Abstract

Traditional pixel-wise image attack algorithms suffer from poor robustness to defense algorithms, i.e. the attack strength degrade dramatically when defense algorithms are applied. Although Generative Adversarial Networks (GAN) can partially address this problem by synthesizing a more semantically meaningful texture pattern, the main limitation is that existing generators can only generate images of a specific scale. In this paper, we propose a scale-free generation-based attack algorithm that synthesize semantically meaningful adversarial patterns globally to images with arbitrary scales. Our generative attack approach consistently outperforms the state-of-the-art methods on a wide range of attack settings, i.e. the proposed approach largely degraded the performance of various image classification, object detection and instance segmentation algorithms under different advanced defense methods.

1. Introduction

Deep Neural Networks (DNNs) are vulnerable to adversarial examples, e.g., images with carefully designed adversarial perturbations can easily mislead well-trained DNNs to output incorrect predictions. To overcome such malicious attacks, several adversarial defense algorithms have been proposed, which, in turn, simulate the development of robust adversarial attack algorithms in order to disrupt these defenses. Therefore, investigating robust and powerful image attack algorithms plays a crucial role in progressing current research towards developing strong defense algorithms.

*Corresponding author.
†Corresponding author.
vignetting [56] and haze [19]) without potential to be further extended. Although some approaches attempt to create local perturbations that are unrelated to the context of the target image [33, 36, 38, 60, 62], they usually make the attacked image easily identifiable to both humans and defense algorithms.

Alternatively, global attack patterns can be achieved by texture modification [28, 61] or color shifting [4], which attacks the target at the whole image-level. Despite such solutions being more robust to defense algorithms compared to pixel-level approaches, they are still perceptually unnatural and identifiable to humans (Problem 3). Besides, other researchers also resort to synthesizing semantically meaningful and global image attack patterns [3, 37, 45]. While these generator-based approaches provide more robust and general image attack patterns, the main shortcoming is that they can only produce a fixed-shape (i.e., image dimensions) global attack pattern depending on the pre-trained generator (Problem 4). Although the generated attack pattern can be resized to the shape of the target image, this would distort the generated attack pattern, resulting in the degradation of the attack strength and visual quality.

In this paper, we aim to address the aforementioned main problems of existing image attack algorithms by developing an PQ-GAN which generates generic and robust attack pattern of any shape.

We first draw our attention towards some natural patterns

We first draw our attention towards the fact that the rain is not only a common natural phenomenon that frequently appears in different real-world scenes, but also is shown to be robust to various defense algorithms [24], i.e., the performance of the predictor usually degrades heavily when images portray synthetically added rain patterns. In this regard, we hypothesize that the rain streaks pattern can be used to attack several types of images. Besides, some patterns result from other natural phenomenon, e.g., rain drops, snow flakes, or from the detects of digital images, e.g., camera lens dirt, also commonly exist in digital images. We suspect that deep neural networks these can also be vulnerable to these types of patterns.

In particular, we propose a photo-realistic adversarial pattern attack approach that fools classifiers by synthesizing realistic, semantically meaningful and whole image-level patterns added to the target images. Realistic and Semantically meaningful pattern ensure the visual quality of the adversarial examples (address and problem 3) and conform the local smoothness of natural images thereby difficult to be defensed (addressing the problem 1). In addition, our methods can be applied to various patterns, e.g., rain streaks, rain drops, snow flakes, lens dirt, etc, which extend the generalization capabilities of existing synthesis-based adversarial attack methods (addressing the problem 2).

To address the problem 4, we extend the basic generative adversarial networks which can only generate patterns of a certain scale to an architecture that is able to generate patterns of arbitrary scale, namely Patch Quilting Generative Adversarial Network (PQ-GAN).

The main contributions of the proposed approach is summarized as follows:

- We propose an adversarial attack strategy that takes the advantage of generative networks (generating images conforming the distribution of the target patterns with great variety) into adversarial attack, further adapting the attack pattern to the target image.

- We propose a novel Patch Quilting Generative Adversarial Network (PQ-GAN) training strategy that learns a set of cascaded generators to manipulate images of varying scales using the realistic and adversarial pattern without any distortion or discontinuity.

- The principle investigation results demonstrate that our approach delivers the state-of-the-art attack strength and our experiments verify the effectiveness of the proposed approach against various types of passive defense algorithms.

2. Related Work

2.0.1 Adversarial Attack & Defense

Depth Neural Networks (DNNs) are vulnerable to adversarial examples, i.e., some carefully designed perturbation being added to the images can mislead the networks to give unexpected outputs. Traditional adversarial attack strategies [5, 12, 14, 22, 54] are extensively researched to generate adversarial examples by adding $L_p$ bounded adversarial perturbations to the benign images. However, these strategies usually lack the robustness to the defense algorithms. [10, 35, 41, 53, 65]. Therefore, it is significant to study the "unrestricted" adversarial examples which contain unrestricted magnitude of perturbation while still preserve perceptual realism. Duan et al. [15] replaces the $l_p$ constraint with the perceptual similarity networks but still use pixel-wise perturbation, which according to Dong et al. [13], violate the local smoothness of natural images, thus still not robust enough. To solve this problem, [13] propose a superpixel-guided adversarial attack algorithms to generate superpixel-wised perturbations which is smooth on most area. To further improve the robustness, some color/texture shifting methods were proposed [4, 26, 28, 52, 61]. However, these methods are either perceptually unnatural or computationally costly due to selective search. Besides, some methods [19, 45, 56] generate patterns that rely on specific math formula while some researchers synthesize semantic patterns [3, 32, 37, 63] using traditional GAN-like structure,
which is only able to attack images of a certain scale or attack a certain type of images (e.g., human face [32, 63], therefore lack generalization capabilities.

Adversarial Defense algorithms can be roughly categorized into two types [34]: active defense (or model optimization) and active defense (or input optimization). Active defense algorithms optimize the model by adding additional training samples (adversarial training [2]) or modifying the network structure (Defensive distillation [46], gradient masking [18], etc.). This process can be costly or hardly be applied to the deployed models. Passive defense algorithms pre-process the input data and remove the artifacts before putting into the reference (victim) network. Some works try to remove the noise by Jpeg compression [10], high-frequency suppression [65], high-level representation de-noising [35], encoder-decoder structure [41], or GAN structure [53]. These noise removal strategies have prominent effects on defending noise-based attacking methods, but less effective on unrestricted semantic attack. Samangouei et al. [49] propose Defense-gan that removes the artifacts on the adversarial examples by learning the distribution of the benign images. However, we experimentally show that our attack method is robust to this defense strategy.

2.0.2 Generative Models and Texture Synthesis:

Generative Adversarial Networks (GAN) [21] has been widely used in many areas due to its capability of learning and generating various data distributions. Some works use hierarchical [6, 29, 50] or growing convolution [30] architectures to generate images with progressive resolution, therefore able to generate images of different resolution by picking the feature map of a certain layer without retraining. However, none of them can control the shape of the view frustum, and in practice, the resolution cannot be increased with no limit due to the capacity of the generation networks. Hence, we aim to develop an algorithm that generate texture patterns of various shapes of view frustum with any resolution in a differentiable manner. In specific, we need to naturally repeat a texture pattern to form a image of an arbitrary shape with great variety and the whole process must be differentiable. For this reason, we develop Patch Quilting GAN (PQ-GAN).

2.0.3 Traditional Image Quilting Algorithms:

Efros et al. [16] propose a seamless image quilting algorithm which quilts two independent patches along the path of minimum pixel difference on the overlapping area. Alternatively, Aguerrebere et al. [1] iteratively samples pixels (original) or patches (optimized) of the top few similarities and copy them to the blank area. As traditional computer vision algorithms, they are non-differentiable, therefore barely implementable in adversarial attack algorithms. In contrast, PQGAN is a GAN-based architecture that can efficiently generate texture patterns of any shape with great variety, and it is differentiable.
3. Methodology

In this section, we present our novel image attack pattern generation approach which can generate a strong and photo-realistic global attack patterns with an arbitrary shape, which are used to attack images with varying dimensions. Building on any pre-trained global image attack generator \( G \) that can produce photo-realistic attack pattern \( F^G \) to attack images of size \( h \times w \), our Patch Quilting Generator (PQG) takes a latent vector \( Z \) and a condition vector \( C \) (optional) as input and generate a set of attack patterns \( F^G_{m,n} \), where \( m, n \) denotes the \( m \)th rows and \( n \)th column while quilting them as an photo-realistic attack pattern of a customized shape \( H \times W \). \( H \) and \( W \) can be adapted to the target image \( I \). Instead of generating them individually and simply concatenating all pattern patches, which may lead discontinuity of patches in the final global attack pattern, making the attacked image unrealistic and easily recognizable to defense algorithms, the proposed concatenation is implemented by our well-trained (PQG), allowing the final produced attack pattern to be smooth (i.e., spatially and semantically continuous).

3.1. Assumption of Conditional Independency

The characteristics of photo-realistic pattern such as rain drops, snow, etc. are local dependency and non-local independency, e.g., for a rain drops pattern, the pixel value of one rain drop is dependent to the value of other pixels of the same rain drop but barely depends on those of other rain drops. Second, the patches of a pattern are repetitive but not identical, e.g., different rain drops patterns has similar but non-identical distributions. Due to these two characteristics, we can treat any pattern \( P \) of shape \( H \times W \) as quilted by small patches \( P_{m,n} \) of shape \( h \times w \). By choosing an appropriate patch size \( h \times w \), we can assume that the distribution of each patch \( P_{m,n} \) is only dependent to their neighbors, i.e.,

\[ P_{m,n} \neq P_{m',n'} \Rightarrow m' \in \{ m \pm 1, m \}, n' \in \{ n \pm 1, n \} \]

Subject to \( m, n \leq H \) and \( m, n \leq W \).

3.2. Patch Quilting generator

Our Patch Quilting (PQ) generator PQG is made up of three cascaded conditional generators \( G_1 \), \( G_2 \) and \( G_3 \), each of which is trained to output a realistic adversarial attack pattern with the shape equaling to \( h \times w \). In particular, the PQG takes the \( C \) and \( Z \) as inputs and outputs an photo-realistic global attack pattern \( P^I \) whose shape is same as that of the target image \( I \). This can be formulated as:

\[ P^I = \text{PQG}(C, Z) \]

\[ \text{PQG} = \{ G_1, G_2, G_3 \} \]  (2)

Supposing that the target image \( I \) has the shape of \( H \times W \), we create a raw pattern \( P^\text{raw} \in \mathbb{R}^{H \times W} \) that has equal or larger shape than the target pattern \( P^I \), which is denoted as:

\[ \hat{H} = \text{ceil} \left( \frac{H}{h} \right) \times h, \quad \hat{W} = \text{ceil} \left( \frac{W}{w} \right) \times w \]  (3)

where ceiling rounds the value to the nearest integer greater than or equal to it. Then, as shown in Fig. 1, the \( G_1 \) generates a set of independent attack patterns of the shape \( h \times w \) based on the \( C \) and \( Z \):

\[ P_{m,n}^G = G_1(C, Z) \]  (4)

where \( P_{m,n}^G = P_{m,n}^G_1, P_{m,n}^G_2, ..., P_{m,n}^G_{N_1} \). The PQG applies the attack patterns in \( P_{m,n}^G \) to fill up a set of non-adjacent regions of the \( P^\text{raw} \) (depicted as blue in Fig. 1).

Then, the \( G_2 \) generates a set of horizontal context-aware realistic adversarial attack patterns \( P^G_{m,n} = P_{m,n}^G_1, P_{m,n}^G_2, ..., P_{m,n}^G_{N_2} \), which fill up each horizontal gap in \( P^\text{raw} \) (depicted as yellow in Fig. 1) by considering not only \( C \) and \( Z \) but also the the horizontal attack pattern neighbours. Specifically, the attack patterns in \( P^G_{m,n} \) are employed to fill up the region \( P^\text{raw}_{m,n} \) in \( P^\text{raw} \), where \( P^\text{raw}_{m,n+1} \) and \( P^\text{raw}_{m,n-1} \) are already filled by attack patterns in \( P^G_{m,n} \) as well as \( C \) and \( Z \):

\[ P^\text{raw}_{m,n} = G_2(C, Z, P^\text{raw}_{m,n-1}, P^\text{raw}_{m,n+1}) \]  (5)

\[ \text{Subject to} \quad P^\text{raw}_{m,n} \in P^G_{m,n} \]

Meanwhile, the \( G_3 \) generates a set of vertical context-aware realistic adversarial attack patterns \( P^G_{m,n} \), targeting with filling up the rest regions (vertical gaps) in \( P^\text{raw} \) (depicted as red in Fig. 1), where the generation of each attack pattern which fills up the \( P^\text{raw}_{a+1,b} \) in \( P^\text{raw} \) is conditioned on its vertical patch neighbours \( P^\text{raw}_{a,b} \) as well as \( C \) and \( Z \):

\[ P^\text{raw}_{m,n} = G_3(C, Z, P^\text{raw}_{m,n-1}, P^\text{raw}_{m,n+1}) \]  (6)

Subject to \( P^\text{raw}_{m,n} \in P^G_{m,n} \)

Consequently, the global photo-realistic adversarial attack pattern \( P^\text{raw} \) is obtained by concatenating attack patterns produced by \( G_1 \), \( G_2 \), and \( G_3 \). We then remove the extra pixels of the \( P^\text{raw} \in \mathbb{R}^{H \times W} \) to make it have the same size \( (H \times W) \) to the target image \( I \), which is denoted as the final \( P^I \). In summary, the proposed PQG can not only synthesize global image attack patterns of any required shape without requiring re-training the network, but also allow the final produced attack pattern to be smooth, continuous and semantically meaningful.

3.3. Generative adversarial training for PQG

This section presents a novel generative adversarial strategy for training our PQG (we coin it as the PQ-GAN in this
Rain drops in this paper), which optimizes $G_1$, $G_2$ and $G_3$ in an end-to-end manner. This strategy would enforce $G_1$, $G_2$ and $G_3$ to generate attack patterns that are spatially correlated, allowing the concatenated global attack pattern to be continuous and smooth.

As illustrated in Fig. 1, the attack patterns produced by each generator are individually fed to a discriminator $D$ to classify whether they have the same distribution as the target pattern. This enforces not only all generators to produce photo-realistic attack patterns, but also the attack patterns produced by $G_2$ and $G_3$ to be smooth and continuous with their adjacent attack patterns in $P^{raw}$. To further ensure the smoothness and continuity of the $P^{raw}$, we also randomly crop a set of patches (denoted as $P^c$ in this paper) from it, all of which have the same size as the outputs of $G_1$, $G_2$ and $G_3$. Each of these cropped patches contains contents from at least two attack patterns in $P^{G_1}$, $P^{G_2}$ and $P^{G_3}$. We then again feed patches in $P^c$ to the discriminator $D$, enforcing the entire raw attack pattern $P^{raw}$ that jointly produced by three generators is photo-realistic, continuous and smooth.

The optimization process of $j$th $(j = 1, 2, 3)$ generator is supervised by two parts: the generative adversarial loss $\mathcal{L}_{PQGAN}$ obtained from the $j$th step in Fig. 1 (denote as $\mathcal{L}^j$) and the joint generative adversarial loss obtained from the $4$th step in Fig. 1 (denote as $\mathcal{L}^c$), which can be formulated as:

$$\mathcal{L}_{PQGAN} = \sum_{j=1}^{3} \mathcal{L}^j + \mathcal{L}^c$$  \hspace{1cm} (7)

where the generative adversarial loss $\mathcal{L}^j$ and the joint generative adversarial loss $\mathcal{L}^c$ are obtained by the generator $G_j$ and the discriminator $D$ as:

$$\mathcal{L}^j = \mathcal{L}^j_{G_j} + \mathcal{L}^j_D$$  \hspace{1cm} (8)

$$\mathcal{L}^c = \mathcal{L}^c_{G_1} + \mathcal{L}^c_{G_2} + \mathcal{L}^c_{G_3} + \mathcal{L}^c_D$$

where $\mathcal{L}^j_A$ denotes the loss of patch $P^a$ being used to update the model $A$; $\mathcal{L}^j_{G_j}$, $\mathcal{L}^j_D$, and $\mathcal{L}^c_D$ denote the standard generator and discriminator losses of the Wasserstein GAN with gradient penalty [23].

### 3.4. Applying PQ generator for image attack

Given a trained scale-free pattern generator PQG, the adversarial attack process is fairly straight-forward. As shown on Fig. 2, the PQG takes a hidden vector $Z$ and a condition vector $C$ as input to generate the adversarial pattern $P^I$. Then, the pattern is synthesized to the benign image $I$ (e.g., pixel-wised addition), resulting adversarial example $I^{adv}$. Then, we put $I^{adv}$ to the victim networks to calculate the adversarial loss to update $Z$. We conduct the standard optimization-based attack process introduced by C&W [5]. Instead of updating the perturbation, we update the hidden vector $Z$ being used in pattern generation. In attacking different victim networks, the adversarial loss are simply the negation of the loss $\mathcal{L}_V$ being used in model training. Mathematically, the attack process is formulated as follow.

$$I^{adv} = \text{Syn}(I, P^I)$$ \hspace{1cm} (Forward)  \hspace{1cm} (9)

$$P^I = \text{PQG}(Z, C, \phi_G)$$

$$Z^{t+1} \leftarrow Z^t + \alpha \nabla_Z \mathcal{L}_{adv}(I^{adv}, y^t, \phi_V)$$ \hspace{1cm} (Backward)  \hspace{1cm} (10)

Where $\phi_G$ and $\phi_V$ are the parameters of PQG and victim network, respectively. $y^t$ represents the ground truth label of the benign image $I$. $t$ is the time-stamp and $\alpha$ denotes the learning rate.

### 4. Experiment

In this paper, we presnet the experimental evaluations across fundamental vision tasks, including images classification, object detection, and semantics segmentation, to demonstrate the effectiveness of our attack. To demonstrate the generalization capabilities, we evaluate our algorithm in four common scenarios corresponding to four different patterns, rain streaks, rain drops, snow flakes, and camera lens dirt. The target patterns are shown on Fig. 3.
with kernel standard deviation $\sigma = 5$. The synthesis strategy of rain drops, snow flakes, and lens dirt patterns are pixel-wise addition, formulated as $I_{adv} = I + \gamma \times P$, where $\gamma = 1$ by default.

We conduct two different patch sizes, $32 \times 32$ and $64 \times 64$. We use patch size of $64 \times 64$ for training the rain streaks and snow flakes patterns and $32 \times 32$ for training the rain drops and lens dirt patterns. The dimension of the latent vector $Z \in \mathbb{N}(0, 1)^{128}$ is 128; an additional condition vector $C \in [0, 1]^4$ of dimension 4 is used for rain streak pattern generation to control the streak direction, heaviness, width, and length (caused camera shutter speed). In each iteration of attack optimization, we used an Adam [31] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and learning rate $lr = 0.01$. We also used cosine annealing scheduler [39].

4.3. Image Classification

4.3.1 White-box and black-box attack.

We first evaluate our attack method on ImageNet validation set, where 5000 validation images are selected, and pre-trained ResNet-18 is employed as our attack network for white-box and surrogate network for black-box attack. We additionally use six networks, e.g., ResNet-50, VGG-19, GoogleNet, InceptionV3, DenseNet-121, and MobileNetV3, as the target attack networks to evaluate the black box attack performance of our approach. We com-
Figure 5. On the ImageNet Dataset, the attack strength of our methods reaches other attack methods when no defense algorithm is applied (A). When the-state-of-the-art defense algorithms is applied (B-F), the attack strength of our method is preserved while that of other methods drops dramatically.

Figure 6. On the ImageNet Dataset, the attack strength is preserved after applying corresponding artifacts removal algorithms.

Figure 7. Evaluation of the defense robustness of our attack method on object detection and instance segmentation tasks.

4.3.2 Robustness against defense strategies.

Our main goal is to achieve robust attack against various defense strategies. To evaluate the robustness, we applied three smoothing-based defense algorithms, Jpeg Compression [10], High Frequency Suppression [65], HGD [35], and two GAN-based defense algorithms Ape-gan [53], Defense-gan [49] to the adversarial examples generated by our method. Shown on Fig. 5(B-F), smoothing-based defense algorithms (B-F) are very effective in defensing traditional noise-based attack algorithms (C&W and FGSM) but less effective in attacking semantic attack algorithms (RAAVA and AdvHaze) and nearly have no effect to our algorithms. GAN-based defense algorithms are more powerful in defensing our attack method but the performance of our method still excess other attack methods undoubtedly.

As attacking using semantic patterns, it is reasonable to suspect that the attack strength can be greatly diminished by removing the patterns. In fact, many great works have been proposed to remove the those semantic patterns (rain streak, rain drops, snow flakes, and camera lens dirt) effectively. To show that our method is robust to those pattern removal algorithms, we pre-process the adversarial examples generated by rain streaks, rain drops, snow flakes and camera lens dirt patterns using the rain streak removal (DID) [64], rain drop removal (CCN) [47], snow removal (HDCW) [7], and lens dust removal [55] algorithms, respectively. As shown on Fig. 6, attack strength is preserved even if the patterns have been visually removed.

4.4. Object Detection and Instance Segmentation

We also extend our attack algorithm to objection detection and instance segmentation tasks to further demonstrate its effectiveness. We attack the object detection model Faster-RCNN [48] and instance segmentation model Mask-RCNN [25]. Models are trained using Cityscapes [8] with standard hyperparameter setup.

Similar to the evaluation on image classification task, we pre-process the adversarial examples with five defense algorithms (Jpeg [10], HCS [65], HGD [35], Ape-gan [53] and Defense-gan [49]) and pattern removal algorithms corresponding to each of the four patterns. As shown on Fig. 7, none of them can effectively defend our attack. The visualization results are shown on Fig. 4 (B). We can see that the object detection model fails to detection of object although the objects are visually obvious.

4.5. Image quality assessment

Besides to visualization results shown on Fig. 4, we would like to numerically assess the image quality of our adversarial examples. BRISQUE [42], NIQE [43] and PIQE [57] provides great reference-free image quality assessment metric. We compare the score of our adversarial examples with those generated by other attack methods. Table 1 shows the methods beat the others.

4.6. Influence of the Latent Vector’s Dimensions

The adversarial attack strength is usually highly affected by the degree of freedom. For examples, in traditional
Table 1. Various reference-free image quality assessment algorithms show that the qualities of our adversarial examples are much higher than those of other attack methods.

| Method            | BRISQUE↓ | NIQE↓ | PIQE↓ |
|-------------------|----------|-------|-------|
| Clean             | 32.6506  | 3.0446| **43.2164** |
| FGSM              | 35.3227  | 3.7291| 51.0106 |
| C&W               | 33.4794  | 2.9942| 46.2374 |
| IadvHaze          | 49.1386  | 5.9577| 67.3326 |
| RA-AVA            | 41.4794  | 4.6255| 53.2247 |
| Rain Streaks (Ours) | 34.1253  | 3.3301| 46.4298 |
| Rain Drops (Ours)  | 32.6792  | 2.7264| 45.4913 |
| Snow Flakes (Ours) | **32.1531** | **2.5649** | **46.4413** |
| Lens Dirt (Ours)   | 32.4270  | 3.0174| 43.4377 |

Figure 8. Relation between the attack strength and dimension $N_Z$ of the latent code $Z$. Experiments includes both the white box attack (left) and the black box attach (right) on different classification models.

noise-based adversarial attack algorithms, tighter pixel-wise $l_p$ constraint usually leads to weaker attack strength. An extreme case is that one-pixel attack algorithms [54] has much lower attack success rate (ASR) than global attack algorithms. Instead of modifying the image pixel-wisely, our method modify the latent vector the of the PQ-Generator. We want to see how the dimension $N_Z$ of the latent code $Z$ affects the attack strength. We can see from Fig. 8 that the classification accuracy decrease as the dimension $N_Z$ increases. It is especially influential when for white box attack and when $N_Z$ is small. The classification accuracy tend to be steady as $N_Z$ goes above 128. By this experiments, we decide to use $N_Z = 128$ for all the other experiments.

5. Conclusion

Adversarial attack based on semantic manipulation was highly limited by the target image scale, i.e., the model architecture is usually designed to attack images of certain scale or different resolutions of specific proportion [6, 29, 30, 50]. In this work, we propose a novel adversarial attack algorithm that can attack images with different scales by generating smooth and continuous photo-realistic adversarial patterns.

We experimentally show its powerful attack strength and its robustness to different potential defense strategies. More importantly, it has been shown both visually and numerically that our method can generate adversarial examples of great image quality by synthesizing different patterns. In fact, the image quality of the adversarial examples can be further improved by more carefully chosen patterns and synthesis strategies, e.g., semantic-aware synthesis. We leave this to the future researchers.

References

[1] Cecilia Aguerrebere, Yann Gousseau, and Guillaume Tar-tavel. Exemplar-based texture synthesis: the efros-leung algorithm. *Image Processing On Line*, 2013:223–241, 2013. 3, 6
[2] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021. 3
[3] Tao Bai, Jun Zhao, Jinlin Zhui, Shoudong Han, Jiefeng Chen, Bo Li, and Alex Kot. Ai-gan: Attack-inspired generation of adversarial examples. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2543–2547. IEEE, 2021.
[4] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and David A Forsyth. Unrestricted adversarial examples via semantic manipulation. *arXiv preprint arXiv:1904.06347*, 2019.
[5] 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. 1, 2, 5, 7
[6] Weimin Chen, Yuqing Ma, Xianglong Liu, and Yi Yuan. Hierarchical generative adversarial networks for single image super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 355–364, 2021. 3, 8
[7] Wei-Ting Chen, Hao-Yu Fang, Cheng-Lin Hsieh, Cheng-Che Tsai, I Chen, Jian-Jiun Ding, Sy-Yen Kuo, et al. All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4196–4205, 2021. 6, 7
[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6, 7
[9] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02990*, 2017.
[10] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E Kounavis, and Duen Horng Chau. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In *Proceedings of the 24th ACM SIGKDD International Conference on*
Knowledge Discovery & Data Mining, pages 196–204, 2018. 1, 2, 3, 7
[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009. 6
[12] Xiaoyi Dong, Dongdong Chen, Jianmin Bao, Chuan Qin, Lu Yuan, Weiming Zhang, Nenghai Yu, and Dong Chen. Greedyfool: Distortion-aware sparse adversarial attack. arXiv preprint arXiv:2010.13773, 2020. 2
[13] Xiaoyi Dong, Jiangfan Han, Dongdong Chen, Jiayang Liu, Huanyu Bian, Zehua Ma, Hongsheng Li, Xiaogang Wang, Weiming Zhang, and Nenghai Yu. Robust superpixel-guided attentional adversarial attack. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12895–12904, 2020. 1, 2
[14] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 9185–9193, 2018. 1, 2
[15] Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A Kai Qin, and Yun Yang. Adversarial camouflage: Hiding physical-world attacks with natural styles. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1000–1008, 2020. 1, 2
[16] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pages 341–346, 2001. 3
[17] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. arXiv preprint arXiv:1703.00410, 2017. 1
[18] Joachim Folz, Sebastian Palacio, Joern Hees, and Andreas Dengel. Adversarial defense based on structure-to-signal autoencoders. In 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 3568–3577. IEEE, 2020. 3
[19] Ruijun Gao, Qing Guo, Felix Juefei-Xu, Hongkai Yu, and Wei Feng. Adhaze: Adversarial haze attack. arXiv preprint arXiv:2104.13673, 2021. 2, 7
[20] Kshitiz Garg and Shree K Nayar. Vision and rain. International Journal of Computer Vision, 75(1):3–27, 2007. 6
[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjl Oizar, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014. 3
[22] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014. 1, 2, 7
[23] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. Advances in neural information processing systems, 30, 2017. 5
[24] Shirsendu Sukanta Halder, Jean-François Lalonde, and Raoul de Charette. Physics-based rendering for improving robustness to rain. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10203–10212, 2019. 2
[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017. 7
[26] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 1614–1619, 2018. 2
[27] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8022–8031, 2019. 6
[28] Zhanhao Hu, Siyuan Huang, Xiaopeng Zhu, Fuchun Sun, Bo Zhang, and Xiaolin Hu. Adversarial texture for fooling person detectors in the physical world. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13307–13316, 2022. 2
[29] Animesh Karnewar and Oliver Wang. Msg-gan: Multi-scale gradients for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7799–7808, 2020. 3, 8
[30] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017. 3, 8
[31] Diederek P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 6
[32] Stepan Komkov and Aleksandr Petushkov. Adhat: Real-world adversarial attack on arcface face id system. In 2020 25th International Conference on Pattern Recognition (ICPR), pages 819–826. IEEE, 2021. 1, 2, 3
[33] Mark Lee and Zico Kolter. On physical adversarial patches for object detection. arXiv preprint arXiv:1906.11897, 2019. 2
[34] Hongshuo Liang, Erlu He, Yangyang Zhao, Zhe Jia, and Hao Li. Adversarial attack and defense: A survey. Electronics, 11(8):1283, 2022. 3
[35] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1778–1787, 2018. 2, 3, 7
[36] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Wang, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 10208–10216, 2019. 2
[37] Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks with momentum. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 1614–1619, 2018. 2
[39] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, 2016. 6

[40] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017. 1

[41] Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, pages 135–147, 2017. 2, 3

[42] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing, 21(12):4695–4708, 2012. 7

[43] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. IEEE Signal Processing Letters, 20(3):209–212, 2012. 7

[44] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2574–2582, 2016. 1

[45] Dantong Niu, Ruohao Guo, and Yisen Wang. Moré attack (ma): A new potential risk of screen photos. Advances in Neural Information Processing Systems, 34:26117–26129, 2021. 2

[46] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In 2016 IEEE symposium on security and privacy (SP), pages 582–597. IEEE, 2016. 3

[47] Ruijie Quan, Xin Yu, Yuanzhi Liang, and Yi Yang. Removing raindrops and rain streaks in one go. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9147–9156, 2019. 2

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 2015. 7

[49] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. arXiv preprint arXiv:1805.06605, 2018. 3, 7

[50] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4570–4580, 2019. 3, 8

[51] Uri Shaham, James Garrittano, Yutaro Yamada, Ethan Weinberger, Alex Cloninger, Xiuyuan Cheng, Kelly Stanton, and Yuval Kluger. Defending against adversarial images using basis functions transformations. arXiv preprint arXiv:1803.10840, 2018. 1

[52] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1151–1160, 2020. 2

[53] Shiwei Shen, Guoqing Jin, Ke Gao, and Yongdong Zhang. Ape-gan: Adversarial perturbation elimination with gan. arXiv preprint arXiv:1707.05474, 2017. 2, 3, 7

[54] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 23(5):828–841, 2019. 1, 2, 8

[55] Tdsepsilon. Automatic sensor dust removal, Aug 2021. 7

[56] Binyu Tian, Felix Juefei-Xu, Qing Guo, Xiaofei Xie, Xiaohong Li, and Yang Liu. AvA: Adversarial vignetting attack against visual recognition. arXiv preprint arXiv:2105.05558, 2021. 2, 7

[57] N Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Chamnapayya, and Swaprup S Medasani. Blind image quality evaluation using perception based features. In 2015 Twenty First National Conference on Communications (NCC), pages 1–6. IEEE, 2015. 7

[58] Zhibo Wang, Mengkai Song, Siyan Zheng, Zhifei Zhang, Yang Song, and Qian Wang. Invisible adversarial attack against deep neural networks: An adaptive penalization approach. IEEE Transactions on Dependable and Secure Computing, 18(3):1474–1488, 2019. 1

[59] Rey Wiyatno and Anqi Xu. Maximal jacobian-based saliency map attack. arXiv preprint arXiv:1808.07945, 2018. 1

[60] Rey Reza Wiyatno and Anqi Xu. Physical adversarial textures that fool visual object tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4822–4831, 2019. 2

[61] Qiuling Xu, Guanhong Tao, Siyuan Cheng, and Xiangyu Zhang. Towards feature space adversarial attack by style perturbation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 10523–10531, 2021. 2

[62] Chenglin Yang, Adam Kortylewski, Cihang Xie, Yinzhao Cao, and Alan Yuille. Patchattack: A black-box texture-based attack with reinforcement learning. In European Conference on Computer Vision, pages 681–698. Springer, 2020. 2

[63] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zhenglong Kong, Shouhong Ding, Jilin Li, and Cong Liu. Advmakeup: A new imperceptible and transferable attack on face recognition. arXiv preprint arXiv:2105.03162, 2021. 1, 2, 3

[64] He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 695–704, 2018. 7

[65] Zhendong Zhang, Cheolkon Jung, and Xiaolong Liang. Adversarial defense by suppressing high-frequency components. arXiv preprint arXiv:1908.06566, 2019. 1, 2, 3, 7