Demiguise Attack: Crafting Invisible Semantic Adversarial Perturbations with Perceptual Similarity

Yajie Wang[*,1], **Shangbo Wu**[*,1], **Wenyi Jiang**[1], **Shengang Hao**[1,3], **Yu-an Tan**[2] and **Quanxin Zhang**[†,1]

[1]School of Computer Science and Technology, Beijing Institute of Technology
[2]School of Cyberspace Science and Technology, Beijing Institute of Technology
[3]School of Computer Science and Technology, Nanyang Normal University
{wangyajie19, shangbo.wu, jiangwenyi2000, haoshengang, tan2008, zhangqx}@bit.edu.cn

Abstract

Deep neural networks (DNNs) have been found to be vulnerable to adversarial examples. Adversarial examples are malicious images with visually imperceptible perturbations. While these carefully crafted perturbations restricted with tight ℓ_p norm bounds are small, they are still easily perceivable by humans. These perturbations also have limited success rates when attacking black-box models or models with defenses like noise reduction filters. To solve these problems, we propose Demiguise Attack, crafting "unrestricted" perturbations with Perceptual Similarity. Specifically, we can create powerful and photorealistic adversarial examples by manipulating semantic information based on Perceptual Similarity. Adversarial examples we generate are friendly to the human visual system (HVS), although the perturbations are of large magnitudes. We extend widely-used attacks with our approach, enhancing adversarial effectiveness impressively while contributing to imperceptibility. Extensive experiments show that the proposed method not only outperforms various state-of-the-art attacks in terms of fooling rate, transferability, and robustness against defenses but can also improve attacks effectively. In addition, we also notice that our implementation can simulate illumination and contrast changes that occur in real-world scenarios, which will contribute to exposing the blind spots of DNNs.

1 Introduction

Precisely crafted perturbations added onto input data can easily fool DNNs [Szegedy *et al.*, 2014; Goodfellow *et al.*, 2015; Carlini and Wagner, 2017; Kurakin *et al.*, 2017]. This vulnerability that inherently exists inside DNNs is often exploited with adversarial examples, which are maliciously modified samples with imperceptible perturbations. Adversarial perturbations are crafted in a sense that they should be constrained within a bound that is *as tight as possible* [Goodfellow *et al.*, 2015; Kurakin *et al.*, 2017; Dong *et al.*, 2018], so that the perturbation is invisible to humans. In light of this, a majority of attacks often measure image similarity with ℓ_p norms — ℓ_0 [Carlini and Wagner, 2017; Papernot *et al.*, 2016], ℓ_2 [Carlini and Wagner, 2017] and ℓ_∞ [Kurakin *et al.*, 2017; Dong *et al.*, 2018]. However, we argue that ℓ_p norms have limitations.

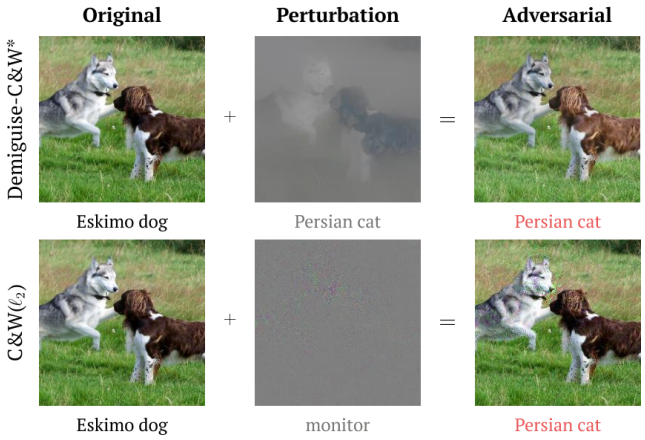


Figure 1: Demonstration of adversarial examples and perturbations crafted by Demiguise-C&W and C&W (ℓ_2) against ResNet-50. C&W (ℓ_2) crafts spottable perturbations with arbitrary noise. Demiguise-C&W crafts much larger perturbations with rich semantic information while maintaining imperceptibility.

Classic per-pixel measurements, such as the ℓ_2 norm distance, are insufficient for assessing structured data included in images because they assume pixel-wise independence. As such, ℓ_p-based perturbations in the RGB color space often have high spatial frequencies, which inevitably changes the spatial frequencies of natural images, making adversarial examples easily spottable by humans. In addition, the adversary would often need to generate larger perturbations in order to achieve higher adversarial strength. This naturally leads to more perceptible changes to the original image that are, again, noticeable by humans. This is the trade-off between adversarial effectiveness and perturbation imperceptibility, which will always exist if ℓ_p norm bounds are applied.

[*]Equal contribution.
[†]Corresponding author.

Many works on adversarial examples have also gained awareness that pixel-wise differences measured with ℓ_p norms coincide deficiently with human perception. Many of the non-ℓ_p solutions execute an attack by modifying the colors of the original image and alleviate the unrealistic factors of the changes with various mitigations, for instance: Semantic Adversarial Examples [Hosseini and Poovendran, 2018] and ColorFool [Shamsabadi *et al.*, 2020]. But for most of the time, these practices are not enough because many of the color changes are still easily perceptible. An optimal unrestricted adversarial attack should be able to exhibit powerful adversarial effectiveness on the premise that the perturbations are perceptually invisible. However, this unique challenge of comparing image similarity is still a wide-open problem. Because not only are visual patterns highly dimensional and highly correlated themselves, but the very measurement of visual similarity is often quite subjective if it's aimed to mimic human judgments.

In this paper, we propose to attenuate this conundrum by using Perceptual Similarity as a measurement for image similarity when crafting adversarial examples. Perceptual Similarity [Zhang *et al.*, 2018] is an emergent property shared across deep visual representations. It is a metric that utilizes deep features from trained CNNs to measure image similarity [Johnson *et al.*, 2016]. Specifically, we optimize perturbations with respect to Perceptual Similarity, creating a novel adversarial attack strategy, namely **Demiguise Attack**. By manipulating semantic information with Perceptual Similarity, Demiguise Attack can perturb images in a way that correlates extraordinary well with human perception such that the perturbations are imperceptible although they are of large magnitudes, as shown in the middle column of Figure 1. Our perturbations with high-order semantic information are even more likely to be classified into the same labels as original images or adversarial examples, which we further demonstrate in Section 4.2. Larger perturbations make the adversarial examples we craft more powerful, robust, and can even transfer from one task to another, which we discuss in Section 4.5. We also demonstrate that the effect of our approach is additive and can be used in combination with existing attacks to improve performances further. What's more, as shown in the rightmost column of Figure 1, by using Perceptual Similarity, our adversarial examples are somewhat able to simulate the illumination changes that occur in natural situations. This could be used for exposing the blind spots inside the target model and thereby helping it improve under real-world scenarios.

Our key contributions in this paper are:

- We propose a novel, unrestricted, black-box adversarial attack based on Perceptual Similarity, called Demiguise Attack. Our approach manipulates semantic information with the HVS-oriented image metric to craft invisible semantic adversarial perturbation.
- The perturbations generated with Perceptual Similarity can simulate the illumination and contrast changes in the real-world and enrich the semantic information lying within original images. This phenomenon implies enhancements that potentially exist for DNNs.
- Extensive experiments show that Demiguise Attack crafted adversarial perturbations both manifest excellent visual quality and boost adversarial strength and robustness when combined with existing attacks. We demonstrate Demiguise Attack's compelling maximum of 50% increase in terms of black-box transferability and the promising nearly 90% successful attacks under cross-task black-box scenarios.

2 Related Work

2.1 Adversarial Attack

ℓ_p norm-based adversarial attack. Most adversarial attacks utilize a form of ℓ_p norm-based distance metric, whether they are used as optimization objectives (such as L-BFGS [Szegedy *et al.*, 2014] and C&W [Carlini and Wagner, 2017]), or direct constraints (such as FGSM [Goodfellow *et al.*, 2015], BIM [Kurakin *et al.*, 2017], and MI-FGSM [Dong *et al.*, 2018]). These adversarial attacks, while being very effective in terms of fooling rate, often generate perturbations that have distinct characteristics — arbitrary multi-colored noise that is very obvious to human perception.

Non ℓ_p norm-based adversarial attack. Recent work suggests a different approach for adversarial attacks: unrestricted, semantic adversarial examples that are non-ℓ_p norm-based. Semantic Adversarial Examples [Hosseini and Poovendran, 2018] changes the colors of the image by shifting its hues and saturations in the HSV color space. ColorFool [Shamsabadi *et al.*, 2020] perturbs images within a chosen natural-color range for specific semantic categories. While these attacks are all non-ℓ_p based, they all fail to generate perturbations that align well with human perception, creating unrealistic adversarial examples easily noticeable by humans.

2.2 Perceptual Distance of Images

Measuring how similar are two images is often a subjective notion. Per-pixel measurements like the ℓ_2 Euclidean distance and Peak Signal-to-Noise Ratio (PSNR) have been proven to lack structural representation and fail to account for the many aspects of human perception. Perceptually motivated distance metrics like SSIM [Wang *et al.*, 2004] and FSIM [Zhang *et al.*, 2011] are still simple static functions that aren't ideal. Perceptual Similarity [Zhang *et al.*, 2018], which utilizes deep features from trained CNNs, models low-level perceptual judgments surprisingly well, outperforming previous widely-used metrics, as we demonstrate in the following section.

3 Methodology

3.1 Measuring Perceptual Similarity

The challenge of image similarity comparison has been a long-standing problem in the field of computer vision. It also has proven itself to be of significance for generating imperceptible adversarial perturbations. As stated in Section 2.2, classical measurements like ℓ_p-norm distances, PSNR and SSIM, are all inadequate to express the similarity of high-dimensional structured data like images that possess rich information. It would be ideal for constructing a "distance met-

ric" that can represent human judgments when measuring image similarity. However, implementing such a metric is challenging, as human judgments are context-dependent, rely on high-order image semantic information, and may not constitute a distance metric. As such, we propose to utilize Perceptual Similarity [Zhang *et al.*, 2018], a novel, HVS-oriented metric, to better craft adversarial examples by manipulating the semantic information lying within images. Here we briefly address concepts of Perceptual Similarity under adversarial settings.

Perceptual Similarity is a novel image quality metric that extracts characteristics from deep feature spaces of CNNs trained on image classification tasks. The metric itself is neither a special function nor a static module; instead, it is a consequence of visual representations tuned to be predictive about real-world structured information. Hence, we would be more capable of cultivating rich semantic information for crafting adversarial perturbation if we were to utilize Perceptual Similarity, thereby calibrating our perturbation to be in line with human perception. Notably, Perceptual Similarity is calculated based on a predefined and pretrained perceptual similarity network. For an original image $\boldsymbol{x}$ and its distorted partner $\boldsymbol{x'}$, with perceptual similarity network $\mathcal{N}$, we compute the distance between $\boldsymbol{x}$ and $\boldsymbol{x'}$ as

$$\mathcal{D}(\boldsymbol{x}, \boldsymbol{x'}) = \sum_{l} \frac{1}{H_l W_l} \cdot \sum_{h,w} \| \omega_l \odot (\hat{\theta}^l_{hw} - \hat{\theta'}^l_{hw}) \|^2. \quad (1)$$

where l is one of the layers in the L layers feature stack in network $\mathcal{N}$, $\hat{\theta}^l, \hat{\theta'}^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ are normalized features extracted from $\boldsymbol{x}$ and $\boldsymbol{x'}$ when they are passing through layer l (with H_l, W_l, C_l representing height, width, and channel for each layer l respectively), and ω_l is the vector that is used to scale the channel-wise activations. For adversarial attacks, $\boldsymbol{x}$ is the input image, and $\boldsymbol{x'}$ is the resulting adversarial example. In this setting, the adversary attempts to make the classifier mispredict by modifying $\boldsymbol{x}$ with a negligible perturbation, producing $\boldsymbol{x'}$ that is as close to $\boldsymbol{x}$ as possible — the optimization goal. The *closeness* between $\boldsymbol{x}$ and $\boldsymbol{x'}$ is, in our case, measured by Eq. 1. Hence, Eq. 1 also constitutes as a perceptual loss function. By optimizing against Perceptual Similarity, we can manipulate deep semantic features that exist within natural images to create perturbations that correlate well with human perception.

3.2 Demiguise Attack

Demiguise-C&W — Combining C&W with Demiguise Attack Demiguise Attack is a universal strategy for integrating Perceptual Similarity into adversarial attacks. We start with Demiguise-C&W, which, as its name suggests, is a variant of Demiguise Attack with its optimisation solved using techniques from C&W [Carlini and Wagner, 2017]. Classic adversarial attacks often use optimization procedures to craft perturbations. In order to satisfy the adversary's goal, one would optimize against an objective function in order to find minimal perturbations as

$$\boldsymbol{x}^{\text{adv}} = \underset{\boldsymbol{x'}: \mathcal{F}(\boldsymbol{x'}) \neq y}{\arg\min} \ \|\boldsymbol{x'} - \boldsymbol{x}\|_p. \quad (2)$$

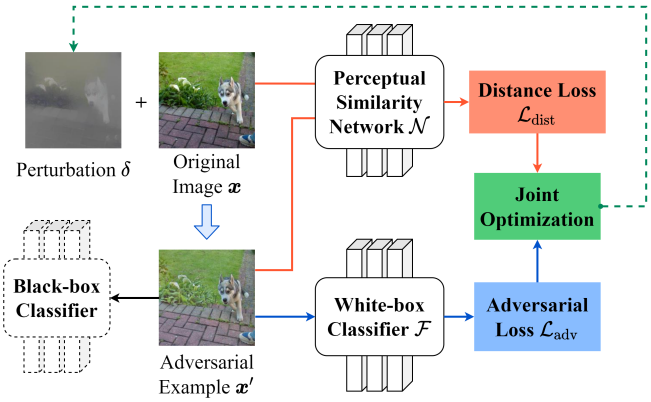


Figure 2: General optimization procedure of Demiguise Attack. For each iteration, we obtain both the $\mathcal{L}_{\text{dist}}$ from Perceptual Similarity network $\mathcal{N}$ as a penalty, and the $\mathcal{L}_{\text{adv}}$ from the classifier. Both of them are then optimized as $\mathcal{L} = \lambda \cdot \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{adv}}$ directly in the back-propagation procedure for crafting adversarial perturbations.

for non-targeted attacks, where $\boldsymbol{x}$ and $\boldsymbol{x'}$ are the original image and its (intermediate) adversarial example respectively, $\mathcal{F}$ is the target classifier, p is the norm used for distance calculation, and $\boldsymbol{x}^{\text{adv}}$ is the final adversarial example. This is basically an optimization against a loss function that perturbs the input image until it is adversarial, while also assuring that the perturbation is minimal. Formally, current adversarial attacks are optimizations against the joint loss of

$$\mathcal{L} = \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{dist}}. \quad (3)$$

in essence, where $\mathcal{L}_{\text{adv}}$ is the adversarial loss that guides the optimization procedure to making the adversarial example *adversarial*, and $\mathcal{L}_{\text{dist}}$ is the distance loss that minimizes the distance between the original input and the adversarial example. However, this optimization problem is often thought to be NP-hard, and as such, various mitigation measures have been proposed to solve it. C&W attack, as one of the most effective approaches, use an f function-interpreted cross-entropy loss as $\mathcal{L}_{\text{adv}}$

$$f(\boldsymbol{x'}) = \max(\max\{Z(\boldsymbol{x'})_i : i \neq t\} - Z(\boldsymbol{x'})_t, 0). \quad (4)$$

where $Z(\boldsymbol{x'})$ is the logit of $\boldsymbol{x'}$. Demiguise Attack applies optimizations against Perceptual Similarity with similar approaches as in C&W, formally expressed as

$$\underset{\boldsymbol{u}}{\text{minimize}} \ \lambda \cdot \mathcal{D}(\boldsymbol{x}, \boldsymbol{x'}) + f(\boldsymbol{x'}). \quad (5)$$

where a change of variables is applied, making $\boldsymbol{x'} = 1/2 \cdot \tanh(\boldsymbol{u}) + 1$, so that we optimise over $\boldsymbol{u}$ instead of $\boldsymbol{x'}$ directly. We take the five convolutional layers from the VGG architecture as the perceptual similarity network $\mathcal{N}$ with pretrained weights. The perceptual similarity network $\mathcal{N}$ exposes distance $\mathcal{D}(\boldsymbol{x}, \boldsymbol{x'})$ along with gradient information $\boldsymbol{g}$ so we can fully utilize it for crafting adversarial perturbations. The general Demiguise-C&W strategy is expressed in detail in Algorithm 1.

Combining other attacks with Demiguise Attack More importantly, we demonstrate that the strategy of Demiguise

Algorithm 1 Demiguise-C&W

Input: Input image $\boldsymbol{x}$, original prediction y, weight of Perceptual Similarity loss λ, number of iterations N;
Output: Adversarial example $\boldsymbol{x'}$;

Initialize: $\boldsymbol{x'_0} \leftarrow \boldsymbol{x}$, $\boldsymbol{u}_0 \leftarrow 0$;
Construct $\mathcal{D}$ — Perceptual Similarity instance with pretrained network $\mathcal{N}$;
for $i = 0$ **to** $N - 1$ **do**
  Initialize perturbation: $\boldsymbol{\delta}_i \leftarrow 1/2 \cdot \tanh(\boldsymbol{u}_i) + 1 - \boldsymbol{x}$;
  $\mathcal{L}_{\text{dist}} \leftarrow \mathcal{D}(\boldsymbol{\delta}_i + \boldsymbol{x}, \boldsymbol{x})$;
  $\mathcal{L}_{\text{adv}} \leftarrow f(\boldsymbol{\delta}_i + \boldsymbol{x})$. f is specified in Equation 4;
  Minimize $\boldsymbol{u}_i$ over $\mathcal{L} = \lambda \cdot \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{adv}}$;
  $\boldsymbol{x'_i} \leftarrow 1/2 \cdot \tanh(\boldsymbol{u}_i) + 1$;
  if $\mathcal{L}$ is not converging **then**
    Early stop and return $\boldsymbol{x'_i}$;
  end if
end for
return $\boldsymbol{x'} \leftarrow \boldsymbol{x'_i}$

Attack is additive, and it can be used in combination with existing attacks to improve performances further. We showcase our strategy's simplicity and universality by illustrating how to combine Demiguise Attack with existing attacks.

In Demiguise Attack, Perceptual Similarity is used directly as the distance penalty for optimization-based attacks. Besides the actual distance, we can also access gradient information from the Perceptual Similarity distance as $\boldsymbol{g} = \nabla_{\boldsymbol{x}} \mathcal{D}(\boldsymbol{x}, \boldsymbol{x'})$, which can then be used in gradient-based attacks. We address our general Demiguise Attack strategy (perturbation optimization procedure) in Figure 2.

With this, we can use Demiguise Attack in combination with existing state-of-the-art attacks, creating Demiguise-{C&W, MI-FGSM, HopSkipJumpAttack}. Here we briefly address their implementations. For C&W and HopSkipJumpAttack [Chen *et al.*, 2020], we combine their optimization procedure with Perceptual Similarity distance $\mathcal{D}$. For MI-FGSM, we update its loss as

$$\mathcal{L} = \nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x'}, y) + \lambda \cdot \nabla_{\boldsymbol{x}} \mathcal{D}(\boldsymbol{x'}, \boldsymbol{x}) \tag{6}$$

where $\mathcal{J}$ is our usual cross-entropy loss. For all Demiguise variants of these attacks, we introduce a new vector λ to balance the joint-optimization of perceptual loss and adversarial loss. Using Perceptual Similarity, Demiguise Attack crafts adversarial examples that dig deep into the rich semantic representations of images, achieving superior adversarial effectiveness while maintaining compelling imperceptibility.

4 Experiments

4.1 Experiment Setup

We use four different architectures of classifiers for evaluation: ResNet-{18, 50, 101, 152} [He *et al.*, 2016a], VGG-{11, 16, 19} [Simonyan and Zisserman, 2015], Inception-v3 [Szegedy *et al.*, 2016], MobileNet-v2 [Sandler *et al.*, 2018]. We use pretrained models from PyTorch's `torchvision` library. We randomly pick 1000 images of 10 separate classes from ImageNet [Deng *et al.*, 2009] that are all classified correctly by the models. All images are resized to $256 * 256$ and center cropped to $224 * 224$ in size, then normalized with `mean=[0.485, 0.456, 0.406]` and `std=[0.229, 0.224, 0.225]`. Our experiments are run on Ubuntu 20.04 LTS with NVIDIA GeForce RTX™ 3090 GPUs and 64GB of memory.

In terms of attacks, we choose a learning rate of 0.2 and a maximum of 1000 iterations for Demiguise-C&W. For Demiguise-MI-FGSM, we choose an ϵ of 0.4, a max iteration of 70 rounds, and a decay factor of 1.0. For Demiguise-HSJA, we choose a max iteration of 2000 queries. We use the ℓ_p-norm based versions of these attacks as baselines (including C&W [Carlini and Wagner, 2017], MI-FGSM [Dong *et al.*, 2018], and HopSkipJumpAttack [Chen *et al.*, 2020]), and we use the same hyper-parameters. Adversarial strength is evaluated in terms of the fooling rate of the attack, i.e., the proportion of successful cases over the whole validation set.

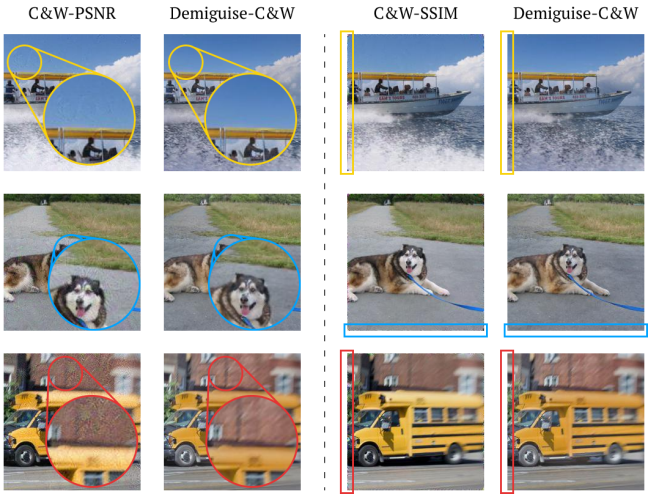


Figure 3: We compare perturbation imperceptibility of our Demiguise-C&W with C&W-PSNR and C&W-SSIM. We find that while all three attacks achieve 100% fooling rate, only Demiguise-C&W crafted adversarial examples are able to truly maintain perturbation imperceptibility.

4.2 Comparison of Different Perceptual Distances

We start our experiments by investigating whether Perceptual Similarity is the optimal metric for crafting adversarial perturbations with respect to human perception. We compare the aforementioned metrics, including PSNR and SSIM, with Perceptual Similarity. We extend C&W attack as C&W-PSNR and C&W-SSIM and attack the same models as Demiguise-C&W with the same parameters.

Here we illustrate a few of the adversarial examples crafted by these attacks. Despite the fact that PSNR and SSIM are some of the most commonly used metrics in computer vision, they still are shallow, facile functions that fail to account for the many factors of human perception. We can see that perturbations crafted by C&W-PSNR and C&W-SSIM contain high spatial frequencies, which changes the spatial frequencies of natural images, as shown in Figure 3. C&W-PSNR crafted examples have obvious arbitrary noise, as shown in the en-

Figure 4: Comparing perturbation crafted by Demiguise-C&W and C&W (ℓ_2) respectively for seven targeted white-box attack scenarios. Demiguise-C&W both generates perturbation that achieves compelling imperceptibility, and succeeds for all seven attacks.

larged region. Although C&W-SSIM's perturbations are less perceptible, a lot of noise is generated along the four sides of the image, as shown in the circled part. These noises make the perturbations spottable by humans and are the major drawbacks of these simple metrics. Conversely, Demiguise-C&W takes advantage of Perceptual Similarity, perturbing images by probing the rich semantic information within high-order structured representations, generating imperceptible perturbations. Besides, we couldn't help but notice that Demiguise Attacks' perturbations can somewhat simulate the natural illumination changes that occur in real-world situations. We believe that this peculiar phenomenon may suggest potential improvements that exist for current DNNs.

We are intrigued to see whether our perturbations crafted by Demiguise Attack actually mean anything to our "smart" classifiers. Thus, we passed our Demiguise Attack's perturbations through the classifiers alone and calculated the ratio of which the "perturbation" itself is classified as the same prediction of the original image or the adversarial example. We find that Demiguise Attack's perturbations attain ratios of about 80% for all four classifiers, more than 40% higher than those of C&W (ℓ_2) as shown in Table 1. Hence, we argue that Demiguise Attack is truly utilizing the high-dimensional semantic information that was not fully used by other attacks. This behavior of semantic information manipulation is what empowers Demiguise Attack to create large feature-rich perturbations that maintain excellent imperceptibility.

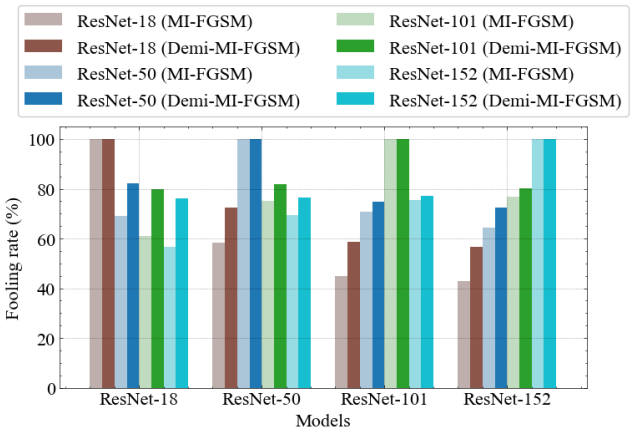
We further discuss the intriguing characteristics of our feature-rich perturbations crafted with respect to perceptual similarity. Here we demonstrate a comparison of perturbation between C&W (ℓ_2) and our Demiguise-C&W under targeted white-box scenarios. For the ground truth image that is correctly classified as "school bus" by ResNet-50, we attack it for a total of seven times, each targeted to a different class. The results are shown in Figure 4. Once again, we observe that our perturbation crafted by Demiguise-C&W supersedes those of C&W (ℓ_2) in terms of perturbation imperceptibility. Not only did Demiguise-C&W succeeded in all seven targeted attacks, but it also managed to maintain outstanding adversarial example quality, not to mention that ℓ_2-based C&W even failed twice in the seven attacks.

Attack	ResNet-50	VGG-19	Inception-v3	MobileNet-v2
Demi-C&W*	**81.11%**	**86.51%**	**72.51%**	**81.35%**
C&W (ℓ_2)	39.86%	48.30%	35.72%	40.60%

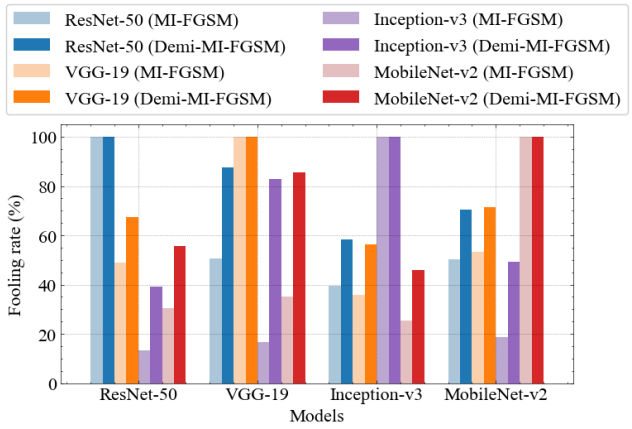
Table 1: We pass our crafted perturbations alone into the same classifiers, and see whether these perturbations contain semantic meanings to our classifiers. Demiguise-C&W crafted perturbations alone are classified into the same label as the original image or the adversarial example for a ratio of around 80%, which is over 40% higher than those crafted by ℓ_2-based C&W.

4.3 Extend Adversarial Attacks with Demiguise Attack

With the excellent performances of Demiguise Attack, we continue to combine our approach with other state-of-the-



(a) Transferability among ResNet models



(b) Transferability among models with different architectures

Figure 5: We compare the transferability of Demiguise-MI-FGSM with ℓ_∞-based MI-FGSM. Fooling rates that reach 100% are white-box attacks. Apart from these, we find that Demiguise-MI-FGSM (solid colors in the chart) outperforms ℓ_∞-based MI-FGSM (translucent colors in the chart) across all models in terms of transferability. Fooling rates can be increased by an average of 10% to 30% by incorporating Demiguise Attack strategy.

art attacks. First, we combine Demiguise Attack with MI-FGSM and investigate whether our strategy can further improve its black-box transferability. For this experiment, we use MI-FGSM and Demiguise-MI-FGSM to attack (1) the ResNet family (including ResNet-{18, 50, 101, 152}), and (2) ResNet-50, VGG-19, Inception-v3, MobileNet-v2. The fooling rates of these attacks under transfer-based black-box scenarios are shown in Figure 5a and Figure 5b respectively, where translucent-colored bars represent fooling rates of ℓ_p-based MI-FGSM, and solid-colored bars represent those of Demiguise-MI-FGSM. We find that the performance of MI-FGSM can be increased by an outstanding margin of 10% to 30% overall by utilizing our Demiguise Attack strategy despite whether models are of the same or different architectures.

Input diversity is another approach that has been proven

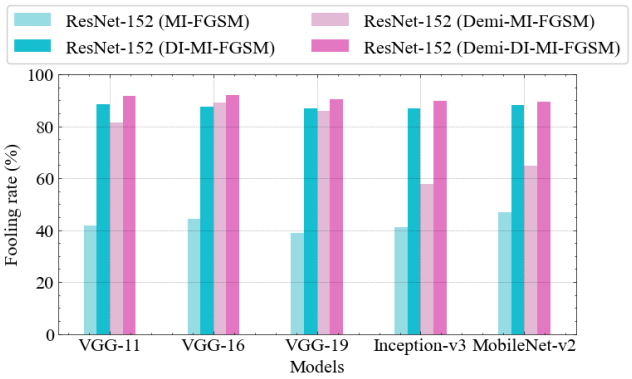


Figure 6: We testify if incorporating input diversity into Demiguise-MI-FGSM (ℓ_∞) can further improve transferability. We find that our final attack: Demiguise-DI-MI-FGSM (ℓ_∞), surpasses all other attacks' fooling rates, improving DI-MI-FGSM's performance by more than 5%, reaching a maximum transfer-based fooling rate of over 90%.

Attack	ResNet-50	VGG-19	Inception-v3	MobileNet-v2
ColorFool	48.1%	35.0%	40.4%	42.9%
HSJA (ℓ_2)	94.9%	92.2%	69.1%	89.7%
Demi-HSJA*	**94.9%**	**92.0%**	**99.1%**	**89.4%**

Table 2: We compare the fooling rates of ℓ_2-based HopSkipJumpAttack, Demiguise-HSJA, and ColorFool under black-box scenarios. The fooling rate of ColorFool barely reaches 50% on 1000 samples, while Demiguise-HSJA keeps up with HopSkipJumpAttack (ℓ_2), reaching fooling rates of over 89%.

to be of great contribution to the transferability of MI-FGSM. [Xie *et al.*, 2019] Here, we attack white-box ResNet-152, and transfer generated adversarial examples to VGG-{11, 16, 19}, Inception-v3 and MobileNet-v2. We compare the transferability of Demiguise-MI-FGSM, Demiguise-DI-MI-FGSM with baseline performances respectively. As shown in Figure 6, we can see that input diversity (DI-MI-FGSM) improves the transferability by a large extent, coming close to or even surpassing the performance of Demiguise-MI-FGSM. By incorporating our approach with input diversity, we can further increase the transferability for more than 5%, reaching fooling rates as high as 92%, creating one of the strongest attacks in terms of black-box transferability.

Finally, we combine Demiguise Attack with one of the most potent decision-based black-box attacks: HopSkipJumpAttack (HSJA). As previously mentioned, ColorFool [Shamsabadi *et al.*, 2020] is a typical non-ℓ_p based black-box adversarial attack where it only changes the colors of specific semantic categories in an image. We compare these attacks under black-box scenarios. A few of the examples crafted by these attacks are shown in Figure 7. We can easily see that ColorFool's color-changing strategy is not exactly ideal in terms of perturbation imperceptibility, generating colors beyond human comprehension, while Demiguise-HSJA crafted perturbations are imperceptible. What's more, ColorFool barely reaches fooling rates of over 50% over 1000 samples. In contrast, Demiguise-HSJA, with fooling

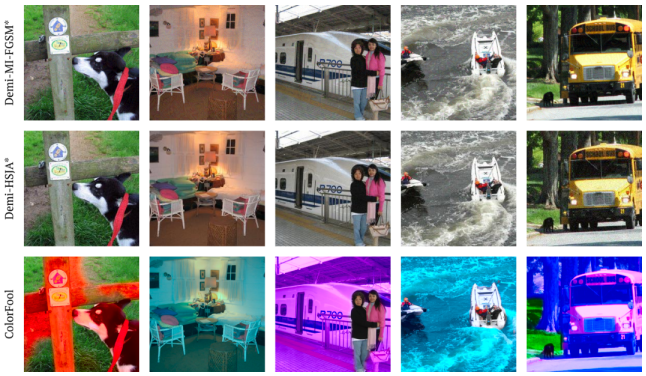


Figure 7: We demonstrate a few of the adversarial examples generated by Demiguise-{MI-FGSM,HSJA}, and ColorFool. Although ColorFool lives up to its promise of modifying specific semantic regions, the colors of its final examples are obvious and misleading.

Models	Attacks	No Defense	JPEG Compress	Binary Filter
ResNet-50	C&W (ℓ_2)	100%	82.09%	81.50%
	Demi-C&W*	100%	**85.10%**	**84.70%**
VGG-19	C&W (ℓ_2)	100%	84.40%	81.59%
	Demi-C&W*	100%	84.40%	**85.80%**
Inception-v3	C&W (ℓ_2)	100%	80.40%	88.80%
	Demi-C&W*	100%	**83.50%**	**92.20%**
MobileNet-v2	C&W (ℓ_2)	100%	76.00%	84.59%
	Demi-C&W*	100%	**76.70%**	79.30%

Table 3: We use JPEG compression and binary filters as defense schemes, and testify the fooling rates of C&W and Demiguise-C&W against these defenses. We find that Demiguise-C&W achieves 3% to 5% better robustness compared with ℓ_p-based C&W for most of the models with active defenses.

rates of over 89%, keeps up with the performance of HopSkipJumpAttack (ℓ_2). Detailed performances of ColorFool and Demiguise-HSJA are shown in Table 2.

4.4 Adversarial Robustness under Defense

In order to gain deeper insight into the effectiveness and strength of Demiguise Attack, we further testify the robustness of our attack against common adversarial defense strategies. Specifically, we utilize both JPEG compression [Das *et al.*, 2018] and binary filter [Xu *et al.*, 2018] as preprocessing measures that we implement directly into our aforementioned classifiers. We set JPEG compression quality as 75 and binary filter bit-depth as 4.

We demonstrate the fooling rates of C&W and Demiguise-C&W in Table 3. Due to the fact that defense schemes like JPEG compression and binary filter try to remove adversarial perturbations by image preprocessing filters, we find that Demiguise-C&W crafted adversarial perturbations tend to stay more intact than ℓ_p-based ones throughout defenses. We observe a 3% to 5% increase in fooling rates for Demiguise-C&W compared with ℓ_p-based C&W when attacking models with defenses.

Moreover, both JPEG Compression and Binary Filter have parameters that we can tweak: the compression ratio for JPEG Compression, and the bit depth for Binary Filter. For JPEG compression, we set the compression ratio from 100 to 10 with a step size of 15. For Binary Filter, we set its bit depth from 7 to 1 with a step size of 1. The results of JPEG Compression as a defense scheme are shown in Figure 8a, and the results of Binary Filter are shown in Figure 8b. We can see for both of these defense strategies, the lower image quality (i.e., lower JPEG compression ratio or the smaller bit depth), the more all fooling rates decreases. This characteristic is especially obvious for Binary Filter, where we see clear drops of our fooling rates when bit depth decreases. Nevertheless, when image quality is still reasonable, we find that Demiguise-C&W still achieves higher fooling rates than vanilla C&W for all four classification models most of the time. We argue that the utilization of perceptual similarity-based optimization rewards us with this boost in adversarial robustness and effectiveness.

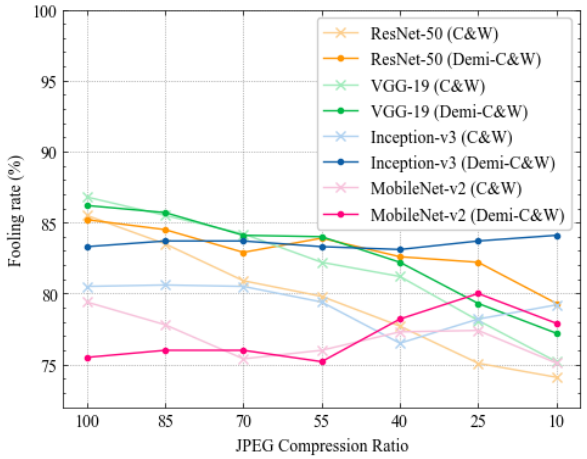
Target Classifiers	Attacks	Black-box Detectors	Label Vanish	Mis-label	Extra Object	Success Cases
ResNet-50	Demi-C&W	YOLOv3	197	86	88	25.0%
		Faster R-CNN	337	215	187	**47.8%**
	Demi-MI	YOLOv3	777	201	17	**89.8%**
		Faster R-CNN	776	111	9	82.7%
VGG-19	Demi-C&W	YOLOv3	413	557	140	**78.8%**
		Faster R-CNN	579	395	107	78.2%
	Demi-MI	YOLOv3	723	282	21	**89.5%**
		Faster R-CNN	746	182	13	82.5%
Inception-v3	Demi-C&W	YOLOv3	404	557	144	**78.2%**
		Faster R-CNN	568	400	101	**78.2%**
	Demi-MI	YOLOv3	766	209	18	**89.1%**
		Faster R-CNN	783	125	7	82.4%
MobileNet-v2	Demi-C&W	YOLOv3	401	561	144	78.0%
		Faster R-CNN	570	409	106	**78.3%**
	Demi-MI	YOLOv3	759	218	17	**89.1%**
		Faster R-CNN	762	126	17	82.6%

Table 4: We investigate the cross-task transferability of our Demiguise Attack on 1000 samples. The number of successful cases is calculated by counting the unique cases of which a sample is either mis-labeled or not labeled at all (label vanish) by the detector.

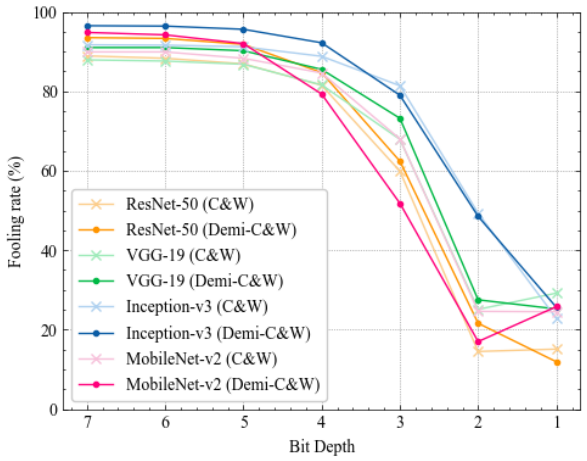
4.5 Cross-task Transferability

Finally, we testify the cross-task transferability of Demiguise Attack. We use Demiguise-{C&W, MI-FGSM} crafted examples on the same aforementioned classifiers and transfer them to black-box object detectors. Specifically, we use PyTorch's official Faster R-CNN [Ren *et al.*, 2015] (with a backbone of ResNet-50) and weights pre-trained on COCO, and we use the original weights for YOLOv3 [Redmon and Farhadi, 2018] as well. We total the number of cases where a sample is (1) not labeled at all (label vanished), (2) mislabeled, and (3) detected with extra objects. We consider the former two scenarios as successful attacks. We sum up the total number of unique samples falling in the former two cases as successful cases.

As in Table 4, both Demiguise-{C&W, MI-FGSM} crafted adversarial examples on all four target classifiers are able to fail detectors to a certain extent under cross-task black-box scenarios. We find that Demiguise-MI-FGSM can craft more transferable perturbations: of the 1000 samples, up to nearly



(a) JPEG Compression



(b) Binary Filter

Figure 8: We compare the adversarial robustness of Demiguise-C&W with C&W (ℓ_2) with defenses of different parameters. When image quality is reasonable, we find that no matter what type of defense scheme is applied, Demiguise-C&W always achieves higher fooling rates than ℓ_2-based C&W.

90% are successful, which is over 100 more than Demiguise-C&W on average. We also argue that attacking white-box classifiers and transferring adversarial examples to black-box detectors with the same backbone networks don't necessarily contribute to more transferability: Demiguise-C&W attacking ResNet-50 don't generate as many samples that can fail detectors as the other three classifiers, even though they share the same backbone network. Nevertheless, we demonstrate that Demiguise Attack is able to achieve outstanding transferability even under cross-task black-box scenarios.

4.6 Competition

Proposed method ranked third in the Security AI Challenger VI Track II: Unrestricted Adversarial Attacks on ImageNet in the CVPR2021 [Dong *et al.*, 2021]. We implement our approach for attacking an ensemble of EfficientNet [Tan and Le, 2019], ResNet-50 [He *et al.*, 2016b], VGG-16 [Simonyan and Zisserman, 2014] and ViT [Dosovitskiy *et al.*, 2020]. And we adopt the ensemble in the logits.

5 Discussion

We mentioned in Section 1 that our approach crafts perturbations with rich semantic information, and can somewhat simulate illumination changes happening in real-world scenarios. We would like to further discuss this notion. Demiguise Attack's adversarial perturbations are optimised against Perceptual Similarity. Perceptual Similarity aligns with human perception because it utilizes deep features inside the VGG network, which learns a representation of the natural world that correlates well with perceptual judgement. Real-world objects constantly appear different because of illumination changes, and we happen to find our perturbations very similar to these. This is the primary reasoning for us to claim that our perturbations simulate some natural phenomena or photographic effects in real life. In addition, as our experiment results in Table 1 show, our perturbations actually contain semantic meanings to classifiers, which further implies that unknown blind spots may inherently exist inside DNN-based classifiers, and may indicate potential robustness enhancements for current widely-used DNNs.

6 Conclusion

In this paper, we propose a novel, unrestricted black-box adversarial attack based on Perceptual Similarity — Demiguise Attack. Our approach can not only craft largely imperceptible perturbation by manipulating deep semantic information in high-dimensional images but also contribute to extensive enhancements when combined with existing state-of-the-art adversarial attacks. Demiguise Attack boosts adversarial strength and robustness, increases transferability for a maximum of 50%, and shows promising cross-task transferability performances. Besides, our findings imply that semantic defenses potentially exist for DNNs, which will be discussed in our future work.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 61876019, No. U1936218, and No. 62072037). We thank the security AI challenger program launched by Alibaba Group and Tsinghua University.

References

[Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[Chen *et al.*, 2020] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pages 1277–1294. IEEE, 2020.

[Das *et al.*, 2018] N. Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, S. Li, L. Chen, Michael E. Kounavis, and Duen Horng Chau. Shield: Fast, practical defense

and vaccination for deep learning using jpeg compression. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[Deng *et al.*, 2009] Jia Deng, W. Dong, R. Socher, L. Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR 2009*, 2009.

[Dong *et al.*, 2018] Y. Dong, Fangzhou Liao, Tianyu Pang, H. Su, J. Zhu, Xiaolin Hu, and J. Li. Boosting adversarial attacks with momentum. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018.

[Dong *et al.*, 2021] Y. Dong, Q. Fu, and X. Yang. Alibaba security: Adversarial robustness benchmark. https://s.alibaba.com/benchmark, 2021. Accessed: 2021-05-16.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Goodfellow *et al.*, 2015] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015.

[He *et al.*, 2016a] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[He *et al.*, 2016b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hosseini and Poovendran, 2018] H. Hosseini and R. Poovendran. Semantic adversarial examples. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1695–16955, 2018.

[Johnson *et al.*, 2016] J. Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[Kurakin *et al.*, 2017] A. Kurakin, Ian J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *ArXiv*, abs/1607.02533, 2017.

[Papernot *et al.*, 2016] Nicolas Papernot, P. McDaniel, S. Jha, Matt Fredrikson, Z. Y. Celik, and A. Swami. The limitations of deep learning in adversarial settings. *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387, 2016.

[Redmon and Farhadi, 2018] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *ArXiv*, abs/1804.02767, 2018.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[Sandler *et al.*, 2018] Mark Sandler, A. Howard, Menglong Zhu, A. Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[Shamsabadi *et al.*, 2020] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and A. Cavallaro. Colorfool: Semantic adversarial colorization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1148–1157, 2020.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Simonyan and Zisserman, 2015] K. Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

[Szegedy *et al.*, 2014] Christian Szegedy, W. Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2014.

[Szegedy *et al.*, 2016] Christian Szegedy, V. Vanhoucke, S. Ioffe, Jon Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[Tan and Le, 2019] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[Wang *et al.*, 2004] Zhou Wang, A. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

[Xie *et al.*, 2019] Cihang Xie, Zhishuai Zhang, Jianyu Wang, Yuyin Zhou, Zhou Ren, and A. Yuille. Improving transferability of adversarial examples with input diversity. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2725–2734, 2019.

[Xu *et al.*, 2018] Weilin Xu, David Evans, and Y. Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *ArXiv*, abs/1704.01155, 2018.

[Zhang *et al.*, 2011] L. Zhang, Lei Zhang, X. Mou, and D. Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20:2378–2386, 2011.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.