GENERATING REALISTIC PHYSICAL ADVERSARIAL EXAMPLES BY PATCH TRANSFORMER NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Physical adversarial attacks apply carefully crafted adversarial perturbations onto real objects to maliciously alter the prediction of object classifiers or detectors. The current standard method for designing physical adversarial patches, i.e. Expectation over Transformations (EoT), simulates real-world environments by random physical transformations, resulting in adversarial examples far from satisfactory. To tackle this issue, we propose and develop a novel network to learn real-world physical transformations from data, including geometric transformation, printer color transformation and illumination adaption. Our approach produces realistic-looking adversarial examples and can be integrated into existing attack generation frameworks to generate adversarial patches effectively. We apply our approach to design adversarial T-shirts worn by moving people, one of the most challenging settings for physical attacks. Experiments show that our approach significantly outperforms the state of the arts when attacking DL-based object detectors in real life. Moreover, we build a first-kind-of adversarial T-shirts dataset to enable effective training of our approach and facilitate fair comparison on physical world attacks by considering a standard patch size, environment changes and object variances. Our code will be made publicly available.

1 INTRODUCTION

Deep learning (DL) has achieved tremendous success in a wide range of applications such as image classification, object detection, and language processing (Alom et al., 2018; Zhao et al., 2019; Minaee et al., 2020; Young et al., 2018). However, a large body of work has shown that deep neural networks (DNNs) are vulnerable to *adversarial attacks or examples* (Goodfellow et al., 2015; Carlini & Wagner, 2017; Papernot et al., 2016; Kurakin et al., 2016), in the sense that inputs with carefully-crafted perturbations can cause erroneous outputs in a DNN model. Such vulnerability has been demonstrated widely in the digital world by the existence of adversarial attacks (Goodfellow et al., 2015; Carlini & Wagner, 2017; Xu et al., 2019; Ilyas et al., 2018; Su et al., 2017; Papernot et al., 2017; Chen et al., 2017; Brown et al., 2017). In recent years, researchers have also shown that adversarial examples can be achieved in the physical world, through the creation of carefully designed physical objects (Thys et al., 2019; Eykholt et al., 2018a; Li et al., 2019; Eykholt et al., 2018b; Xu et al., 2020; Chen et al., 2018; Sharif et al., 2016; Wu et al., 2020; Athalye et al., 2018; Sitawarin et al., 2018). These physical attacks pose serious security and safety concerns to DNN-based computer vision systems such as autonomous driving and video surveillance.

While powerful physical adversarial examples exist, their design is currently less effective than that of digital attacks, even in the setting where perturbation strength is unconstrained. In contrast to digital adversarial attacks , physical adversarial attacks (PAAs) must deal with varying environment conditions in the real world under which physical adversaries are viewed. A standard method so-called Expectation over Transformations (EoT) (Athalye et al., 2018) designs PAAs by learning a *universal* adversarial perturbation (a.k.a adversarial patch (Brown et al., 2017)) across *all* possible physical transformations including background noise, translation, rotation, lighting, and contrast. The EoT has demonstrated success in a number of scenarios from 3D printed objects (Athalye et al., 2018), to stop/traffic signs (Eykholt et al., 2018b; Evtimov et al., 2017), eyeglass frames (Sharif et al., 2016), adversarial cardboard (Thys et al., 2019) and the more recent adversarial T-shirts (Xu et al., 2020) and invisibility cloaks (Wu et al., 2020). However this approach has some inherent

limitations. First, the randomness of the approach implies a large search space of transformations for finding an optimized adversarial patch, making the training inefficient or even infeasible. Second, as illustrated in Fig. 1, the adversarial examples generated by the EoT for training look substantially different from their physical realizations, indicating that the EoT lacks capability to simulate realistic environments, such as geometric consistency and lighting changes. Due to that, *large discrepancies between training data (digital world) and test data (physical world) exist, leading to a significant performance gap of adversarial patches between the digital world and physical world.*

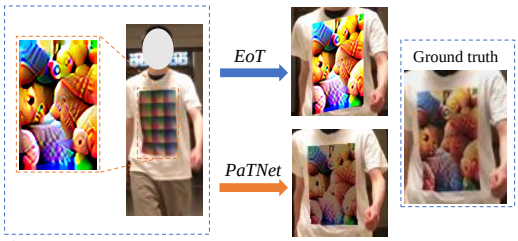


Figure 1: Generating adversarial examples in this work involves inserting an adversarial patch (left) into a specified region of a T-shirt. Shown here are two examples produced by Expectation of Transformation(EoT) (Athalye et al., 2018) and our approach *PaTNet*, respectively. Clearly, *PaTNet* produces a more realistic patch than the EoT, based on the ground truth (right).

In this paper, we ask if it is possible to learn physical transformations automatically from data and use them to generate environment-resilient adversarial examples for designing physical adversaries. To this end, we aim to develop a patch simulator which can infer the physical-world conditions from an image and adapt an adversarial patch to such conditions in a realistic way (see Fig. 1). Our problem is closely related the image inpainting task (Pathak et al., 2016) that learns to repair a "lost" part of an image, but differs in that the content (i.e. the adversarial patch) is provided in our case. Since an adversarial patch is dynamically updated during the attack generation, the key challenge to our problem is to ensure spatial invariance and illumination consistency in simulation for realism, even for novel unseen patches. For this reason, we propose to disentangle the simulation task into a sequence of physical transformations including geometric transformation, printer color transformation and illumination adaption, and jointly learn them by a novel patch transformer network, herein referred to as *PaTNet*. As demonstrated later, *PaTNet* produces realistic adversarial examples and enables design of robust adversaries that produce strong attacks in the physical world.

We apply our approach to create adversarial T-shirts (or "invisibility cloaks") (Xu et al., 2020; Wu et al., 2020), arguably the most challenging setting among existing adversarial attacks. Here, the attack generator designs adversarial patches on a non-rigid object, i.e. a T-shirt, to fool object detectors and allow moving people to evade detection. Learning all transformations aforementioned in an end-to-end way is non-trivial. While public datasets such as CoCo (Lin et al., 2014) and Inria (Dalal & Triggs, 2005) have been shown useful for attack generation (Thys et al., 2019; Wu et al., 2020), they don't provide patch location information needed for learning physical transformations effectively, as indicated in (Xu et al., 2020). We thus collected a new dataset dedicated for designing adversarial T-shirts and facilitating fair evaluation of adversarial attacks. The dataset contains an uniform size of the adversarial pattern with varied objects and scenes, and is available at https://github.com/Anonymous1125/patnet_dataset. We further develop a two-stage training algorithm to learn *PaTNet* robustly with only minimal annotations required (i.e. patch location). We summarize our contributions as below.

- We develop a novel framework *PaTNet* to generate realistic-looking examples for training physical adversarial attack models. To the best of our knowledge, it is one of the very few approaches that aim to reduce the performance gap of adversarial patches between the digital world and physical world.
- Our approach does not require strong supervision, converges fast in training, and generalizes well to novel adversarial patches (disentangled transformations). All this enables design of strong physical adversarial attacks against realistic environments from large-scale datasets.
- We apply the proposed framework to design adversarial T-shirts to attack multiple popular detectors and evaluate the efficacy of our approach under challenging real-world scenarios. Our adversarial patches outperform the SOTA approaches by an absolute 30%.
- We created a first-kind-of real world dataset to facilitate fair comparison on adversarial T-shirts generated from different approaches.

2 Related Work

Physical Adversarial Attacks There exist many different types of adversarial attacks against image classification or object detectors in the real world. In (Sharif et al., 2016), the first physical attacks against face recognition systems by designing adversarial eyeglass frames was demonstrated. Another well-known success of physical adversarial attacks is the generation of adversarial stop signs that can fool state-of-the-art object detectors (Eykholt et al., 2018b; Evtimov et al., 2017; Chen et al., 2018). In (Athalye et al., 2018), a framework known as Expectation over Transformation (EoT) was proposed to synthesize adversarial examples robust to a set of physical transformations. EoT has become a standard way to craft physical adversarial attacks such as adversarial vehicle camouflages (Zhang et al., 2019) and adversarial cloths (Thys et al., 2019; Wu et al., 2020; Xu et al., 2020).

Most existing work has focused on the design of adversarial patches (or stickers) on rigid physical objects, e.g., eyeglass frames (Sharif et al., 2016), traffic signs (Eykholt et al., 2018b; Evtimov et al., 2017; Chen et al., 2018; Duan et al., 2020), cameras (Li et al., 2019), vehicles (Zhang et al., 2019) and pieces of cardboard (Thys et al., 2019), except for (Wu et al., 2020; Xu et al., 2020), which address non-rigid objects. The most relevant work to ours is (Thys et al., 2019; Wu et al., 2020; Xu et al., 2020). In (Thys et al., 2019), the problem of physical adversarial attacks against person detectors was studied for the first time. Although the proposed attack only applied to a rigid object—cardboard, held by a stationary person, it successfully fools person detectors. Spurred by (Thys et al., 2019), the work (Wu et al., 2020; Xu et al., 2020; Huang et al., 2020) attempted to design wearable adversarial examples. In (Wu et al., 2020), a comprehensive empirical study was made to demonstrate the effectiveness of an adversarial cloak to fool various object detectors. In (Xu et al., 2020), by utilizing Thin Plate Spline (TPS) based transformation to model fabric distortions in an EoT framework, an adversarial T-shirt was created that could fool object detectors even for moving individuals. In (Huang et al., 2020), a simulated 3D world environment was built to support generating an Universal Physical Camouflage Attack. The method of (Jan et al., 2019) applies the conditional GAN technique to simulate the physical transformations for generating robust adversarial examples, with a focus on digital adversarial examples.

Image Generation The pioneering work of Generative Adversarial Network (GAN) (Goodfellow et al., 2014) has led to a series of interesting approaches in image generation which produce realistic content such as cycle-GAN (Zhu et al., 2017), star-GAN(Choi et al., 2018) and style-GAN (Karras et al., 2019). These techniques have also been widely applied to other topics related to image manipulation such as image inpainting/completion (Pathak et al., 2016; Yu et al., 2018; Portenier et al., 2019; 2018) and view synthesis (Park et al., 2017; Lin et al., 2018). Our problem of generating adversarial examples is closely related to image inpainting in the sense of completing a missing region in an image. (Pathak et al., 2016) train a GAN in combination with a pixel-wise reconstruction loss to reconstruct the missing part of an image. In (Yu et al., 2018), the idea of context-aware attention is proposed to encourage spatial consistency for image completion. In our case, since the content to be filled in is provided by a digital patch already, we emphasize ensuring a generated example to be spatially and color consistent with the source image.

3 Proposed Approach

The adversarial example generation process is similar to image inpainting, which aims to reconstruct lost or deteriorated parts of an image. In our case, the "missing" content is provided by the adversarial patch,thus our primary interest is to ensure spatial invariance and illumination consistency in the modification for realism. In other words, the generated adversarial examples should look realistic under varying environment conditions in order for the attack to achieve robustness against real-world scenarios. Such a task is challenging as the physical realization of a digital patch undergoes a number of underlying physical transformations involving a printer, a camera, a scene as well as the patch itself. There is no known method to model such a complex process precisely in a principled way.

In this work, we propose a general framework for physical adversarial attacks with a particular focus on making adversarial examples resilient to real-world environments. It is an important problem for physical adversarial attacks (PAAs) but has not received much attention. We demonstrate the effectiveness of our idea by the example of adversarial T-shirts. However, it should be straightforward to generalize our approach to other types of physical adversarial attacks. Before detailing our approach, we first briefly describe how Expectation over Transformation (EoT) (Athalye et al., 2018),

the existing most successful framework for PAAs, is used to create adversarial patches for attacking object detectors.

Preliminaries: Adversarial Patch by EoT. Let $\boldsymbol{\delta}$ be an adversarial patch (or perturbation) masked by M_i. An image $\mathbf{x}_i$ from a training set D results in an adversarial example $\mathbf{x}'_i = (1 - M_i) \cdot \mathbf{x}_i + M_i \cdot t(\boldsymbol{\delta})$ by a set of predefined transformations $t \in \mathcal{T}$. The adversarial patch $\boldsymbol{\delta}$, in the context of object detection, can be generated by optimizing the following objective (Eykholt et al., 2018a),

$$\arg\min_{\boldsymbol{\delta}} \lambda ||\boldsymbol{\delta}||_{\text{TV}} + \mathcal{L}_c(\boldsymbol{\delta}) + \mathbb{E}_{t\in\mathcal{T},\mathbf{x}_i\in\mathcal{D}} J(f_\theta(\mathbf{x}'_i, \mathbf{y}') \tag{1}$$

The first term of (1) is the total variation (TV) of $\boldsymbol{\delta}$ in the ℓ_p norm to ensure smoothness of the patch. The second term is a loss function that measures the printing quality of the patch, such as the non-printability score (NPS) proposed in (Sharif et al., 2016). The last term J indicates the attack loss, which measures the difference between the detection score of a victim detector f_θ and a target label $\mathbf{y}'$.

The currently existing approaches of PAAs solve Eq. 1 by the EoT (Athalye et al., 2018), which applies random image transformations such as scaling, rotation and lighting from $\mathcal{T}$ to $\boldsymbol{\delta}$. Clearly, the randomness of transformations, even if $\mathcal{T}$ is small, still results in a large search space for optimizing $\boldsymbol{\delta}$ and slow convergence in training. In addition, the randomly sampled transformations are far from being representative of the real physical world, thus unable to provide training examples with good quality for adversarial path generation. As a result, adversarial examples based on the EoT often perform well in the digital world, but cannot live up to expectations in practice. Regardless of these limitations, the EoT has demonstrated success in a variety of interesting applications (Biggio et al., 2018).

Overview of *PaTNet*. To address the limitations of the EoT, we propose to learn physical transformations in a data-driven way and then apply them to generate adversarial examples. As mentioned earlier, our problem is closely related to image inpainting (Yu et al., 2018; 2019b), which can be potentially used for generating adversarial examples. However, image inpainting approaches heavily rely on local and global contexts to fill in the holes of an image with semantically consistent content. They do not tend to generalize well to new data, as shown later in Figure 3. In addition, training these approaches is computationally costly and requires a large amount of data.

In this work, we construct adversarial examples from an adversarial patch by considering three fundamental physical transformations, namely *geometric transformation* ($\mathcal{T}_{\text{gt}}$), *printer color transformation* ($\mathcal{T}_{\text{pct}}$), and *illumination adaptation* ($\mathcal{T}_{\text{ia}}$). Among them, $\mathcal{T}_{\text{gt}}$ controls how the patch is geometrically mapped to a target image; $\mathcal{T}_{\text{pct}}$ attempts to make the patch colors more reproducible by the printer; and finally $\mathcal{T}_{\text{ia}}$ adapts the patch to fit the illumination of the scene. As illustrated in Figure 2, each of these transformations is implemented as a separate network module, which is then integrated into one network to jointly learn all the transformations with minimal supervision from the person and patch locations only. We refer to this network as Patch Transformer Network (*PaTNet*) for convenience.

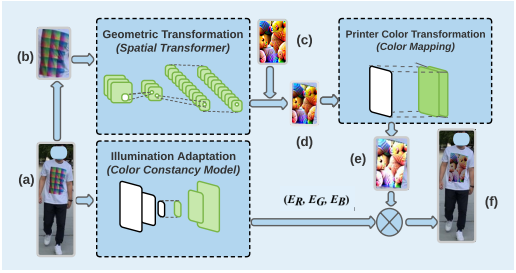


Figure 2: An overview of our approach. The network takes as the input a person image (a), a cropped patch (b) from (a) and a binary mask of the patch in (b) (not drawn here), transforming a digital patch (c) to replace (b) on the T-shirt. The patch (c) goes through three fundamental physical transformations, i.e. geometric transformation (d), printer color transformation (e) and illumination adaption to ensure realism on (f). (E_R, E_G, E_B) is an illuminant inferred from a color constancy model to adjust the brightness of (e).

Note that the aforementioned transformations can by no means precisely simulate the complex process underlying the physical-world conditions. For example, most security cameras have automatic white balance, which is not considered here. However, by combining together all the transformations described above, we demonstrate that they can be jointly optimized and yield visually more satisfied adversarial examples, which in turn improve the effectiveness of adversarial attacks. We also would like to point out that disentangling these transformations for simulation rather than learning them in a blackbox setting like a GAN-based image generator (Yu et al., 2018; 2019b), is essential for ensuring *PaTNet* to have good generalizability to novel adversarial patterns (see Fig. 3 for comparison).

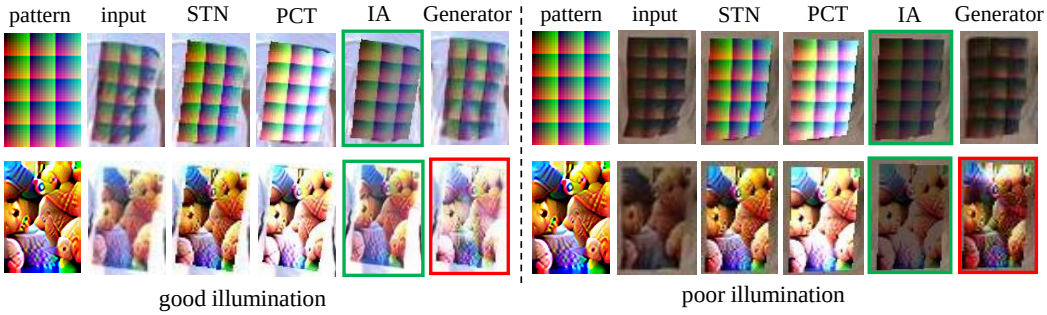


Figure 3: Simulated results of each transformation (STN, PCT and IA) under two different lighting conditions. The last column shows examples generated by an image generator (Yu et al., 2018) (see Figure 4). The resulting output of our approach (i.e. IA) demonstrates good adaptation capability to different lighting conditions for both patches while the generator (red box) is unable to generalize to the novel adversarial patch.

As shown in Figure 2, the input to *PaTNet* includes an adversarial patch, a person image and a binary mask M_i [1] indicating where to insert the patch in the image. The output is a transformed image represented by

$$\mathbf{x}_i' = (1 - M_i) \cdot \mathbf{x}_i + M_i \cdot \mathcal{T}_{\text{ia}}(\mathcal{T}_{\text{pct}}(\mathcal{T}_{\text{gt}}(\boldsymbol{\delta}))) \tag{2}$$

Geometric Transformation. The task of geometric transformation is to determine how to place a patch onto a target object (i.e. T-shirt in our case) through finding a geometric mapping between them. Depending on application, different geometric relationships can be applied. Accurate patch localization is crucial for designing robust physical adversarial attacks. For example, affine transformation is used for attaching adversarial stickers to stop signs (Eykholt et al., 2018b) while Thin-Plate-Spline (TPS) based mapping is shown more effective in modeling fabric variations like cloth deformation caused by a moving person (Xu et al., 2020). To the best of our knowledge, all existing approaches rely on either manual labeling or random sampling to obtain spatial transformations to create physical adversarial examples.

In this work, we apply Spatial Transformer Networks (STN) (Jaderberg et al., 2015) to automatically learn the geometric transformation between a patch and an object. As shown in Figure 2, the input to STN is a cropped region (b) from an image, which is used to infer the spatial transformation between itself and a digital pattern (c). The learned geometric model is then applied to transform the original patch onto a T-shirt. Learning STN requires supervision, which usually comes from a different task. In our case, we train STN by minimizing the difference between the reconstructed patch and the cropped region (b), which is detailed at the end of this section. We refer readers to the Appendix for quantitative localization results of STN.

Printer Color Transformation. Instead of recognizing red, green and blue (RGB) values, high-end digital printers usually present colors using a color wheel of cyan, magneta and yellow (CMYK). Because of these variances, some RGB colors are outside of the CMYK gamut and cannot be made with CMYK. This results in differences (sometimes significant) between on-screen colors and what result in post-printing. The printer color issue has been well known to cause performance degradation in physical adversarial attacks (Eykholt et al., 2018b; Xu et al., 2020). Prior works such as (Sharif et al., 2016) use the non-printability score (NPS) to guide the perturbation to choose RGB values closer to printable colors. More recently, a simpler but effective method is developed in (Xu et al., 2020). It directly calibrates a color mapping between a digital color palette representing 960 colors (Figure 3) and an image of the patch taken under a natural lighting condition.

Inspired by the idea in (Xu et al., 2020), we insert a network module into our framework to learn a mapping between digital and printable colors. In other words, we look for a function $\mathcal{T}_{\text{pct}}$ that can translate a digital color to its printable one. We adopt a 3×3 matrix $\mathcal{M}$ to transform a digital patch P_i by $\hat{P}_i^{\text{pct}} = \mathcal{T}_{\text{pct}}(P_i) = \mathcal{M}P_i$. We also experimented with a quadratic polynomial function, which yields similar results.

[1] A mask is created by using the four corner points of the patch on the T-shirt, which can be either manually annotated or automatically identified by image matching or detected by a detector.

Illumination Adaption. Our approach so far can paste a digital patch onto an object (i.e. T-shirt) with good localization. However the PCT discussed above only partially addresses the color inconsistency between the patch and its physically realized content. As illustrated in Figure 3, the output of PCT is still not satisfactory due to the lack of adaptation to the environmental illumination. Further addressing this issue is related to a long-standing problem in vision called computational color constancy (CCC) (Gijsenij et al., 2011). CCC involves removing illumination color casts in an image so that the image can be perceived the same way regardless of illumination changes. In other words, the goal of CCC is to infer the illuminant with which the image is lit, i.e. the missing information needed to recover the true color of a transformed patch by PCT. We briefly describe CCC below.

The *diagonal matrix* is a simple but widely used method in CCC, which computes the color of an object under an illuminant by a 3×3 diagonal matrix. Let $(R_{\mathrm{i}}, G_{\mathrm{i}}, B_{\mathrm{i}})$ be the RGB value of a pixel i on an image and $(E_{\mathrm{R}}, E_{\mathrm{G}}, E_{\mathrm{B}})$ is the unknown illuminant. The diagonal model is then given by

$$[R_{\mathrm{i}}, G_{\mathrm{i}}, B_{\mathrm{i}}]^T = diag(E_{\mathrm{R}}, E_{\mathrm{G}}, E_{\mathrm{B}}) \cdot [R_{\mathrm{i}}^{\mathrm{c}}, G\mathrm{i}^{\mathrm{c}}, B_{\mathrm{i}}^{\mathrm{c}}]^T \tag{3}$$

where T denotes Transpose of matrix and $(R_{\mathrm{i}}^{\mathrm{c}}, G_{\mathrm{i}}^{\mathrm{c}}, B_{\mathrm{i}}^{\mathrm{c}})$ is the color under a canonical light source, a.k.a white light. In (Barnard & Barnard, 1995), the problem of CCC is defined as estimating the canonical illuminant E. It should be clear that color constancy is the key problem to resolve if we treat the output of PCT as $(R_{\mathrm{i}}^{\mathrm{c}}, G_{\mathrm{i}}^{\mathrm{c}}, B_{\mathrm{i}}^{\mathrm{c}})$ in Eq. 3.

The solution to color constancy has witnessed significant progress in recent years by deep learning based approaches (Barron, 2015; Hu et al., 2017; Yu et al., 2019a). In our work, we adopted FC^4, a CNN-based approach (Hu et al., 2017), for illumination estimation. FC^4 is a fully convolution network that can be easily integrated into our approach for end-to-end training. Figure 3 illustrates the simulated results of two patches under good and poor lighting conditions. The color palette is included in the training data while the adversarial patch is novel to the transformers. It is clear that the final results of *PaTNet* (i.e. IA) inherit both the geometry and illuminations well from the input images, demonstrating the efficacy and generalizability of our approach. In contrast, a powerful image generator from a GAN-based approach (Yu et al., 2018) (see the next session for detail), while showing excellent simulation for the color palette, is unable to produce satisfactory results for a novel adversarial patch. Especially, the bottom two images highlighted by the red boxes indicate that the generator lacks the ability of lighting adaptation demonstrated by *PaTNet*.

Two-stage Training of *PaTNet*. We learn our proposed *PaTNet* by matching a reconstructed patch from *PaTNet* to its physical example in the real world (i.e. an input image). The learning is mostly done in an unsupervised way and the only information required is the patch location. The *PaTNet* integrates a sequence of transformation modules into one network, and is designed in a way that end-to-end learning is feasible. However, we found that learning all the transformations jointly does not converge well as expected in practice. This is largely because the color constancy model (FC^4), which is dedicated to estimate an illuminant, appears too weak to provide strong supervision on learning the geometric mapping (i.e. STN). Moreover, the success of FC^4 itself requires accurate pixel-level localization provided by STN. We thus develop a two-stage approach below to train *PaTNet* effectively.

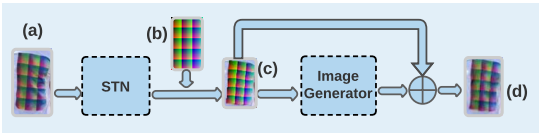


Figure 4: Learning STN using an image generator. The network transforms a patch (b) to visually match its physical realization (a) on the T-shirt (training data). The Image generator learns the difference between (c) and (d).

At the first stage, we focus on learning STN by coupling it with a powerful image generator to create adversarial examples (no printer color transformation and illumination adaptation), as shown in Fig. 4. The image generator used in our work is the fine encoder-decoder model developed in (Yu et al., 2018). Different from (Yu et al., 2018), the generator in our case learns to infer the residual between the reconstructed patch and the original one in an attempt to reduce the dependency on image context (Fig. 4). At the second stage, we train the entire *PaTNet* end-to-end with STN fixed. Both stages optimize the following objective function,

$$\underset{\{\theta_{\mathrm{pct}}, \theta_{\mathrm{ia}}\}}{\arg\min}\ E_{\mathbf{x}_i \in \mathcal{D}} \|\mathbf{x}_i' - \mathbf{x}_i\|_1 \tag{4}$$

where θ_{pct} and θ_{ia} are the model parameters of the PCT and IA modules respectively, and $\mathbf{x}_i$ is a ground-truth adversarial patch from a T-shirt while $\mathbf{x}_i'$ is the final output of *PaTNet*.

Table 1: Digitial-world ASR under different transformations.

Methods	YOLOv2		YOLOv3	
	self	*PaTNet*	*self*	*PaTNet*
PatNet-s	92.2	54.5	66.5	10.5
PatNet-sp	93.1	78.0	50.0	30.4
PatNet	86.5	86.5	42.6	42.6

Table 2: Digital-world attack transferrability of *PaTNet*

	Victim Detectors				
Patch	YOLOv2	YOLOv3	SSD	FRCNN	RetinaNet
YOLOv2	**86.5**	1.1	7.3	0.0	0.2
YOLOv3	27.8	**42.6**	0.6	0.0	0.8
SSD	11.4	0.0	**57.6**	0.0	0.0
FRCNN	9.0	0.8	4.0	**34.3**	20.7
RetinaNet	10.0	0.1	0.5	0.00	**83.2**

Adversarial Patch Generation. Integrating *PaTNet* into an adversarial patch generator is straightforward. Following (Xu et al., 2020), we define the attack loss J of Eq. 1 by

$$J(\mathbf{x}_i') = \max_{B_j \cap P_i > \eta} \max(p(B_j), \nu)\} \tag{5}$$

where $p(\cdot)$ denotes the confidence score of the j^{th} 'person' bounding box B_j and ν is a confidence threshold. The use of $\max\{p(\cdot), \nu\}$ enforces the optimizer to minimize the bounding boxes of high probability ($\geq \nu$). $|B_j \cap P_i| > \eta$ indicates that B_j has at least η-overlapping with the person P_i wearing an adversarial T-shirt. η is set to be a small number 0.1 in both training and test. The rational behind Eq. 5 is that the probability of person detection would be suppressed only if a 'person' bounding box associated with the adversarial T-shirt has a high confidence.

4 Experiments

4.1 Experimental Setup

In what follows, we illustrate the effectiveness of *PaTNet* in design of adversarial T-shirt (Xu et al., 2020) to fool person detectors. Our experiments consider 5 advanced detectors including YOLOv2 (Redmon & Farhadi, 2017), YOLOv3 (Redmon & Farhadi, 2018), SSD (Liu et al., 2016), Faster RCNN (Ren et al., 2015), and RetinaNet (Lin et al., 2017).

Dataset. We built a large-scale dataset for evaluating our proposed approach. We used a checkerboard and color palette T-shirts to collect a total of 105 videos from four indoor and two outdoor scenes by iPhone cameras. We uniformly sampled video frames at a rate of $1/5$, which results in $4,892$ frames. Among these frames, $2,830$ of them were used for training and $1,688$ for validation. Some samples frames are shown in Fig. 5. We refer readers to the Appendix for more detail of the dataset and annotations.

Figure 5: Data used for training and validating *PaTNet*.

Training Details. We conducted all our experiments on compute nodes with 6 V100 GPUs and a total of 96GB GPU memory. We chose the Adam optimizer and trained all the models using a batch size of 72 for 500 epochs. For *PaTNet*, we fix the parameters of Adam ($\beta_1 = 0.5$, $\beta_2 = 0.9$ and $lr = 1 \times 10^{-4}$) while employing adaptive learning in adversarial patch training.

Baselines. We conduct comprehensive experiments on both digital and physical worlds with three variants of our proposed *PaTNet*, 1) ***PatNet-s*** using STN only; 2) ***PatNet-sp*** including STN and PCT; 3) ***PatNet*** with all three transformations, which is our proposed approach. Furthermore, we compare with other existing methods of designing adversarial T-shirts: ***adv-patch*** (Thys et al., 2019), ***advT*** (Xu et al., 2020) and ***inv-cloak*** (Wu et al., 2020). Note that different approaches are trained on different datasets and designed for different scenarios with different sizes/numbers of patches, it's challenging to make strictly fair comparison of these approaches, especially in the real-world scenarios. Nevertheless, we managed to compare our approach with other existing methods under some reasonable assumptions, in both the digital world and physical world (see Section 4.2).

4.2 Experimental Results

Following (Xu et al., 2020), We set the minimum detection threshold to 0.7 for fair comparison, and use Attack Success Rate (ASR) to report results in percentage (%).

Digital-world Test. We first investigate the effects of physical transformations. We generate adversarial patches against YOLOv2 and YOLOv3 using different patch transformers *PatNet-s*,

PatNet-sp and *PatNet*. The evaluations are conducted in two different ways: *self* means the test transformers are the same transformers as used in training; *PaTNet* indicates the test results are based on our proposed transformations, STN + PCT + IA. Interestingly, as shown in Table 1, when training and evaluation use the same setting (*self*), *PatNet-s* and *PatNet-sp* outperform *PatNet* on both YOLOv2 and YOLOv3. This can be explained by the fact that fewer transformations imply fewer constraints, or a larger perturbation space that allows for better optimization. On the other hand, when tested by a more realistic transformer *PaTNet*, *PatNet* itself performs the best as expected.

To understand the efficacy and transferability of *PatNet* on stronger detectors, we compare the vulnerability of 5 advanced object detectors in Table 2. All detectors are trained on the COCO dataset (Lin et al., 2014). SSD (Liu et al., 2016) and Faster RCNN (FRCNN) (Ren et al., 2015) use VGG-16 (Simonyan & Zisserman, 2014) as the backbone network while RetinaNet (Lin et al., 2017) is based on ResNet-101 (He et al., 2016). The results indicate that YOLOv2 is the most vulnerable while Faster RCNN being the most robust. Table 2 also shows that adversarial T-shirts have limited transferrability, which is only observed when the backbone networks are similar, such as YOLOv2 and YOLOv3.

Physical-world Test. The efficacy of any physical adversarial attack needs to be carefully validated in the physical world. We collected new test data in 4 scenes with completely new environments different from training data for the physical-world experiments. Each video captures two actors walking in parallel. Both actors wear adversarial T-shirts, one of which is designed by our proposed approach *PatNet* and the other T-shirt generated by a baseline method. The actors then swap positions or T-shirts and repeat the acting, resulting in a total of 4 videos per scene. This allows us to make a direct pairwise comparison between our approach and a baseline method.

Table 3: Pysical-world attack results (ASR) of *PaTNet* against YOLOv2 in different unseen scenes.

Scenes	*PatNet-s*	*PatNet*	*PatNet-sp*	*PatNet*
outdoor 1	9.5	39.2	20.4	47.6
outdoor 2	52.2	56.3	29.9	52.0
indoor 1	18.4	33.8	2.9	31.2
indoor 2	36.2	52.2	14.0	50.2
overall†	27.7	**45.3**	17.9	**46.0**

† overall results are computed as the total number of frames detected over the total number of frames.

Table 3 lists the attack results of different approaches in different scenes. Our approach demonstrates stronger attack ability than *PatNet-s* (geometric transformation only) and *PatNet-sp* (geometric and printer color transformations) by a large margin of $20\% \sim 30\%$, suggesting that the realism of adversarial examples is crucial for robust physical adversarial attacks in real-world scenarios. We refer readers to Figure A3 in the Appendix for examples of our proposed approach. Demo clips can also be found in the supplemental materials.

We also conduct experiments to evaluate the performance of other detectors in the real world. The results show that the ASRs are 13.1%, 18.6%, 1.0% and 40.0% for YOLOv3, SSD, Faster RCNN and RetinaNet respectively. There seems a tendency that stronger detectors are less vulnerable to adversarial attacks, which is aligned with the finding in the digital test. Faster RCNN is almost immune to adversarial attack in our experiments, possibly because the attack strategy in Eq. 5 is not effective on the two-stage architecture of Faster RCNN. While a larger-scale evaluation is desired for better understanding of the vulnerability of these detectors, the physical-world experiments actually echo the digital world results, suggesting that *PaTNet* provides a promising way for determining the attack strength of an adversarial model in the real world from digital-world results.

Comparison with Other Approaches. We further compare our approach with object detector attacking methods including adv-patch (Thys et al., 2019), advT (Xu et al., 2020) and inv-cloak (Wu et al., 2020). For the physical world experiments, we obtained the adversarial patterns designed by these approaches from their arxived papers and printed the patterns out on T-shirts. These T-shirts were used to collect test data with the T-shirt designed by our approach at the same time and in the same scenes. So the experimental setting is fair for all methods. All scenes consider the lightning, distance and poses changes. For the digital world experiments, we use the public datasets COCO (Lin et al., 2014) and Inria (Dalal & Triggs, 2005), and apply all patches to the largest person object in an image by an affine transformation, follow the setting in (Thys et al., 2019).

Table 4: Comparisons with other approaches . (all based on YOLOv2)

Methods	Digital World		Physical World	
	COCO	Inria	Indoor	Outdoor
adv-patch (Thys et al., 2019)	65.3	**71.4**	12.2	8.9
advT (Xu et al., 2020)	68.2	69.5	18.7	45.0
inv-cloak (Wu et al., 2020)	**81.2**	67.0	27.3	34.1
PatNet	67.5	56.2	**67.3**	**58.4**

In Table 4, the results clearly show that *PatNet* achieves the best performance in the real world scenarios. Moreover, since all methods trained on different datasets, we cross-validate these methods on COCO (Lin et al., 2014) (used in (Wu et al., 2020)) and Inria (Dalal & Triggs, 2005) (used in (Thys et al., 2019)). Not surprisingly, *inv-cloak* (Wu et al., 2020) and *adv-patch* (Thys et al., 2019) perform well in the datasets they use for training and testing. Neverthless, our approach outperforms all the other approaches in the physical world test, clearly indicating that realistic simulation of physical transformations is critical for robust adversarial attacks in the physical world. By comparing with the results in the physical world and digital world, we can conclude that: 1) the results generated on the digital world datasets like COCO and Inria are not representative of real-world performance; 2) our approach *PatNet* is carefully designed for physical world attack and thus performs effectively in the realistic scenarios.

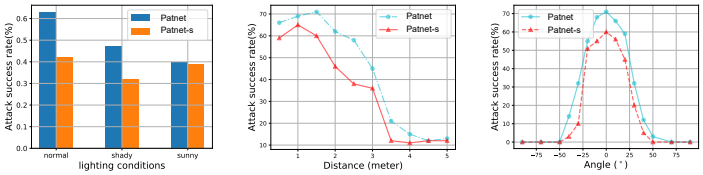


Figure 6: ASR (%) of *PatNet-s* and *PatNet* under different real-world scenarios.

Effects of Environment Conditions. To better understand the effects of environmental conditions on model performance, we conduct experiments dissected by camera distance, camera views and lighting conditions using the test videos (Figure 6). Specifically, when considering the lighting effect, the distance from the actor to the camera ranges from 1.5m to 3m and the camera angle is around $0°$. Similarly, the angle effect is evaluated under normal lighting conditions and a fixed distance. As can be observed in Figure 6, our approach demonstrates strong improvements ($5\% - 30\%$) over the baseline methods in almost all categories. The much better performance on various lightening conditions indicates simulating illumination changes in adversarial examples is of significance for an adversarial patch to survive in realistic environments. Our approach is also more robust to camera views up to 30°, though such robustness is unclear as the camera distance increases.

Especially, we notice that strong sun lights (i.e. a sunny condition) result in overexposed frames so that the ASR in this case are usually lower than that under other lighting conditions. This is a challenging problem that currently cannot be effectively addressed by our approach. Also, distance is an issue that future research should look into for developing stronger adversarial attack methods.

Detection Robustness Against Distance. It is under the impression that the robustness of a detector against adversarial attack increases by distance due to the negative effect of image resolution on the efficacy of the attack. We plot in Figure 7 the detection scores w/o adversarial attack (orange), the ASRs (green) and detection scores (blue) under attack by the height of the adversarial patch (a smaller height indicates a farther distance) for several detectors. Surprisingly, both YOLOv2 and YOLOv3 shows more resilience to attack at an intermediate distance around $180 \sim 200$ pixels. This is more differentiating for YOLOv2, indicated by a notable sag on its ASR curve. While there is no clear reason why this occurs to these detectors, we conjecture that it might be related to a combined effect of the test size of the detector and the patch size. Another insightful observation is that compared to YOLOv3 and RetinaNet101, YOLOv2 is more vulnerable to the physical adversarial attack (characterized by ASR) even at a long distance (corresponding to small patch size).By contrast, all the detectors can be fooled at a short distance.

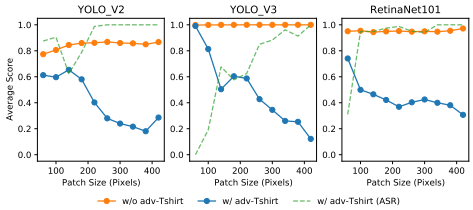


Figure 7: Detection scores w/o adversarial attack (orange), ASRs (green) and detection scores under attack (blue) w.r.t. patch size.

5 Conclusions

We have presented a unified framework for generation of physical adversarial patch, and demonstrated that it is much powerful compared with EoT, the current standard approach for physical adversarial example generation. Our framework can serve as a benchmark for future physical world patch simulation. While our architecture provides an end-to-end, consolidated approach to the multiple, complex transformations found in real world images, generating physical adversarial examples that work at long ranges remains an area for future improvement.

REFERENCES

Md Zahangir Alom, Tarek M Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S Awwal, and Vijayan K Asari. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, 2018.

A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 284–293, 10–15 Jul 2018.

Name Barnard and Kobus Barnard. Computational color constancy: Taking theory into practice. 09 1995.

Jonathan T Barron. Convolutional color constancy. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 379–387, 2015.

Battista Biggio, Fabio Roli, and xxx. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pp. 39–57. IEEE, 2017.

Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples. *arXiv preprint arXiv:1709.04114*, 2017.

Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 52–68. Springer, 2018.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, 2018.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pp. 886–893. Ieee, 2005.

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*, pp. 1000–1008, 2020.

Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*, 2017.

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

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Physical adversarial examples for object detectors. In *12th USENIX Workshop on Offensive Technologies (WOOT 18)*, 2018b.

Arjan Gijsenij, Theo Gevers, and Joost Van De Weijer. Computational color constancy: Survey and experiments. *IEEE Transactions on Image Processing*, 20(9):2475–2489, 2011.

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*, pp. 2672–2680, 2014.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *2015 ICLR*, arXiv preprint arXiv:1412.6572, 2015.

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

Yuanming Hu, Baoyuan Wang, and Stephen Lin. Fc4: Fully convolutional color constancy with confidence-weighted pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4085–4094, 2017.

Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 720–729, 2020.

A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. Black-box adversarial attacks with limited queries and information. *arXiv preprint arXiv:1804.08598*, 2018.

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

Steve T.K. Jan, Joseph Messou, Yen-Chen Lin, Jia-Bin Huang, and Gang Wang. Connecting the digital and physical world: Improving the robustness of adversarial attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):962–969, Jul. 2019. doi: 10.1609/aaai.v33i01.3301962. URL https://ojs.aaai.org/index.php/AAAI/article/view/3926.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

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

Juncheng Li, Frank Schmidt, and Zico Kolter. Adversarial camera stickers: A physical camera-based attack on deep learning systems. In *International Conference on Machine Learning*, pp. 3896–3904, 2019.

Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9455–9464, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pp. 21–37. Springer, 2016.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning based text classification: A comprehensive review. *arXiv preprint arXiv:2004.03705*, 2020.

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

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pp. 506–519. ACM, 2017.

Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 3500–3509, 2017.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Tiziano Portenier, Qiyang Hu, Attila Szabo, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. Faceshop: Deep sketch-based face image editing. *ACM Transactions on Graphics (TOG)*, 37(4):99, 2018.

Tiziano Portenier, Qiyang Hu, Paolo Favaro, and Matthias Zwicker. Smart, deep copy-paste. *CoRR*, abs/1903.06763, 2019. URL http://arxiv.org/abs/1903.06763.

Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

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*, pp. 1528–1540. ACM, 2016.

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

C. Sitawarin, A. N. Bhagoji, A. Mosenia, P. Mittal, and M. Chiang. Rogue signs: Deceiving traffic sign recognition with malicious ads and logos. *arXiv preprint arXiv:1801.02780*, 2018.

Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi. One pixel attack for fooling deep neural networks. *arXiv preprint arXiv:1710.08864*, 2017.

Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2020.

Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, and Xue Lin. Structured adversarial attack: Towards general implementation and better interpretability. In *International Conference on Learning Representations*, 2019.

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*, pp. 665–681. Springer, 2020.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational intelligenCe magazine*, 13(3):55–75, 2018.

Huanglin Yu, Ke Chen, Kaiqi Wang, Yanlin Qian, Zhaoxiang Zhang, and Kui Jia. Cascading convolutional color constancy. *arXiv preprint arXiv:1912.11180*, 2019a.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5505–5514, 2018.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4471–4480, 2019b.

Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJgEl3A5tm.

Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

APPENDIX

Dataset We built a large-scale dataset for evaluating our proposed approach. Similar to (Xu et al., 2020), we used a checkerboard T-shirts to collect 66 videos from four indoor and two outdoor scenes by iPhone cameras. Each video contains one, two, or three persons walking side by side towards a hand-held camera. In addition, to enable learning printer color transformation, we captured another 39 videos with a single person wearing a color-palette T-shirt (see Figure 3 and 5 in the main paper). The data present significant challenges such as distance, pose changes and varied illuminations. We uniformly sampled video frames at a rate of $1/5$, which results in $4,892$ frames. Among these frames, $2,830$ of them were used for training and $1,688$ for validation. All video frames are provided with the coordinates of the 4 corner points of the patch on the T-shirt and the bounding box of the person who wears an adversarial T-shirt. Some samples frames are shown in Fig. 5 in the main paper. Only the corner points were manually annotated. The person bounding boxes are provided by a detector based on Faster-RCNN.

Training Convergence In Fig. A1, we compare the convergence rate of our approach with *advP* (Thys et al., 2019). The EoT-based training converges much slower than the training driven by *PatNet*, suggesting that EoT learns both less efficiently and less effectively than our approach. Moreover, our approach benefits from the realistic examples generated by *PatNet*, demonstrating good learning capability on a large dataset with great complexity from real-world scenarios.

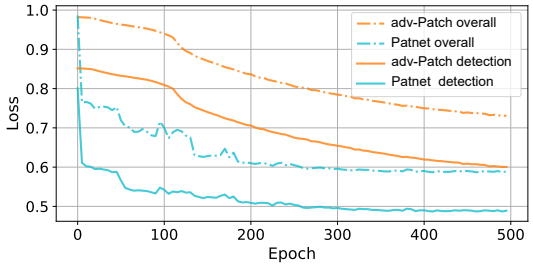


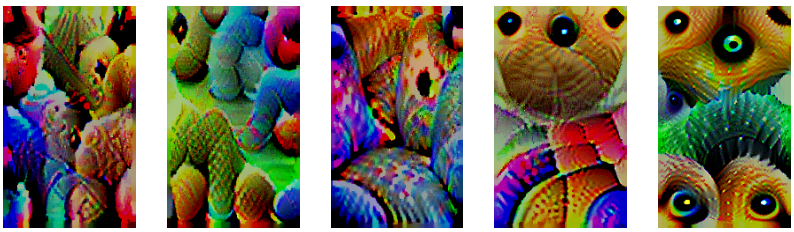
Figure A1: The overall and detection losses v.s. epoch for *advP* (Thys et al., 2019) and *advPat*. Our proposed approach converges much faster than *advP* based on EoT.

Localization We use ResNet18 as the backbone for the Spatial Transformer Network (i.e. STN), and define the Thin Plate Spline (TPS) transformation by 200 control points. We apply AlexNet to learn illumination adaptation in the color constancy model. As discussed in the main paper, learning our proposed *PaTNet* is based on self supervision through minimizing the ℓ_1 distance of the output of the IA module and the corresponding input image (See Fig 4 in the main paper). Table A1 shows the quality of patch transformation at each stage of *PaTNet*, i.e. geometric transformation (STN), printer color transformation (PCT) and illumination adaptation (IA), under two similarity metrics, namely the Structural Similarity Index Measure (SSIM) (Wang et al., 2004) and ℓ_1 distance. Not surprisingly, the IA module (the minimum ℓ_1 distance and the highest SSIM) produces the best visual similarity to the ground truth images. The high matching costs of PCT indicate that it does not transform digital colors close to their printer colors, as discussed in Section 3.2 in the main paper. In addition, the results confirm that TPS leads to better modeling of cloth deformation than Affine Transformation (Xu et al., 2020).

Table A1: Quality assessment of different transformations in *PaTNet*

Geometric	SSIM			ℓ_1		
Transformation	STN	PCT	IA	STN	PCT	IA
Affine	0.496	0.425	0.574	0.242	0.295	0.157
TPS	0.554	0.478	**0.645**	0.233	0.293	**0.140**

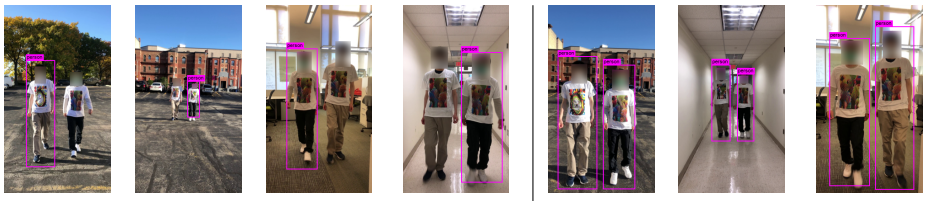
Adversarial Patches We show the adversarial patches generated by our approach against different detectors in Figure A2. The performance of these patches in the digital and physical worlds can be found in Table 3 and Table 5 of the main paper, respectively.



YOLOv2 YOLOv3 SSD FRCNN RetinaNet

Figure A2: Adversarial patches generated by *PaTNet* against different detectors.

Illustrative Examples Below illustrated in Fig A3 are examples of adversarial attacks from our approach, including both successful and failure cases.



L: *advT* R: *PatNet-s* L:*PatNet-s* R: *PatNet-sp* | L: *advT* L: *PatNet-s* L: *PatNet-sp*

Figure A3: Examples of adversarial attacks in the physical world. Successful cases: images 1-4 from left to right; failure cases: images 5-7. **L** or **R** indicates the location of an adversarial T-shirt generated from a baseline approach, and the other one is our approach *PatNet*.