0.948**	**0.928**	**1.569**	**3.80**

Comparison with DIM Attack [15]: In the DIM method, the inputs to the model are stochastically transformed copies of the original image to increase the adversarial transferability. At each iteration, correspondingly the gradient is updated with the transformation with a probability p. In the experiments, we selected p as 0.7 which was reported to achieve the highest averaged attack success rates [15]. Based on the DIM method, we have derivations of our proposed structure-aware variants, i.e., SSA-DIM and FSA-DIM methods.

Table 5: Attack success rate comparisons between DIM and the proposed SSA-DIM and FSA-DIM methods. The attack success rate is in percent (%).

Attack	ResNet-101	ResNet-152	Inc-v4	Inc-v3$_{adv}$	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	Avg ASR
DIM	**64.2**	62.8	73.6	31.6	32.6	32.1	49.5
SSA-DIM	62.8	**63.1**	**73.7**	33.6	35.3	33.4	50.3
FSA-DIM	53.0	52.2	60.5	**45.3**	**43.0**	**49.2**	**50.5**

To make the proposed attacks comparable with DIM [15] in ASR, we adopt $\alpha_0 = 2.35$, $\beta_0 = 6.5$ and show the attack success rate comparison between DIM, SSA-DIM and FSA-DIM methods in Table 5. The averaged attack success rates are 49.5%, 50.3% and 50.5%, respectively.

Compared with vanilla MIM (Table 3), vanilla DIM improves the averaged ASR by about 10% (Table 5). Correspondingly, we observe that the proposed attacks (i.e. SSA-DIM and FSA-DIM) also improve their ASRs over MIM-based methods (i.e. SSA-MIM and FSA-MIM) by a similar margin. This observation confirms that our proposed structure-aware strategies are indeed independent of gradient-based methods, i.e., the incorporation of perceptual models into existing attacks can still maintain their

Table 6: Visual quality comparisons between DIM, SSA-DIM and FSA-DIM methods. The symbol "↑" ("↓") indicates that a higher (lower) value is better in perceptual quality.

Attack	multiSim3 (↑)	FSIMc (↑)	NIQE (↓)	MOS (↑)
DIM	0.906	0.816	2.431	2.15
SSA-DIM	0.926	0.851	2.112	3.15
FSA-DIM	**0.941**	**0.921**	**1.684**	**3.58**

Table 7: ASR improvement comparisons between the baseline attacks and their SSA/FSA versions, with the comparable visual quality.

Attacks	SSA equivalent		FSA equivalent	
	FSIMc	ΔASR (%)	FSIMc	ΔASR (%)
FGSM	0.801	1.9	0.913	10.1
MIM	0.851	3.0	0.924	11.5
DIM	0.851	4.5	0.923	13.1

attack ability.

Meanwhile, we notice that vanilla DIM also suffers from the visual quality problem as reported in Table 6, e.g., the FSIMc is only 0.816. By contrast, SSA-DIM improves the metric by 3.9% and FSA-DIM further boosts its FSIMc metric to be 0.921.

Finally, we show several typical adversarial examples for qualitative visual comparison in Fig. 2. The first row depicts the clean images, and the last three rows display adversarial examples generated from DIM, SSA-DIM and FSA-DIM, respectively. For DIM, we observe the perceptual degradation phenomenon in adversarial images, especially in the smooth regions. In detail, the texture-like distortions make adversarial examples visually unpleasant and easily to be spotted (please zoom in Fig. 2 for better comparison). Compared with DIM, SSA-DIM clearly improves the perceptual quality by re-allocating larger perturbation budgets to those visual insensitive regions based on spatial perceptual models. In the last row, we observe that the proposed FSA-DIM produces adversarial examples with almost imperceptible visual quality. Therefore, with the proposed structure-aware strategies (i.e., SSA and FSA), we can achieve comparable attack success rates yet with significantly higher visual quality over baseline methods, both quantitatively and qualitatively.

In conclusion, this section shows experimental results and compare the perceptual improvement of the proposed structure-aware attacks with baseline attacks respectively. Overall, both the proposed spatial perceptual and frequency perceptual approaches can clearly improve the visual quality of adversarial examples with comparable average ASRs. Particularly for the frequency perceptual attacks, our proposed methods can generate almost imperceptible adversarial examples for each compared baseline attack.

5.4. ASR Improvement Assessment

In general, for the same attack, we can always maintain better visual quality (with less adversarial perturbations) at the expense of lower attack success rates [9]. In this section, we compare ASRs of three baseline methods and our proposed methods. To be specific, we decrease the perturbation budget ϵ of each baseline attack to make their IQA metrics comparable with the proposed methods individually. The IQA values of SSA and FSA have been reported in Table 2 - Table 6 as the comparison reference.

With comparable visual quality (e.g. FSIMc), we show the ASR improvement ΔASR in Table 7. For the spatial perception-based methods, ASR improvement ranges from 1.9% to 4.5%. For the frequency perception-incorporated methods, ASR improves over baselines by 10.1% to 13.1%. This comparison result reveals another superiority of the proposed methods: by incorporating the proposed structure-aware strategies, we can achieve higher ASRs than baselines with comparably good visual quality. Therefore we can conclude that, compared with baseline attacks, the proposed methods manage to obtain a better trade-off between the attack success rates and perceptual quality.

5.5. Perturbation Residues

To better understand the perceptual-based attacks, as an example we visualize perturbation residues of the DIM-based methods in Fig. 3. The parameters of attacks are the same as in Table 6. In general, we observe that DIM uniformly perturbs all pixels of the image which accounts for the visual degradation issue. By contrast, SSA-DIM mainly perturbs the visual insignificant regions which can be computed from spatial perceptual structures. Meanwhile, FSA-DIM approach perturbs the clean images with frequency insensitive adversarial perturbations in frequency perceptual bands, which generally appears invisible in the spatial domain.

5.6. Parameter Sensitivity

In this section, we study the effect of hyperparameters ϵ, α_0 and β_0 in the proposed attacks. To better illustrate the comparison trend of visual quality with respect to different hyperparameters, we normalize the NIQE values to be NIQE′: $NIQE' = 1 - NIQE/NIQE_{ub}$, where $NIQE_{ub}$ is an upper bound of $NIQE$ values for all experiments we conducted. A higher NIQE′ value indicates the better visual quality or vice versa.

In Fig. 4, we depict the parameter sensitivity curves for three baseline attacks (FGSM, MIM and DIM) and their perception-incorporated SSA/FSA based methods. The normalization factor $NIQE_{ub}$ equals 3.5 in the figures. In general, for each attack method, as hyperparameters (perturbation budgets) increase, the averaged ASRs increase at the expense of degraded visual quality (i.e. lower multiSim3, NIQE′ and FSIMc indices). We also observe that with comparable ASRs, the proposed methods consistently outperform their baselines. For instance, DIM achieves averaged ASR as 43.5% at $\epsilon = 10$ and its FSIMc equals 0.880. As a comparison, SSA-DIM produces averaged ASR to be 44.4% with FSIMc as 0.896 at $\alpha_0 = 1.75$. Meanwhile, FSA-DIM attains its ASR as **45.0**% with FSIMc equals **0.941** at $\beta_0 = 5.0$. The comparison results answer the question that *it is indeed possible to achieve a high ASR with improved visual quality*.

Fig. 2: Examples of perceptual image quality comparison between DIM, SSA-DIM and FSA-DIM methods. We recommend to zoom into the digital version for better visual comparison.

6. Conclusions and future work

In this work, we present two novel approaches to improve the perceptual quality of adversarial examples for deep networks in the transfer-based black-box setting. Since the existing uniform perturbation constraint does not align well with human visual systems, we explicitly consider the regional and structural information of images and incorporate the perceptual models into adversarial attacks. Specifically, we firstly introduce a spatial perceptual model and propose a structure-aware adversarial attack framework in the spatial domain. This framework is general and is compatible with all gradient-based attack methods. Further, we propose an adversarial attack framework by perturbing images in the frequency perceptual domain. Due to the structural constraints we explicitly consider, compared with baseline attacks, we demonstrate that adversarial examples produced by the proposed methods can generally have imperceptible or higher natural visual quality than the original attack methods with comparable attack success rates. Moreover, with comparable perceptual quality, the proposed methods produce higher attack success rates than baseline methods. In the future work, we plan to investigate and extend the proposed structure-aware frameworks to related tasks, e.g., imperceptible physical adversarial attacks.

Appendix A. Spatial JND Filters

In the spatial perceptual model (Section 3.1), the four high-pass oriented filters used in Eq.(3) are,

$$\boldsymbol{h}_1 = \frac{1}{16}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 3 & 8 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & -3 & -8 & -3 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \boldsymbol{h}_2 = \frac{1}{16}\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 8 & 3 & 0 & 0 \\ 1 & 3 & 0 & -3 & -1 \\ 0 & 0 & -3 & -8 & 0 \\ 0 & 0 & -1 & 0 & 0 \end{bmatrix}$$

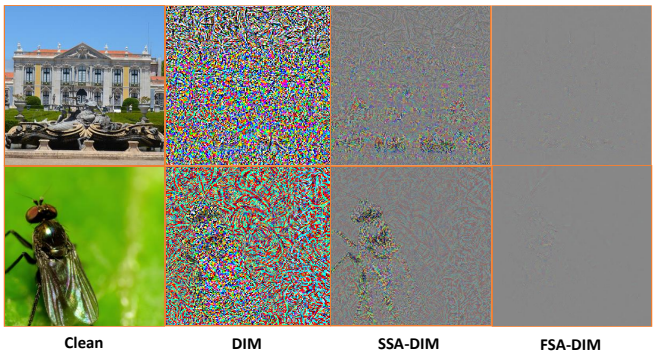

Fig. 3: Comparison of perturbation residues between DIM ($\epsilon = 14$), SSA-DIM ($\alpha_0 = 2.35$) and FSA-DIM ($\beta_0 = 6.5$) attacks on example images.

$$\boldsymbol{h}_3 = \frac{1}{16}\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 3 & 8 & 0 \\ -1 & -3 & 0 & 3 & 1 \\ 0 & -8 & -3 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{h}_4 = \frac{1}{16}\begin{bmatrix} 0 & 1 & 0 & -1 & 0 \\ 0 & 3 & 0 & -3 & 0 \\ 0 & 8 & 0 & -8 & 0 \\ 0 & 3 & 0 & -3 & 0 \\ 0 & 1 & 0 & -1 & 0 \end{bmatrix}$$

The parameters for the Gaussian low-pass filer $\boldsymbol{l_g}$ used in Eq.(4) are: 3×3 Gaussian kernel with mean as 0 and standard deviation as 0.5.

Appendix B. The MOS Test

To test the perceptual improvement of our proposed framework, we design a subjective test for perceptual image quality evaluation. We invite 10 volunteers to score the visual quality of the adversarial images. In each series of comparisons from Section 5.3, i.e. FGSM series, MIM series and DIM series, we randomly choose 50 adversarial examples from each adversarial attack method. For example, in the FGSM series, we randomly select 50 adversarial images generated by FGSM, 50 images generated by SSA-FGSM, and 50 images generated by FSA-FGSM. During the subjective test, we show a volunteer one pair of images and give her/him two seconds to review. The pair of images include an adversarial image and its corresponding clean image as reference. Finally the volunteer rates the adversarial image with a score. We repeat this process until all selected images are reviewed by this volunteer. The order of images for different volunteers are different. In this experiment, we employ the commonly used absolute category rating principle [42], with image quality score ranging from 1 to 5. The scores indicate:

- score = 1: visually bad and very disturbing;
- score = 2: poor visual quality with disturbing visual artifacts;
- score = 3: fair visual quality with acceptable perceptual distortion;
- score = 4: good visual quality with slight perceptual distortion;
- score = 5: excellent visual quality with almost imperceptible distortion.

Mean opinion score (MOS) is computed by averaging subjective scores from all volunteers for each adversarial attack method.

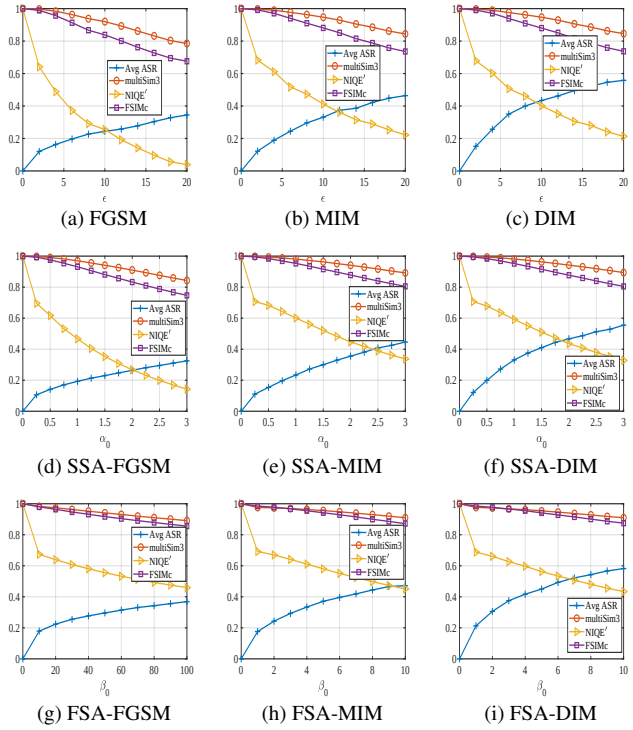

Fig. 4: Parameter sensitivity comparisons between the baseline methods (i.e. subfigure (a)-(c)) and the proposed SSA (i.e. (d)-(f)) and FSA (i.e. (g)-(i)) based approaches.

References

[1] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[2] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[3] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99.

[4] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[5] Y. Wang, H. Palangi, Z. J. Wang, H. Wang, Revhashnet: Perceptually dehashing real-valued image hashes for similarity retrieval, Signal processing: Image communication 68 (2018) 68–75.

[6] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, P. H. Torr, Fast online object tracking and segmentation: A unifying approach, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 1328–1338.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: International Conference on Learning Representations, 2014.
URL `http://arxiv.org/abs/1312.6199`

[8] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: 2017 ieee symposium on security and privacy (sp), IEEE, 2017, pp. 39–57.

[9] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, J. Li, Boosting adversarial attacks with momentum, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9185–9193.

[10] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, Robust physical-world attacks on deep learning visual classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1625–1634.

[11] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples (2015).

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, International Conference on Learning Representations (2018).

[13] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, A. Swami, The limitations of deep learning in adversarial settings, in: 2016 IEEE European symposium on security and privacy (EuroS&P), IEEE, 2016, pp. 372–387.

[14] Y. Qin, N. Carlini, I. Goodfellow, G. Cottrell, C. Raffel, Imperceptible, robust, and targeted adversarial examples for automatic speech recognition, nternational Conferenceon Machine Learning (2019).

[15] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, A. L. Yuille, Improving transferability of adversarial examples with input diversity, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2730–2739.

[16] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3207–3216.

[17] C.-H. Chou, Y.-C. Li, A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile, IEEE Transactions on circuits and systems for video technology 5 (6) (1995) 467–476.

[18] X. Yang, W. Ling, Z. Lu, E. P. Ong, S. Yao, Just noticeable distortion model and its applications in video coding, Signal Processing: Image Communication 20 (7) (2005) 662–680.

[19] N. Jayant, J. Johnston, R. Safranek, Signal compression based on models of human perception, Proceedings of the IEEE 81 (10) (1993) 1385–1422.

[20] X. Zhang, W. Lin, P. Xue, Improved estimation for just-noticeable visual distortion, Signal Processing 85 (4) (2005) 795–808.

[21] W. Lin, C.-C. J. Kuo, Perceptual visual quality metrics: A survey, Journal of visual communication and image representation 22 (4) (2011) 297–312.

[22] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, On evaluating adversarial robustness, arXiv preprint arXiv:1902.06705 (2019).

[23] A. Kurakin, I. Goodfellow, S. Bengio, Adversarial examples in the physical world, International Conference on Learning Representations Workshop (2017).

[24] B. Luo, Y. Liu, L. Wei, Q. Xu, Towards imperceptible and robust adversarial example attacks against neural networks, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[25] F. Croce, M. Hein, Sparse and imperceivable adversarial attacks, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4724–4732.

[26] Y.-J. Chin, T. Berger, A software-only videocodec using pixelwise conditional differential replenishment and perceptual enhancements, IEEE Transactions on Circuits and Systems for Video Technology 9 (3) (1999) 438–450.

[27] I. Hontsch, L. J. Karam, Locally adaptive perceptual image coding, IEEE Transactions on Image Processing 9 (9) (2000) 1472–1483.

[28] J. Canny, A computational approach to edge detection, IEEE Transactions on pattern analysis and machine intelligence (6) (1986) 679–698.

[29] A. Netravali, Digital pictures: representation and compression, Springer Science & Business Media, 2013.

[30] Y. Tsuzuku, I. Sato, On the structural sensitivity of deep convolutional networks to the directions of fourier basis functions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 51–60.

[31] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, K. Q. Weinberger, Simple black-box adversarial attacks, nternational Conferenceon Machine Learning (2019).

[32] Y. Sharma, G. W. Ding, M. Brubaker, On the effectiveness of low frequency perturbations, Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (2019).

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, 2019, pp. 8024–8035.

[35] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-first AAAI conference on artificial intelligence, 2017.

[36] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, P. McDaniel, Ensemble adversarial training: Attacks and defenses, International Conference on Learning Representations (2018).

[37] S. Escalera, M. Weimer, The NIPS'17 Competition: Building Intelligent Systems, Springer, 2018.

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International journal of computer vision 115 (3) (2015) 211–252.

[39] Z. Wang, E. P. Simoncelli, A. C. Bovik, Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Vol. 2, Ieee, 2003, pp. 1398–1402.

[40] L. Zhang, L. Zhang, X. Mou, D. Zhang, Fsim: A feature similarity index for image quality assessment, IEEE transactions on Image Processing 20 (8) (2011) 2378–2386.

[41] A. Mittal, R. Soundararajan, A. C. Bovik, Making a "completely blind" image quality analyzer, IEEE Signal processing letters 20 (3) (2012) 209–212.

[42] R. C. Streijl, S. Winkler, D. S. Hands, Mean opinion score (mos) revisited: methods and applications, limitations and alternatives, Multimedia Systems 22 (2) (2016) 213–227.