Meta-Attack: Class-agnostic and Model-agnostic Physical Adversarial Attack

Weiwei Feng[1], Baoyuan Wu[2,3,†], Tianzhu Zhang[1,†], Yong Zhang[4], Yongdong Zhang[1]
[1] University of Science and Technology of China
[2]School of Data Science, The Chinese University of Hong Kong, Shenzhen, China
[3]Shenzhen Research Institute of Big Data, Shenzhen, China [4] Tencent AI Lab
fengww@mail.ustc.edu.cn wubaoyuan@cuhk.edu.cn {tzzhang, zhyd73}@ustc.edu.cn
zhangyong201303@gmail.com

Abstract

Modern deep neural networks are often vulnerable to adversarial examples. Most exist attack methods focus on crafting adversarial examples in the digital domain, while only limited works study physical adversarial attack. However, it is more challenging to generate effective adversarial examples in the physical world due to many uncontrollable physical dynamics. Most current physical attack methods aim to generate robust physical adversarial examples by simulating all possible physical dynamics. When attacking new images or new DNN models, they require expensive manually efforts for simulating physical dynamics and considerable time for iteratively optimizing for each image. To tackle these issues, we propose a class-agnostic and model-agnostic physical adversarial attack model (Meta-Attack), which is able to not only generate robust physical adversarial examples by simulating color and shape distortions, but also generalize to attacking novel images and novel DNN models by accessing a few digital and physical images. To the best of our knowledge, this is the first work to formulate the physical attack as a few-shot learning problem. Here, the training task is redefined as the composition of a support set, a query set, and a target DNN model. Under the few-shot setting, we design a novel class-agnostic and model-agnostic meta-learning algorithm to enhance the generalization ability of our method. Extensive experimental results on two benchmark datasets with four challenging experimental settings verify the superior robustness and generalization of our method by comparing to state-of-the-art physical attack methods.

1. Introduction

Deep neural networks (DNNs) have been widely used in various fields and shown exceptionally good performance. However, adversarial examples (adding small-magnitude perturbations to the original input image) have been a severe threat against DNN models, and have been extensively studied in recent years [12, 3, 26, 8, 33, 5, 13]. Most existing attack works focus on the scenario of *digital attack*, which is based on an assumption that the adversarial images are directly fed into the attacked model. Only limited works [21, 2, 22, 47, 30] focus on adversarial attack in the physical world, where the above assumption is unrealistic. The physical attack is more in line with the real word and may cause security problems in practical scenarios, *e.g.*, autonomous driving or face recognition. For example, a carefully crafted physical adversarial image can mislead the autonomous driving system to behave in abnormal and potentially dangerous ways.

However, a series of recent works [2, 21] prove that physical attack is more difficult than digital attack because of many uncontrollable physical dynamics (*e.g.*, varying distances and view-angles, and characteristics of printing device). Specifically, as shown in Figure 1(a), physical attack involves with multiple stages: *(1)* Given an attacked image and a specific target DNN model, attackers craft an adversarial example in the digital space. *(2)* Attackers print the digital adversarial example out into an object (*e.g.*, 2D photo or 3D object). *(3)* The printed adversarial object is captured by a camera or scanner. Then the captured image (*i.e.*, the physical adversarial image) is fed to the DNN model. The sequential operation of printing and capturing is called as the *digital-to-physical* (D2P) transformation [21]. Note that the D2P transformation usually causes significant color and shape distortions, due to the characteristics of printing device and the relative location between printed objects and capturing devices. These distortions make physical adversarial images inconsistent with digital adversarial images. Therefore, adversarial images may become ineffective in the physical world due to the D2P transformation.

To handle this problem, recent works [46, 2, 21] propose to generate physical adversarial examples that are robust to the distortions caused by the D2P transformation. As shown in Figure 1(b), the basic idea is to simulate the possible distortions during generating adversarial examples. For example, given an attacked image and a target DNN model, attackers firstly attempt to collect large-scale physi-

† indicates corresponding authors. This work corresponds to wubaoyuan@cuhk.edu.cn and tzzhang@ustc.edu.cn

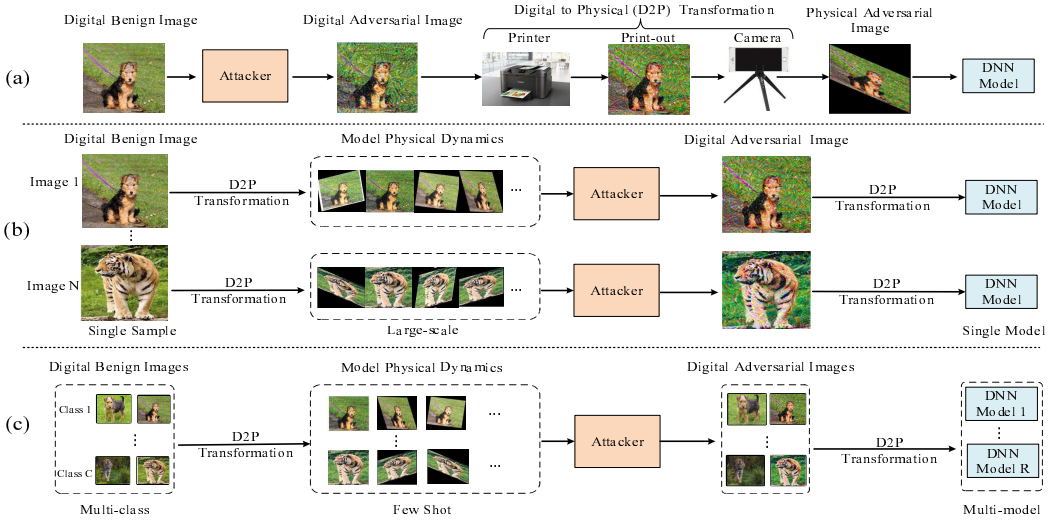


Figure 1. (a) The pipeline of physical attack. (b) The overview of existing physical attacks [11], which need to model physical dynamics manually by repeatedly performing D2P transformation. (c) The motivation of class-agnostic and model-agnostic physical attack.

cal images for simulating all possible distortions caused by physical dynamics. Then based on these collected images, attackers craft a robust physical adversarial image by iteratively optimizing for attacking the given DNN model. However, the input attacked images are diverse in the real world. When given N attacked images, existing methods [11, 2] have to spend expensive manual efforts to repeat the above attack operation for each image, which consumes considerable time and storage. Furthermore, most previous methods craft adversarial examples based on a premise that a specific target DNN model is given. However, there are numerous images and various DNN models in the physical world. The existing methods cannot handle with this complex situation, due to the expensive manually efforts for simulating physical dynamics and substantial time for optimizing.

To deal with the above issues, an intuitive idea is to formulate the physical attack as a problem of few-shot learning, with the aim of saving expensive manual efforts. The few-shot learning assumes that only a few images are available for each class. Our goal is to generate robust physical adversarial examples by accessing only few physical images. To learn a generalized attack method on these few shot physical images, we take advantage of meta-learning algorithms [14, 34] that need only a few updating steps on few shot images to achieve good performance on unseen images or even unseen DNN models (see Figure 1(c)).

To this end, we propose a class-agnostic and model-agnostic physical adversarial attack method, denoted as Meta-Attack, which is able to not only generate robust physical adversarial examples, but also generalize to attacking novel images and novel DNN models. To achieve this goal, we design a generative attack model and a class-agnostic and model-agnostic meta-learning (CMML) algorithm. (1) The generative attack model aims at generating robust physical adversarial examples by simulating color and shape distortions resulted from the D2P transformation. Concretely, we utilize a cycle-consistent adversarial network (CycleGAN) to simulate the patterns of color distortions caused by the printers. Besides, we embed the expectation over spatial transformations (EOT) into the adversarial loss to capture shape distortions. Then, the CycleGAN and EOT losses are combined together to form the generative attack model. (2) The class-agnostic and model-agnostic meta-learing algorithm (CMML) is designed to enhance the generalization ability of the proposed generative attack model across different attacked images and DNN models by using few collected physical images. To be specific, distinguished from the few-shot learning setting in the popular meta-learning algorithm Model-Agnostic Meta-Learning (MAML) [14], we design an attack task composed of a support set, a query set and a target attacked model. During the meta-training, the generative attack model is iteratively updated to acquire a good model initialization, by minimizing the validation loss on the query set of multiple attack tasks. Then during the meta-testing stage, the meta generative attack model (Meta-Attack) can rapidly adapt to novel attack tasks by fine-tuning the parameters for few steps on a support set. As a result, the CMML algorithm contributes to a class-agnostic and model-agnostic generative attack model with good performance on new images or new DNN models, by use of only a handful of digital and physical images for simulating physical dynamics.

In summary, our main contributions are three-fold: (1) We propose a class-agnostic and model-agnostic physical adversarial attack method (Meta-Attack), which is able to not only produce robust physical adversarial examples by simulating color and shape distortions, but also adapt to attacking new images from unseen classes and new DNN models by accessing a few digital and physical images. (2) We formulate the physical attack as a few-shot learning problem, where the training task is redefined as the composition of a support set, a query set, and a target DNN model. Then, we design a class-agnostic and model-agnostic meta-learning algorithm to enhance the generalization ability of

the generative attack model under the few-shot setting. (3) Extensive experimental results on two datasets with four challenging experimental settings verify the superior robustness and generalization of our method by comparing to state-of-the-art physical attack methods.

2. Related Work

In this section, we briefly review related works on the digital and physical adversarial attack.

Digital Attack. Since the pioneering work [41] revealed that deep neural networks are vulnerable to adversarial examples, various attack methods have been proposed [33, 7, 1, 4, 19, 27, 32, 35], which can be divided into three categories: gradient-based [16, 24], optimization-based [3, 41], and GAN-based methods [18, 29, 42]. Gradient-based methods perform gradient update to generate adversarial examples, such as FGSM [16] and BIM [25]. Optimization-based methods generate adversarial examples by solving an optimization problem, such as C&W [3]. Different from these methods, GAN-based methods [45] can directly transform input images into adversarial examples using feed-forward networks. There are already many works [9, 17, 20, 37, 50] studying adversarial examples in the digital space, while only limited works [2, 22, 46] study adversarial examples in the physical world.

Physical Attack. In [25], it's the first work that shows the adversarial examples generated by the digital attack method are also possible to attack the model in the physical scenario. Recently, many researchers start to study generating adversarial examples in the physical world [28, 43, 44, 23, 38, 31]. The main difficulties of generating adversarial examples in the physical world are the distortions from both the printing process and the spatial transformation. For the first issue, Sharif et al. [38] observe that the distortion between the digital image and physical image is partially caused by the fact that the color space of the printer is only a subset of the whole RGB space, as a result, the pixel values out of the color space are clipped when printing. Based on this observation, the non-printability score (NPS) is introduced in [38] for improving the printability of adversarial examples. To address the second issue, the method called expectation over transformation (EOT) is proposed in [2] to improve the robustness to the spatial transformation. It redefines the adversarial loss in the digital attack to an expectation over spatial transformations with respect to the original adversarial loss. And the EOT loss is further extended in some recent works [11, 46, 36, 44, 48, 10, 22]. For example, the RP2 [11] improves EOT by adding physical images to the transformation sets, together with the NPS score. Different from these works, another recent work called D2P [21] firstly trains a generative adversarial network (GAN) [15] to transform the original digital image into one image that is similar to its physical image. Then, the EOT method is adopted on the generated image to produce an adversarial image. Although both RP2 and D2P consider the distortions from both the printing and the spatial transformation, these two distortions are captured in separate stages. As a result, the generated adversarial examples may not be robust to both distortions. Moreover, most existing physical attack approaches [11, 31, 23, 21] are only tested on limited cases or only designed for attacking a specific DNN model or a specific image. Different from the above methods, we consider both the color and shape distortions in one unified model, such that the robustness to both distortions could be simultaneously simulated. Besides, we first formulate the physical attack as a problem of few-shot learning. And we design a class-agnostic and model-agnostic meta-learning algorithm to solve this problem, which can also improve the generalization ability of physical attack by accessing a few digital and physical images.

3. Our Approach

In this section, we present the overall scheme of Class-agnostic and Model-agnostic Physical Adversarial Attack (Meta-Attack) method, as shown in Figure 2.

3.1. Notations and Preliminaries

We denote the benign image as $\boldsymbol{x} \in \mathcal{X}$, whose physical image is denoted as $\boldsymbol{x}_p \in \mathcal{X}$. The attacked model is denoted as $f : \mathcal{X} \to \mathcal{Y}$, with $\mathcal{Y}$ being the output space. Our goal is to learn a generative network $G_{\boldsymbol{\theta}} : \boldsymbol{x} \to \boldsymbol{x}_{adv}$, which could produce an adversarial image $\boldsymbol{x}_{adv} = G_{\boldsymbol{\theta}}(\boldsymbol{x})$ to fool the model f in the targeted mode[1]. We formulate the problem of targeted attack as $f(\boldsymbol{x}_{adv}) = y_t \neq f(\boldsymbol{x})$, where y_t is the given target label. We denote the physical image of $\boldsymbol{x}_{adv}$ as $\boldsymbol{x}^p_{adv}$, so a robust physical adversarial example can be formulated as $f(\boldsymbol{x}^p_{adv}) = y_t \neq f(\boldsymbol{x})$.

However, as shown in Figure 1(a), there are always distortions between the digital image and its corresponding physical image. To generate robust physical adversarial images, we need to analyze the reasons behind such distortions in details. We factorize the D2P transformation into two transformations. **(1)** 1:1 Digital-to-Physical transformation (1:1-D2P transformation) means that the digital image is printed and scanned with 1:1 scale. It does not change the shape of the original digital image, but the pixel values may be changed due to the characters of the printer and scanner/camera. **(2)** Spatial transformations occur during capturing the adversarial image due to the varying relative locations and view-angles between the printed photo and the capturing device. They may change the shape, scale, and location of the captured adversarial image.

[1]Like other physical attacks [2, 21], in this work we focus on the targeted attack. Because changing the original prediction at random in physical attack is easy, which may be due to the D2P transformation, rather than the adversarial perturbation.

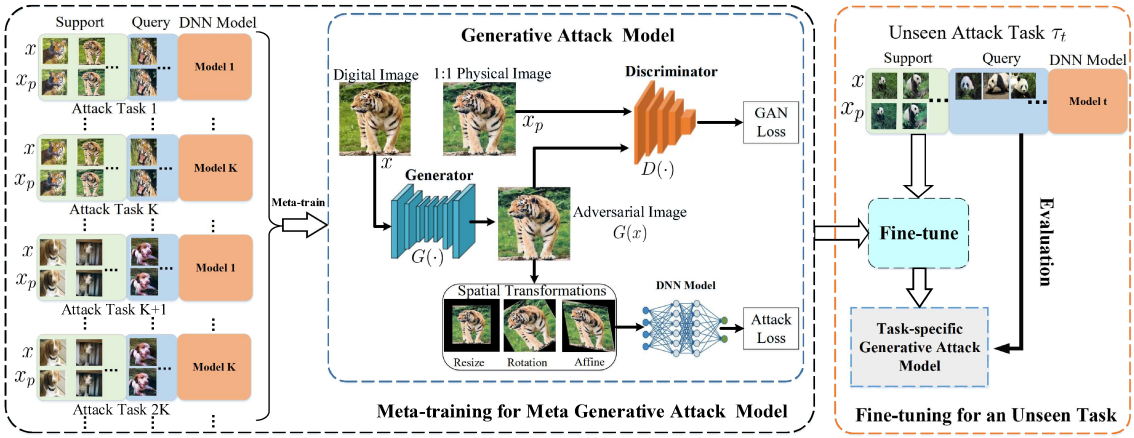


Figure 2. The detailed scheme of class-agnostic and model-agnostic physical adversarial attack method. During the meta-training, training tasks are fed into the generative attack model to learn a meta generative attack model (*i.e.*, G^{meta}, D^{meta}). In the meta-test phase, the meta generative attack model is fine-tuned on the support set of an unseen attack task τ_t, and then evaluated on a query set.

3.2. Meta-Attack

Architecture Overview. The detailed scheme of our method is shown in Figure 2, which is composed of a robust generative attack model and a **C**lass-agnostic and **M**odel-agnostic **M**eta-**L**earning algorithm (CMML). The generative attack model is to generate robust physical adversarial images by simultaneously simulating both color and spatial distortions. While the class-agnostic and model-agnostic meta-learning algorithm can improve the generalization ability of the generative attack model by use of only a handful physical images for simulating physical dynamics. The generative attack model trained by CMML can achieve the goal of class-agnostic and model-agnostic physical adversarial attack.

Generative Attack Model. As shown in Figure 2, the generative attack model $G(\cdot)$ takes the benign image $\boldsymbol{x}$ as input and generates an adversarial example $\boldsymbol{x}_{adv}$ (*i.e.*, $G(\boldsymbol{x})$) to fool the target DNN model f. Our goal is to enhance the robustness of the adversarial example $\boldsymbol{x}_{adv}$, which can remain effectiveness in the physical world. As discussed in Section 3.1, a successful and robust physical adversarial example $\boldsymbol{x}_{adv}$ should satisfy three requirements including ***(1)*** *Successful digital attack*, ***(2)*** *Robustness to 1:1-D2P transformation*, and ***(3)*** *Robustness to spatial transformations*.

(1) Successful Digital Attack. To satisfy the first requirement, we should solve the following optimization problem,

$$\min_{G} \mathcal{L}_{adv}\big(f(G(\boldsymbol{x})), y_t\big), \tag{1}$$

where $\mathcal{L}_{adv}$ is the cross entropy loss for the targeted attack, which corresponds to the attack loss in Figure 2.

(2) Robustness to 1:1-D2P Transformation. To ensure the robustness of $\boldsymbol{x}_{adv}$ (*i.e.*, $G(\boldsymbol{x})$) to color distortions caused by the 1:1-D2P transformation, following [21], we encourage $\boldsymbol{x}_{adv}$ to be close to the corresponding physical image $\boldsymbol{x}_p$. From this perspective, $G(\cdot)$ can be seen as the domain transfer from the digital domain $\mathcal{X}_d$ to the physical domain $\mathcal{X}_p$, which can be formulated as:

$$\min_{G}\max_{D} \mathcal{L}_{\text{x-GAN}}(\boldsymbol{x}, \boldsymbol{x}_p; G, D), \tag{2}$$

where D indicates the discriminator, $\boldsymbol{x} \in \mathcal{X}_d$, $\boldsymbol{x}_p \in \mathcal{X}_p$ [2]. We denote the GAN loss as $\mathcal{L}_{\text{x-GAN}}$ generally, as any GAN model can be adopted, which is specified in Section 4.1.

(3) Robustness to Spatial Transformations. Inspired by the expectation over transformations (EOT) method [2], we introduce several spatial transformations to model varying distances and view-angles during generating adversarial images. Then we aim to improve the robustness of adversarial examples based on a series of synthetically transformed images. Thus, we have the following optimization problem:

$$\min_{G} \mathbb{E}_{t\in T}\big[\mathcal{L}_{adv}\big(f(t(G(\boldsymbol{x}))), y_t\big)\big], \tag{3}$$

where T denotes a chosen transformation distribution of the transformation function t. In practice, the distribution T can model shape distortions such as random rotation, translation, resize, or affine.

Full Objective Function. In order to achieve a robust physical adversarial attack, we combine Eq. (1), Eq. (2) and Eq. (3) to derive the final objective function:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{x-GAN}}(\boldsymbol{x}, \boldsymbol{x}_p; G, D) \\ &+ \lambda \cdot \mathbb{E}_{t\in T}\big[\mathcal{L}_{adv}\big(f(t(G(\boldsymbol{x}))), y_t\big)\big] \\ &+ c \cdot ||G(\boldsymbol{x}) - \boldsymbol{x}_p||_p,\end{aligned} \tag{4}$$

where $\lambda > 0$ denotes the trade-off parameter to balance the robustness to 1:1-D2P transformation and the attack performance under the spatial transformation, which will be specified in experiments. And the last term of Eq. (4) is to guarantee that the adversarial examples are imperceptible by human eyes. Therefore, we aim to solve the following optimization problem:

$$G^*, D^* = \arg\min_{G}\max_{D} \mathcal{L}. \tag{5}$$

[2] In the following pages, except for special statements, physical images represent the images printed and scanned with 1:1 scale.

Class-agnostic and Model-agnostic Meta-Learning. To train the proposed generative attack model, according to Eq. (2), we need to collect large-scale training image pairs (*i.e.*, $\boldsymbol{x}$ and $\boldsymbol{x}_p$). However, it is unrealistic to craft physical images for each images, which requires printing and scanning repeatedly. Therefore, we formulate it as a few-shot learning problem and just craft few-shot image pairs for each class. To better solve this few-shot learning problem, we propose a Class-agnostic and Model-agnostic Meta-Learning (CMML) algorithm. In order to make the CMML algorithm suitable for the scenario of physical adversarial attack, we design a new setting of the training task τ that is different from the seminal work [14] of meta-learning. We define the attack task τ that consists of a support set $\mathcal{D}_s$, a query set $\mathcal{D}_q$, and a target attacked model f, where both $\mathcal{D}_s$ and $\mathcal{D}_q$ are composed of digital images $\boldsymbol{x}$ and corresponding physical images $\boldsymbol{x}_p$. The CMML algorithm can help the proposed generative attack model to produce robust physical adversarial examples by using the crafted few-shot image pairs. Besides, thanks to the concept "learning to learn" of meta learning [14, 34], which aims at generalizing to new tasks and new environments that have never been encountered during training, our CMML algorithm can improve the generalization ability of the generative attack model on attacking unseen images and unseen target DNN models.

As Figure 2 shows, our CMML algorithm consists of two stages including the meta-training step for a meta generative attack model and the fine-tuning step for unknown attack tasks. At the meta-training step, our CMML algorithm takes $\{\tau_i\}_{i=1}^N = \{\mathcal{D}_s^{\tau_i}, \mathcal{D}_q^{\tau_i}, f_i\}_{i=1}^N$ as inputs and produces a meta generative attack model (*i.e.*, G^{meta}, D^{meta}) by minimizing the validation loss on query sets of multiple attack tasks. When at the fine-tuning step, given a new attack task $\tau_t = \{\mathcal{D}_s^{\tau_t}, \mathcal{D}_q^{\tau_t}, f_t\}$, the meta generative attack model should be fine-tuned for few steps on few-shot image pairs (*i.e.*, $\mathcal{D}_s^{\tau_t}$), which avoids collecting a lot of digital and physical image pairs for simulating physical dynamics. After fine-tuning, a task-specific generative attack model (*i.e.*, G^{τ_t}, D^{τ_t}) is evaluated on $\mathcal{D}_q^{\tau_t}$, which just consists of several digital images. As a consequence, the proposed CMML algorithm is able to not only effectively deal with the few-shot problem, but also improve the generalization ability of the generative attack model on unseen attack tasks.

3.3. Training and Inference

Meta-Training. The meta generative attack model (*i.e.*, G^{meta}, D^{meta}) is trained by our proposed CMML algorithm to find a sensitive and transferable initial parameters such that a few gradient updating steps on few-shot digital and physical image pairs can lead to good performance on a new attack task. The entire pipeline is illustrated in Figure 2 and Algorithm 1. Each training task τ_i is composed of a query set, a support set and a target attacked model: $\{\mathcal{D}_s^{\tau_i}, \mathcal{D}_q^{\tau_i}, f_i\}$. Let $\mathcal{L}_{\tau_i}(\boldsymbol{\theta}_G, \boldsymbol{\theta}_D, f_i)$ denote the loss of task τ_i, $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_D$ are the parameters of G and D. Following the practice in MAML [14], an update step of the parameters of task τ_i with respect to the support set can be represented by:

$$\boldsymbol{\theta}_{G,\tau_i}^{'} = \boldsymbol{\theta}_G - \alpha \nabla_{\boldsymbol{\theta}_G} \mathcal{L}_{\tau_i}(\boldsymbol{\theta}_G, \boldsymbol{\theta}_D, f_i, \mathcal{D}_s^{\tau_i}), \tag{6}$$

$$\boldsymbol{\theta}_{D,\tau_i}^{'} = \boldsymbol{\theta}_D + \alpha \nabla_{\boldsymbol{\theta}_D} \mathcal{L}_{\tau_i}(\boldsymbol{\theta}_G, \boldsymbol{\theta}_D, f_i, \mathcal{D}_s^{\tau_i}), \tag{7}$$

where α is a learning rate. The query set $\mathcal{D}_q^{\tau_i}$ is used to evaluate the effectiveness of the updated parameters, *i.e.*, $\mathcal{L}_{\tau_i}(\boldsymbol{\theta}_{G,\tau_i}^{'}, \boldsymbol{\theta}_{D,\tau_i}^{'}, f_i, \mathcal{D}_q^{\tau_i})$. Hence, the objective function of meta learning is defined as

$$\mathcal{L}^{\text{meta}} = \sum_{\tau_i \in p(\tau)} \mathcal{L}_{\tau_i}(\boldsymbol{\theta}_{G,\tau_i}^{'}, \boldsymbol{\theta}_{D,\tau_i}^{'}, f_i, \mathcal{D}_q^{\tau_i}), \tag{8}$$

where $p(\tau)$ is the distribution of the constructed tasks. As images in each task are randomly sampled, $p(\tau)$ follows a uniform distribution. The update of meta parameters is defined as:

$$\boldsymbol{\theta}_G = \boldsymbol{\theta}_G - \beta \nabla_{\boldsymbol{\theta}_G} \mathcal{L}^{\text{meta}}, \tag{9}$$

$$\boldsymbol{\theta}_D = \boldsymbol{\theta}_D + \beta \nabla_{\boldsymbol{\theta}_D} \mathcal{L}^{\text{meta}}, \tag{10}$$

where β is the learning rate.

Meta-Testing. During the meta-testing stage, for an unseen target attack task $\tau_t = \{\mathcal{D}_s^{\tau_t}, \mathcal{D}_q^{\tau_t}, f_t\}$, we iteratively fine-tune the meta generative attack model (*i.e.*, G^{meta}, D^{meta}) by Eq. (6) and Eq. (7) for few steps on $\mathcal{D}_s^{\tau_t}$, where only few-shot digital and physical image pairs are needed. Then, the images from $\mathcal{D}_q^{\tau_t}$ (the physical images are not required) can be directly fed to the fine-tuned model to get corresponding robust physical adversarial examples.

Algorithm 1 Our proposed Meta-Attack Method

Input: Training task set $\{\tau_i\}_{i=1}^N = \{\mathcal{D}_s^{\tau_i}, \mathcal{D}_q^{\tau_i}, f_i\}_{i=1}^N$, target unseen task $\tau_t = \{\mathcal{D}_s^{\tau_t}, \mathcal{D}_q^{\tau_t}, f_t\}$.
Output: fine-tuned generator G and discriminator D.

```
Initial G parameters θ_G and D parameters θ_D;
/⋆ meta-train ⋆/
for iter = 1, 2, 3... do
    Sample a mini-batch of k tasks;
    for all k tasks do
        Compute θ'_{G,τ_i}, θ'_{D,τ_i} by Eq. (6), (7) on D_s^{τ_i};
    end for
    Compute L^{meta} according to Eq. (8);
    Update θ_G, θ_D according to Eq. (9) and Eq. (10);
end for
/⋆ meta-test on an unseen task τ_t = {D_s^{τ_t}, D_q^{τ_t}, f_t} ⋆/
Fine-tune θ_G, θ_D by Eq. (6) and Eq. (7) on D_s^{τ_t}.
```

4. Experiments

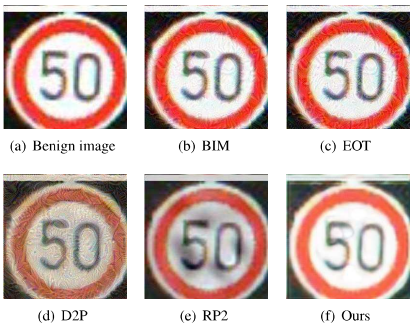
In this section, we conduct experiments with four different settings to evaluate the performance of the proposed method in both digital domain and physical domain. Please refer to the **Supplementary Material** for more results.

4.1. Detailed Settings

Datasets. We evaluate our approach on ImageNet [6] and GTSRD [40]. **Attacked Model.** If not particularly specified, we choose VGG-16 [39] as the target model by default. **Implementation.** Our generative network is implemented based on Cycle-GAN [49]. The trade-off parameters λ, c in Eq. (4) are set to 20 and 10, respectively. Task-level learning rate α in Eqs. (6), (7) and meta-level learning rate β in Eqs. (9), (10) are set to 0.0001 and 0.0002, respectively.

Baseline Methods. We compare our method with several baseline approaches including BIM [26], EOT [2], D2P [21], and RP2 [11]. Here, we set ϵ = 40/255, α = 2/255, steps = 30 for BIM. We follow [21] to configure the settings of EOT and D2P schemes, including scaling, rotation, translation, noise level $\epsilon = 40/255$ and step size $\alpha = 0.5$. Similarly, our configuration for RP2 is the same as the original paper [11], while the NPS (non-printability score) term is modified by us to fit our printing equipment.

Evaluation Metrics. For each generated adversarial image, we test it in both digital and physical domains (i.e., printing, scanning or photographing). Then, we report the attack success rate (ASR) and the average confidence of the target label overall tested images (120 images).



(a) Benign image (b) BIM (c) EOT

(d) D2P (e) RP2 (f) Ours

Figure 3. Adversarial images generated by different methods. (from "SpeedLimit 50" to "SpeedLimit 30".)

4.2. Results and Analysis

As discussed in section 3.2, we define the task τ as $\{\mathcal{D}_s, \mathcal{D}_q, f\}$. Therefore, to evaluate the effectiveness of our model, we conduct four different experiments including ***(1)*** attacking a seen model on images of a seen class, ***(2)*** attacking a seen model on images of an unseen class, ***(3)*** attacking an unseen model on images of a seen class, and ***(4)*** attacking an unseen model on images of an unseen class.

Exp.1 Attack a seen model on images of a seen class.
Setting. Given K images of one class, we perform the 1:1-D2P transformation to get their physical images. Attacking each individual image is treated as a task. In meta-train phase, for each task, a randomly selected pair from K training pairs is treated as the support set. We scale and crop the digital and physical images. Then, the pair of the cropped digital and physical patches is treated as the query set. In meta-test phase, a target image and its physical image from the same class are treated as the support set for fine-tuning. After fine-tuning for M steps, we perform attack on the target image with the fine-tuned model. We specify the parameters $K = 50$, $M = 30$.

Table 1. Comparison results of **Exp.1** on GTSRD.

Domain →	Digital		Physical	
Attack↓	ASR	Conf	ASR	Conf
BIM [26]	**1.0**	**1.0**	0.213	0.188
EOT [2]	**1.0**	**1.0**	0.846	0.546
D2P [21]	**1.0**	**1.0**	0.909	0.787
RP2 [11]	**1.0**	**1.0**	0.838	0.665
Ours	**1.0**	**1.0**	**0.952**	**0.936**

Table 2. Photographing with different view angles on GTSRD.

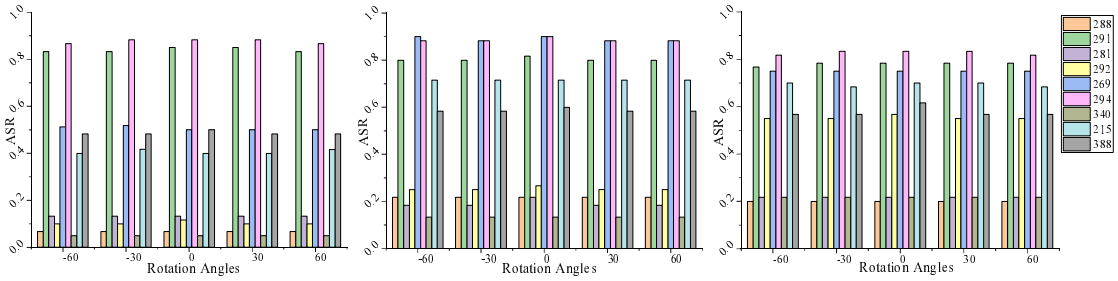
View Angles	−45°		0°		45°	
Attack↓	ASR	Conf	ASR	Conf	ASR	Conf
EOT [2]	0.415	0.377	0.633	0.401	0.403	0.373
D2P [21]	0.537	0.389	0.700	0.408	0.517	0.386
RP2 [11]	0.497	0.298	0.667	0.398	0.502	0.364
Ours	**0.667**	**0.599**	**0.817**	**0.647**	**0.657**	**0.558**

Results. We show the results of this setting on both ImageNet and GTSRD. Figure 3 presents some adversarial examples. From Table 1, our model outperforms the other methods by a large margin in terms of ASR in the physical domain, and achieves a high ASR of 95.2%. We also evaluate the robustness of adversarial examples on GTSRD by changing viewing angles. In Table 2, We observe that our method has a better performance compared to EOT, D2P and RP2. At the frontal view, the ASR of our method is 81.7%, while the ASR of EOT, D2P, RP2 is 63.3%, 70.0%, 66.7%, respectively. When the image is turned by -45 degrees, the ASR of ours is still 66.7%, which is higher than others. The results on ImageNet are shown in Table 3. Although the ASR of all methods in the digital domain can reach 100%, our method outperforms other methods in the physical domain under all spatial transformations, which means our adversarial examples are more robust.

Exp.2 Attack a seen model on images of an unseen class.
Setting. Given images from C classes, we use one class for evaluation in meta-test phase and the other $C-1$ classes for meta-train. Firstly, we perform the 1:1-D2P transformation to get the physical images of given images. In meta-train phase, to construct a task, two pairs of digital and physical image are randomly selected from one of the $C-1$ classes. One pair is treated as the support set while the other is treated as the query set. In meta-test phase, several pairs of the target class are treated as the support set for fine-tuning. After fine-tuning for M steps, we perform evaluations on the other images of the target class with the adjusted model. We configure the parameter $C = 9$, which represents the number of the images classes sampled from ImageNet. (label list [288,291,281,292,269,294,340,215,388]).
Results. Table 4 and Figure 4 report the generalization and

Table 3. Results on **Exp.1** under different spatial transformations in the physical domain on ImageNet database.

Attack →	BIM		EOT		RP2		D2P		Ours	
Spatial Transformation ↓	ASR	Conf	ASR	Conf	ASR	Conf	ASR	Conf	ASR	Conf
Digital domain	**1.0**	0.932	**1.0**	0.908	0.980	0.806	**1.0**	0.925	**1.0**	**0.994**
Resize 1 + Rotation 0°	0.067	0.023	0.667	0.594	0.680	0.444	0.733	0.694	**0.850**	**0.834**
Resize 1 + Rotation 20°	0.0	0.0	0.617	0.574	0.676	0.425	0.717	0.686	**0.837**	**0.829**
Resize 1 + Rotation −20°	0.0	0.0	0.667	0.596	0.676	0.436	0.717	0.686	**0.842**	**0.833**
Resize 1.2 + Rotation 0°	0.0	0.0	0.767	**0.679**	0.675	0.435	0.667	0.619	**0.817**	0.673
Resize 1.2 + Rotation 20°	0.0	0.0	0.757	**0.668**	0.667	0.433	0.650	0.608	**0.807**	0.665
Resize 1.2 + Rotation −20°	0.0	0.0	0.753	0.645	0.675	0.428	0.667	0.619	**0.817**	**0.673**
Resize 0.8 + Rotation 0°	0.0	0.0	0.633	0.583	0.667	0.441	0.667	0.669	**0.783**	**0.765**
Resize 0.8 + Rotation 20°	0.0	0.0	0.617	0.575	0.643	0.431	0.650	0.653	**0.783**	**0.765**
Resize 0.8 + Rotation −20°	0.0	0.0	0.633	0.587	0.643	0.428	0.667	0.669	**0.767**	**0.757**



(a) Resize 0.8 (b) Resize 1.0 (c) Resize 1.2

Figure 4. The adversarial images from different classes against different spatial transformations in the physical domain under **Exp.2**.

Table 4. Experimental results of **Exp.2** on ImageNet dataset.

Domain →	Digital		Physical	
Source Label ↓	ASR	Conf	ASR	Conf
288	0.750	0.688	0.217	0.146
291	0.900	0.900	0.817	0.682
281	0.283	0.281	0.217	0.113
292	0.817	0.789	0.267	0.209
269	0.983	0.926	0.900	0.690
294	0.967	0.972	0.900	0.771
340	0.717	0.683	0.133	0.117
215	0.933	0.887	0.716	0.470
388	0.900	0.877	0.600	0.513
ave	0.806	0.778	0.530	0.406

robustness of adversarial examples on ImageNet database. The results on GTSRD will appear in the **Supplementary Material**. The class label shown in the first column of Table 4 is the unseen class. From Table 4, it is easy to find that the model can quickly adapt to the unseen class and get the ASR with high values. In particular, when we treat 269 as the unseen class, our ASR reaches 98.3% in the digital domain and 90.0% in the physical domain. To verify the robustness of adversarial examples, we also evaluate adversarial examples from various unseen classes under different spatial transformations in the physical domain. Figure 4 shows that when our adversarial examples undergo spatial transformations, the fluctuation of ASR is very small. This reveals that the spatial transformation has little effect on adversarial examples, which is attributed to EOT [2].

Exp.3 Attack an unseen model on images of a seen class.

Setting. Given R target models, one is used for evaluation in the meta-test phase while the other $R-1$ models are used for meta-training. Given K images from one class, we perform the 1:1-D2P transformation to get their physical images, forming K pairs of digital and physical images. In meta-train phase, to construct a task, we randomly select a pair from the K pairs as the support set and another pair as the query set while we randomly select a model from $R-1$ models as the target model. In meta-test phase, we randomly select a few target images and their physical images from the same class as the support set to fine-tune the generative model to attack the testing target model. After fine-tuning for M steps, we perform attack on the digital images of the same class with the adjusted model. The hyper-parameters are specified as follows: $R=3, K=50, M=30$. These 3 target models include VGG-16, VGG-19 and ResNet-50.

Table 5. Result of adapting to different attacked DNN models under different spatial transformations on ImageNet dataset.

Attacked Model →	VGG-16		VGG-19		ResNet-50	
Spatial Transformation ↓	ASR	Conf	ASR	Conf	ASR	Conf
Digital domain	0.850	0.732	0.867	0.811	0.683	0.615
Resize 1 + Rotation 0°	0.383	0.365	0.583	0.312	0.308	0.288
Resize 1 + Rotation 20°	0.367	0.353	0.567	0.309	0.292	0.276
Resize 1 + Rotation −20°	0.367	0.352	0.567	0.310	0.308	0.281
Resize 1.2 + Rotation 0°	0.483	0.330	0.683	0.353	0.317	0.294
Resize 1.2 + Rotation 20°	0.483	0.318	0.667	0.349	0.292	0.267
Resize 1.2 + Rotation −20°	0.483	0.317	0.667	0.349	0.292	0.277
Resize 0.8 + Rotation 0°	0.367	0.316	0.567	0.343	0.250	0.233
Resize 0.8 + Rotation 20°	0.367	0.309	0.558	0.334	0.241	0.227
Resize 0.8 + Rotation −20°	0.350	0.304	0.558	0.342	0.241	0.236

Results. Exp.3 is to examine the generalization ability upon the attacked models. Table 5 presents the results on ImageNet database. The **Supplementary Material** will provide the results on GTSRD database. When adapting to VGG-16, in Table 5, the ASR in the digital domain is 85.0%. If we replace the attacked model with VGG-19, the ASR is also with a high value of 86.7% in the digital domain. However the physical attack performance degrades to some extent, but the performance is still acceptable, especially on VGG-19. For example, when adapting to VGG-19, and rotating the image by 20 degrees, scaling it by 0.8, the ASR is still 55.8%. This indicates that our model has a good generalizable characteristics and the generated adversarial images are robust against spatial transformations.

Table 6. Result of adapting to different attacked classes and different DNN models on ImageNet dataset.

Source Label → / Attacked Model ↓			215	291	388
VGG-16	Digital	ASR	0.850	0.833	0.867
		Conf	0.776	0.807	0.834
	Physical	ASR	0.537	0.417	0.467
		Conf	0.487	0.381	0.381
VGG-19	Digital	ASR	0.875	0.867	0.883
		Conf	0.831	0.838	0.841
	Physical	ASR	0.592	0.583	0.492
		Conf	0.547	0.557	0.451
ResNet-50	Digital	ASR	0.683	0.637	0.717
		Conf	0.637	0.586	0.659
	Physical	ASR	0.307	0.313	0.283
		Conf	0.276	0.278	0.247

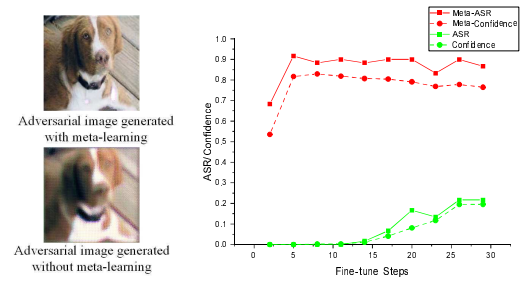
Exp.4 Attack an unseen model on images of an unseen class.

Setting. Given images from C classes and R target models, we use one class and one model for evaluation in meta-test phase and the other $C-1$ classes and $R-1$ models for meta-train. Similarly, we perform the 1:1-D2P transformation to get the physical images of given images. In meta-train phase, to construct a task, two pairs of digital and physical image are randomly selected from one of the $C-1$ classes. One pair is treated as the support set while the other is treated as the query set, and a randomly selected model from $R-1$ models as the target model. In meta-test phase, several pairs of the target class and the target model are used for fine-tuning. After fine-tuning for M steps, we attack the target model on the other images of the target class with the adjusted model. The hyper-parameters are specified as follows: $R=3, M=30, C=3$. (label list [215, 291, 388]). Target models include VGG-16, VGG-19 and ResNet-50.

Results. Exp.4 is to evaluate the performance of attacking an unseen model on images from unseen class. Table 6 reports the results on ImageNet. We can observe that in most cases, the performance of our method is acceptable. For example, when we treat VGG-19 and 291 as the unseen DNN model and unseen class, respectively, the ASR in the digital domain is 86.7%, and the physical ASR achieves 58.3%. When we focus on ResNet-50 and label 215, the digital ASR reaches 68.3%, and the ASR in the physical domain is 30.7%. We find that the ASR of VGG is overall higher than that of ResNet, which indicates that ResNet is more robust than VGG. Consistently across all cases, our method performs well on different classes and different DNN models, which proves the good generalization of our method.

4.3. Ablation Studies

In this section, we explore the influence of the proposed class-agnostic and model-agnostic meta-learning (CMML) algorithm. To show the performance difference, we design a comparative experiment. We use two generative attack models to generate adversarial images of the same images for attacking a given DNN model, one of which is randomly initialized, while the other is pre-trained by CMML. Figure 5 shows the performance comparison of the these two generative attack models. From the comparison of the two images in Figure 5(a), we can easily find that the upper is clearer, which is generated by the pre-trained model with the proposed meta-learning strategy. In addition, Figure 5(b) presents that the pre-trained model significantly outperforms the model without pre-training on ASR and confidence metrics in digital domain. The pre-trained model can achieve a much higher ASR of 90%, while the model without pre-training can hardly attack successfully.



(a) Generated image comparision (b) ASR and confidence comparison

Figure 5. Performance comparison of the generative attack model with or without CMML pre-training in the digital domain.

5. Conclusion

In this work, we present a class-agnostic and model-agnostic physical adversarial attack method, which is able to not only generate robust physical adversarial examples, but also show good generalization. Firstly, we propose a generative attack model by combining the CycleGAN and EOT loss together to simulate color and spatial distortions. Then we formulate the physical attack as a problem of few-shot learning, and propose a class-agnostic and model-agnostic meta-learning algorithm, which can enhance the generalization ability of the generative attack model on attacking novel images or novel DNN models. Comprehensive experimental results on two datasets with four experimental settings demonstrate the superiority of the proposed attack method with good robustness and generalization.

6. Acknowledgment

This work was partially supported by the National Key Research and Development Program under Grant No. 2018YFB0804204, National Defense Basic Scientific Research Program (JCKY2020903B002), National Nature Science Foundation of China (Grant 62022078, 62021001), and Youth Innovation Promotion Association CAS 2018166. Baoyuan Wu is supported by the National Natural Science Foundation of China under grant No.62076213, the university development fund of the Chinese University of Hong Kong, Shenzhen under grant No.01001810, the special project fund of Shenzhen Research Institute of Big Data under grant No.T00120210003, and Shenzhen Science and Technology Program under grant No.GXWD20201231105722002-20200901175001001.

References

[1] N. Akhtar and A. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 3

[2] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018. 1, 2, 3, 4, 6, 7

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2017. 1, 3

[4] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey, 2018. 3

[5] Weilun Chen, Zhaoxiang Zhang, Xiaolin Hu, and Baoyuan Wu. Boosting decision-based black-box adversarial attacks with random sign flip. In *European Conference on Computer Vision*, pages 276–293. Springer, 2020. 1

[6] 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

[7] 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. 3

[8] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019. 1

[9] Jiawei Du, Hu Zhang, Joey Tianyi Zhou, Yi Yang, and Jiashi Feng. Query-efficient meta attack to deep neural networks. In *International Conference on Learning Representations*, 2020. 3

[10] 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. 3

[11] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018. 2, 3, 6

[12] Yanbo Fan, Baoyuan Wu, Tuanhui Li, Yong Zhang, Mingyang Li, Zhifeng Li, and Yujiu Yang. Sparse adversarial attack via perturbation factorization. In *Proceedings of European Conference on Computer Vision*, 2020. 1

[13] Yan Feng, Baoyuan Wu, Yanbo Fan, Li Liu, Zhifeng Li, and Shutao Xia. Boosting black-box attack with partially transferred conditional adversarial distribution. *arXiv preprint arXiv:2006.08538*, 2020. 1

[14] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017. 2, 5

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 3

[17] Ying Guo, Xingxing Wei, Guoqiu Wang, and Bo Zhang. Meaningful adversarial stickers for face recognition in physical world. *arXiv preprint arXiv:2104.06728*, 2021. 3

[18] Jiangfan Han, Xiaoyi Dong, Ruimao Zhang, Dongdong Chen, Weiming Zhang, Nenghai Yu, Ping Luo, and Xiaogang Wang. Once a man: Towards multi-target attack via learning multi-target adversarial network once. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5158–5167, 2019. 3

[19] Han Xu Yao Ma Hao-Chen, Liu Debayan Deb, Hui Liu Ji-Liang Tang Anil, and K Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178, 2020. 3

[20] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 3

[21] Steve TK Jan, Joseph Messou, Yen-Chen Lin, Jia-Bin Huang, and Gang Wang. Connecting the digital and physical world: Improving the robustness of adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 962–969, 2019. 1, 3, 4, 6

[22] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. *arXiv preprint arXiv:1908.08705*, 2019. 1, 3

[23] Zelun Kong, Junfeng Guo, Ang Li, and Cong Liu. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14254–14263, 2020. 3

[24] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 3

[25] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 3

[26] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 1, 6

[27] Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. In *Advances in neural information processing systems*, pages 10408–10418, 2019. 3

[28] Aishan Liu, Jiakai Wang, Xianglong Liu, Bowen Cao, Chongzhi Zhang, and Hang Yu. Bias-based universal adversarial patch attack for automatic check-out, 2020. 3

[29] Xuanqing Liu and Cho-Jui Hsieh. Rob-gan: Generator, discriminator, and adversarial attacker. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11234–11243, 2019. 3

[30] Bo Luo and Qiang Xu. Region-wise attack: On efficient generation of robust physical adversarial examples. *arXiv preprint arXiv:1912.02598*, 2019. 1

[31] Jinqi Luo, Tao Bai, Jun Zhao, and Bo Li. Generating adversarial yet inconspicuous patches with a single image. 2020. 3

[32] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 3

[33] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks, 2016. 1, 3

[34] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018. 2, 5

[35] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4422–4431, 2018. 3

[36] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International Conference on Machine Learning*, pages 5231–5240. PMLR, 2019. 3

[37] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditional image editing. *arXiv preprint arXiv:1906.07927*, 2019. 3

[38] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016. 3

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 6

[40] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, (0):–, 2012. 6

[41] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *Computer Science*, 2013. 3

[42] Xiaosen Wang, Kun He, and John E Hopcroft. At-gan: A generative attack model for adversarial transferring on generative adversarial nets. *arXiv preprint arXiv:1904.07793*, 3(4), 2019. 3

[43] Zhibo Wang, Siyan Zheng, Mengkai Song, Qian Wang, Alireza Rahimpour, and Hairong Qi. advpattern: Physical-world attacks on deep person re-identification via adversarially transformable patterns. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8341–8350, 2019. 3

[44] Zuxuan Wu, Ser-Nam Lim, Larry Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. *arXiv preprint arXiv:1910.14667*, 2019. 3

[45] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018. 3

[46] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *European Conference on Computer Vision*, pages 665–681. Springer, 2020. 1, 3

[47] Yan Xu, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, and Wei Liu. Exact adversarial attack to image captioning via structured output learning with latent variables. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4135–4144, 2019. 1

[48] Hiromu Yakura and Jun Sakuma. Robust audio adversarial example for a physical attack. *arXiv preprint arXiv:1810.11793*, 2018. 3

[49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 6

[50] Daniel Zügner and Stephan Günnemann. Adversarial attacks on graph neural networks via meta learning. *arXiv preprint arXiv:1902.08412*, 2019. 3