Evading Real-Time Person Detectors by Adversarial T-shirt

Kaidi Xu[1] Gaoyuan Zhang[2] Sijia Liu[2] Quanfu Fan[2] Mengshu Sun[1]
Hongge Chen[3] Pin-Yu Chen[2] Yanzhi Wang[1] Xue Lin[1]

[1]Northeastern University, USA
[2]MIT-IBM Watson AI Lab, IBM Research, USA
[3]Massachusetts Institute of Technology, USA

October 25, 2019

Abstract

It is known that deep neural networks (DNNs) could be vulnerable to adversarial attacks. The so-called *physical adversarial examples* deceive DNN-based decision makers by attaching adversarial patches to real objects. However, most of the existing works on physical adversarial attacks focus on static objects such as glass frame, stop sign and image attached to a cardboard. In this work, we propose *Adversarial T-shirt*, a robust physical adversarial example for evading person detectors even if it suffers from deformation due to a moving person's pose change. To the best of our knowledge, the effect of deformation is first modelled for designing physical adversarial examples with respect to non-rigid objects such as T-shirts. We show that the proposed method achieve 79% and 63% attack success rates in digital and physical worlds respectively against YOLOv2. In contrast, the state-of-the-art physical attack method to fool a person detector only achieves 27% attack success rate. Furthermore, by leveraging min-max optimization, we extend our method to the ensemble attack setting against object detectors YOLOv2 and Faster R-CNN simultaneously.

1 Introduction

The vulnerability of deep neural networks (DNNs) against adversarial attacks (namely, perturbed inputs deceiving DNNs) has been found in applications spanning from image classification to speech recognition [1, 2, 3, 4, 5, 6, 7]. Early works studied adversarial examples in the digital space only. Recently, some works showed that it is possible to create adversarial perturbations on physical objects and fool DNN-based decision makers under a variety of real-world conditions [8, 9, 10, 11, 12, 13, 14, 15, 16]. However, most of the studied *physical adversarial attacks* encounter two limitations: a) the physical objects are usually considered being *static*, and b) the possible *deformation* of adversarial pattern attached to a moving object (e.g., due to pose change of a moving person) is commonly neglected. In this paper, we propose a new type of physical adversarial attack, *adversarial T-shirt*, to evade a real-time person detector when the person moves and wears the adversarial T-shirt; see the second and the third rows of Figure 1 for illustrative examples.

Most of the existing physical adversarial attacks were generated against image classifiers and object detectors. In [8], a face recognition system is fooled by a real eyeglass frame designed under a crafted adversarial pattern. In [9], a stop sign is misclassified by adding black or white stickers on it against image classification system. In [16], an image classifier is fooled by placing a crafted sticker at the lens of a camera. In [10], a so-called Expectation over Transformation (EoT) framework was proposed to synthesize adversarial examples robust to a set of physical transformations such as rotation, translation, contrast, brightness, and random noise. Compared to attacking image classifiers, generating physical adversarial attacks against object detectors is more challenging since the adversary is required to mislead both bounding box detector and object classifier. A well-known success is the generation of adversarial stop sign [11], which deceives state-of-the-art object detectors such as YOLOv2 [17] and Faster R-CNN [18].

The most relevant work to ours is [14], in which a person detector is fooled when the person holds a cardboard plate printed by an adversarial patch. However, such a physical attack restricts the adversarial patch to be attached to a *rigid* carrier (cardbolad), and is not directly applied to the design of adversarial T-shirt. We show that the attack proposed by [14] becomes ineffective when the adversarial patch is attached to a T-shirt (rather than a cardboard) and worn by a moving person (see the fourth row of Figure 1). At the technical side, different from [14] we propose a thin plate spline (TPS) based transformer to model the deformation effect of a non-rigid object, and we develop an ensemble physical attack that fools object detectors YOLOv2 and Faster R-CNN simultaneously. We highlight that the proposed adversarial T-shirt is not just a T-shirt with printed adversarial patch for clothing fashion, it is a physical adversarial wearable designed for evading person detectors in a real world.

Our work is also motivated by the importance of person detection on intelligent surveillance. DNN-based surveillance systems have significantly advanced the field of object detection [19, 20]. Efficient object detectors such as faster R-CNN [18], SSD [21], and YOLOv2 [17] have been deployed for human detection. Thus, one may wonder whether or not there exists a security risk for intelligent surveillance systems caused by adversarial human wearables, e.g., adversarial T-shirt. However, paralyzing a person detector in the physical world requires substantially more challenges such as low resolution, pose change and occlusion.

Contribution. We summarize our contributions as follows.

- We develop a TPS based transformer to model the temporal deformation of adversarial T-shirt caused by pose change of a moving person. We also show its importance to ensure the effectiveness of adversarial T-shirt in the physical world.
- We propose a general optimization framework for design of adversarial T-shirt in both single-detector and multiple-detector settings.
- We conduct experiments in both digital and physical worlds and show that the proposed adversarial T-shirt achieves 79% and 63% attack success rates respectively when attacking YOLOv2. By contrast, the physical adversarial patch [14] printed on a T-shirt only achieves 27% attack success rate. Some of our results are highlighted in Figure 1.

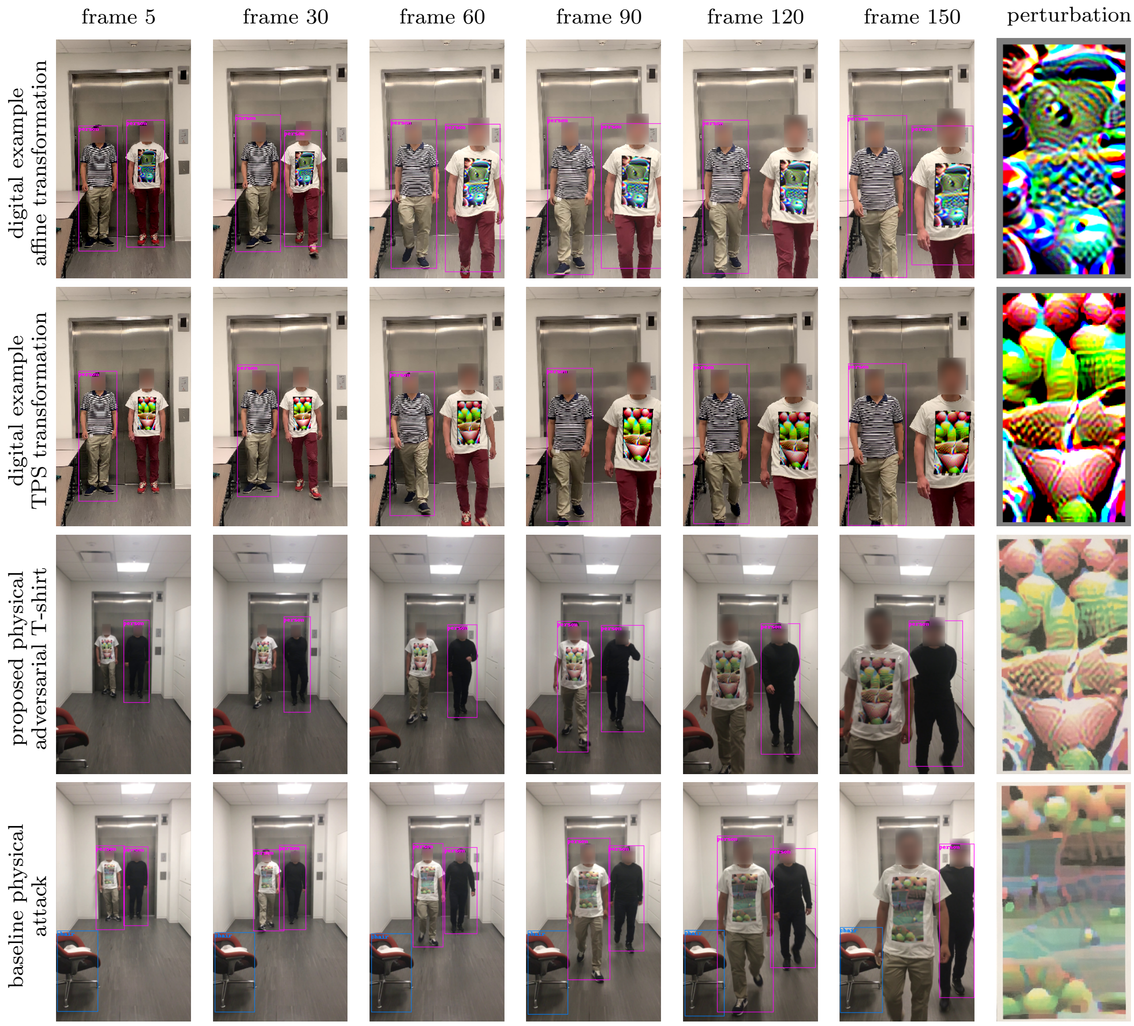


Figure 1: Evaluating effectiveness of adversarial T-shirt to evade person detection by YOLOv2. Each row corresponds to a specific attack method, and each column denotes a video frame except the last column, which shows the generated adversarial pattern. There exist two persons at each frame, and only one person wears the adversarial T-shirt. First row: digital adversarial T-shirt generated with affine transformation (namely, in the absence of modeling deformation). Second row: digital adversarial T-shirt generated using TPS. Third row: physical adversarial T-shirt generated by our method. Fourth row: physical adversarial patch generated by [14] printed on a T-shirt.

2 Modeling Deformation of A Moving Object by Thin Plate Spline Mapping

In this section, we begin by reviewing some existing transformations required in the design of physical adversarial examples. We then elaborate on Thin Plate Spline (TPS) mapping used to model the possible deformation encountered by a moving and non-rigid object.

Let $\mathbf{x}$ be an original image (or a video frame), and $t(\cdot)$ be the physical transformer. The transformed image $\mathbf{z}$ under t is given by

$$\mathbf{z} = t(\mathbf{x}). \tag{1}$$

Existing transformations. In [10], the parametric transformers include scaling, translation, rotation, brightness and additive Gaussian noise; see details in [10, Appendix D]. In [22], the geometry and lighting transformations are studied via parametric models. Other transformations including perspective transformation, brightness adjustment, resampling (or image resizing), smoothing and saturation are considered in [23, 24]. All the existing transformations are included in our library of physical transformations. However, they are not sufficient to model the cloth deformation caused by pose change of a moving person. For example, the first and fourth rows of Figure 1 show that adversarial T-shirts designed against only existing physical transformations yield low attack success rates.

TPS transformation for cloth deformation. A person's movement can result in significant and constantly changing wrinkles (aka deformations) in her clothes. This makes it challenging to develop adversarial T-shirt effectively in the real world. To circumvent this challenge, we employ TPS mapping [25] to model the cloth deformation caused by human body movement. TPS has been widely used as the non-rigid transformation model in image alignment and shape matching [26]. It consists of an affine component and a non-affine warping component. We will show that the non-linear warping part in TPS can provide an effective means of modeling cloth deformation for learning adversarial patterns of non-rigid objects.

TPS learns a parametric deformation mapping from an original image $\mathbf{x}$ to a target image $\mathbf{z}$ through a set of control points with given positions. Let $\mathbf{p} := (\phi, \psi)$ denote the 2D location of an image pixel. The deformation from $\mathbf{x}$ to $\mathbf{z}$ is then characterized by the displacement of every pixel, namely, how a pixel at $\mathbf{p}^{(x)}$ on image $\mathbf{x}$ changes to the pixel on image $\mathbf{z}$ at $\mathbf{p}^{(z)}$, where $\phi^{(z)} = \phi^{(x)} + \Delta_\phi$ and $\psi^{(z)} = \psi^{(x)} + \Delta_\psi$, and Δ_ϕ and Δ_ψ denote the pixel displacement on image $\mathbf{x}$ along ϕ direction and ψ direction, respectively. Given a set of n control points with locations $\{\hat{\mathbf{p}}_i^{(x)} := (\hat{\phi}_i^{(x)}, \hat{\psi}_i^{(x)})\}_{i=1}^n$ on image $\mathbf{x}$, TPS provides a parametric model of pixel displacement when mapping $\mathbf{p}^{(x)}$ to $\mathbf{p}^{(z)}$ [27]

$$\Delta(\mathbf{p}^{(x)}; \boldsymbol{\theta}) = a_0 + a_1\phi^{(x)} + a_2\psi^{(x)} + \sum_{i=1}^{n} c_i U(\|\hat{\mathbf{p}}_i^{(x)} - \mathbf{x}\|_2), \tag{2}$$

where $U(r) = r^2 \log(r)$ and $\boldsymbol{\theta} = [\mathbf{c}; \mathbf{a}]$ are the TPS parameters, and $\Delta(\mathbf{p}^{(x)}; \boldsymbol{\theta})$ represents the displacement along either ϕ or ψ direction.

To determine $\boldsymbol{\theta}$ in (2), TPS resorts to a regression model given the locations of control points on the transformed image $\mathbf{z}$, namely, $\{\hat{\mathbf{p}}_i^{(z)}\}_{i=1}^n$. This then yields a regression problem which minimizes the distance between $\{\Delta_\phi(\mathbf{p}_i^{(x)}; \boldsymbol{\theta}_\phi)\}_{i=1}^n$ and $\{\hat{\Delta}_{\phi,i} := \hat{\phi}_i^{(z)} - \hat{\phi}_i^{(x)}\}_{i=1}^n$ and the distance between $\{\Delta_\psi(\mathbf{p}_i^{(x)}; \boldsymbol{\theta}_\psi)\}_{i=1}^n$ and $\{\hat{\Delta}_{\psi,i} := \hat{\psi}_i^{(z)} - \hat{\psi}_i^{(x)}\}_{i=1}^n$, respectively. Here TPS (2) is applied to coordinate ϕ and ψ separately (corresponding to parameters $\boldsymbol{\theta}_\phi$ and $\boldsymbol{\theta}_\psi$). The regression problem can be solved by the following linear system of equations [28]

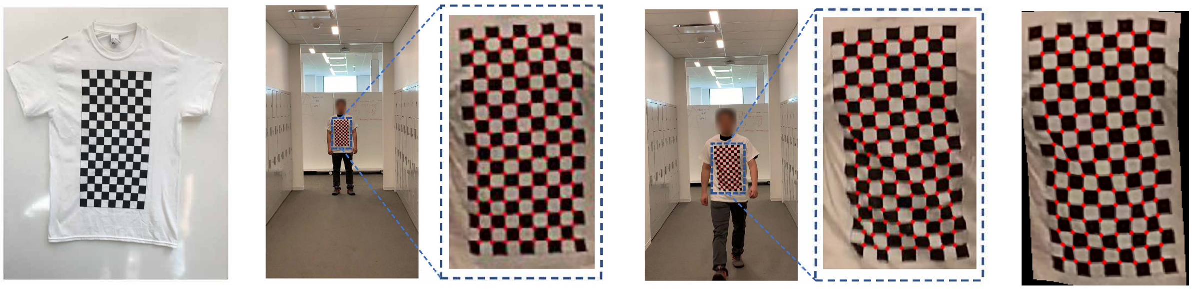
$$\begin{bmatrix} \mathbf{K} & \mathbf{P} \\ \mathbf{P}^T & \mathbf{0}_{3\times3} \end{bmatrix} \boldsymbol{\theta}_\phi = \begin{bmatrix} \hat{\mathbf{\Delta}}_\phi \\ \mathbf{0}_{3\times1} \end{bmatrix}, \quad \begin{bmatrix} \mathbf{K} & \mathbf{P} \\ \mathbf{P}^T & \mathbf{0}_{3\times3} \end{bmatrix} \boldsymbol{\theta}_\psi = \begin{bmatrix} \hat{\mathbf{\Delta}}_\psi \\ \mathbf{0}_{3\times1} \end{bmatrix} \tag{3}$$

where the (i,j)th element of $\mathbf{K} \in \mathbb{R}^{n\times n}$ is given by $K_{ij} = U(\|\hat{\mathbf{p}}_i^{(x)} - \hat{\mathbf{p}}_j^{(x)}\|_2)$, the ith row of $\mathbf{P} \in \mathbb{R}^{n\times 3}$ is given by $P_i = [1, \hat{\phi}_i^{(x)}, \hat{\psi}_i^{(x)}]$, and the ith elements of $\hat{\mathbf{\Delta}}_\phi \in \mathbb{R}^n$ and $\hat{\mathbf{\Delta}}_\psi \in \mathbb{R}^n$ are given by $\hat{\Delta}_{\phi,i}$ and $\hat{\Delta}_{\psi,i}$, respectively.

The difficulty of implementing TPS for design of adversarial T-shirt is how to determine the set of control points and obtain positions $\{\hat{\mathbf{p}}_i^{(x)}\}$ and $\{\hat{\mathbf{p}}_i^{(z)}\}$ in both original and target images. Spurred by [29] for camera calibration, we print a *checkerboard* on a T-shirt and use it to collect control points and their positions between two video frames. In practice, we selected one frame as the anchor frame $\mathbf{x}$, then generate TPS from other frames. Figure 2 shows the T-shirt with the checkerboard pattern, where each intersection between two checkerboard grid regions is selected as a control point. We remark that the considered control points can be accurately detected using the Matlab vision toolbox [30], and the videos used to generate TPS transformations are independent of testing data for evaluation of the proposed adversarial T-shirt.

3 Generation of Adversarial T-shirt: An Optimization Perspective

In this section, we begin by formalizing the problem of adversarial T-shirt and introducing notations used in our setup. We then propose to design a *universal* perturbation used in our adversarial T-shirt deceiving a *single* object detector. We lastly propose a min-max (robust) optimization framework to design the universal adversarial patch against *multiple* object detectors.



(a) (b) (c) (d)

Figure 2: (a): examples of our T-shirt with printed checkerboard to construct control points for TPS transformation. (b) and (c): two frames with checkerboard detection results. (d): result of applying TPS transformation from (b) to (c).

Let $\mathcal{D} := \{\mathbf{x}_i\}_{i=1}^M$ denote M video frames extracted from one or multiple given videos, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the ith frame. Let $\boldsymbol{\delta} \in \mathbb{R}^d$ denote the universal adversarial perturbation applied to $\mathcal{D}$. The adversarial T-shirt is then characterized by $M_{c,i} \circ \boldsymbol{\delta}$, where $M_{c,i} \in \{0,1\}^d$ is a bounding box encoding the position of the cloth region to be perturbed at the ith frame, and $\circ$ denotes element-wise product. *The goal of adversarial T-shirt is to design $\boldsymbol{\delta}$ such that the perturbed frames of $\mathcal{D}$ are mis-detected by object detectors.*

Fooling a single object detector. We generalize the Expectation over Transformation (EoT) method in [31] for design of adversarial T-shirt. Note that different from the conventional EoT, a transformers' composition is required for generating adversarial T-shirt. For example, a perspective transformation on the bounding box of T-shirt is composited with TPS transformation on the perturbed cloth region.

Let us begin by considering two video frames, an anchor image $\mathbf{x}_0$ (e.g., the first frame in the video) and a target image $\mathbf{x}_i$ for $i \in [M]$[1]. Given the bounding boxes of the person ($M_{p,0} \in \{0,1\}^d$) and the T-shirt ($M_{c,0} \in \{0,1\}^d$) at $\mathbf{x}_0$, we apply the perspective transformation from $\mathbf{x}_0$ to $\mathbf{x}_i$ to obtain the bounding boxes $M_{p,i}$ and $M_{c,i}$ at image $\mathbf{x}_i$. In the *absence* of physical transformations, the perturbed image $\mathbf{x}_i'$ with respect to (w.r.t.) $\mathbf{x}_i$ is given by

$$\mathbf{x}_i' = \underbrace{(\mathbf{1} - M_{p,i}) \circ \mathbf{x}_i}_{\text{A}} + \underbrace{M_{p,i} \circ \mathbf{x}_i}_{\text{B}} - \underbrace{M_{c,i} \circ \mathbf{x}_i}_{\text{C}} + \underbrace{M_{c,i} \circ \boldsymbol{\delta}}_{\text{D}}, \tag{4}$$

where the term A denotes the background region outside the bouding box of the person, the term B is the person-bounded region, the term C erases the pixel values within the bounding box of the T-shirt, and the term D is the newly introduced additive perturbation. Without taking into account physical transformations, Eq. (4) simply reduces to the conventional formulation of adversarial example $(1 - M_{c,i}) \circ \mathbf{x}_i + M_{c,i} \circ \boldsymbol{\delta}$.

We next consider *two categories* of physical transformations: a) TPS transformation $t_{\text{TPS}} \in \mathcal{T}_{\text{TPS}}$ applying to the adversarial perturbation $\boldsymbol{\delta}$ for modeling the effect of cloth deformation, and b) conventional physical transformation $t \in \mathcal{T}$ applying to the region within the person's bounding box, namely, $(M_{p,i} \circ \mathbf{x}_i - M_{c,i} \circ \mathbf{x}_i + M_{c,i} \circ \boldsymbol{\delta})$. Here $\mathcal{T}_{\text{TPS}}$ denotes the set of possible non-rigid transformations, and $\mathcal{T}$ denotes the set of commonly-used physical transformations, e.g., scaling, translation, rotation, brightness, blurring and contrast. A modification of (4) under different sources of transformations is then given by

$$\mathbf{x}_i' = t_{\text{env}}\left(\underbrace{(\mathbf{1} - M_{p,i}) \circ \mathbf{x}_i}_{\text{A}} + t_{\text{person}}\left(\underbrace{M_{p,i} \circ \mathbf{x}_i}_{\text{B}} - \underbrace{M_{c,i} \circ \mathbf{x}_i}_{\text{C}} + \underbrace{M_{c,i} \circ t_{\text{TPS}}(\boldsymbol{\delta} + \mathbf{v})}_{\text{D}}\right)\right), \quad t_{\text{env}}, t_{\text{person}} \in \mathcal{T},\ t_{\text{TPS}} \in \mathcal{T}_{\text{TPS}} \tag{5}$$

where $\mathbf{v}$ is an additive random noise that allows the variation of pixel values, e.g., due to the mismatch between the digital color and the printed color, t_{env} is a transformer modelling the environmental condition and we set it as brightness transformation in practice, and t_{person} denotes a transformer applied to the image region characterized by a person's bounding box.

With the aid of (5), the EoT formulation to fool a single object detector is cast as

$$\begin{array}{ll} \underset{\boldsymbol{\delta}}{\text{minimize}} & \frac{1}{M}\sum_{i=1}^M \mathbb{E}_{t \in \mathcal{T}, t_{\text{TPS}} \in \mathcal{T}_{\text{TPS}}}\left[f(\mathbf{x}_i')\right] + \lambda g(\boldsymbol{\delta}) \\ \text{subject to} & \boldsymbol{\delta} \in \mathcal{C}, \end{array} \tag{6}$$

where f denotes an attack loss for misdetection, g is the total-variation norm that enhances perturbations' smoothness [11], $\lambda > 0$ is a regularization parameter, and $\mathcal{C}$ signifies additional constraints that $\boldsymbol{\delta}$ should follow, e.g., a discrete set of printable color options for generating physical adversarial examples [8].

[1] $[M]$ denotes the integer set $\{1, 2, \ldots, M\}$.

Min-max optimization for fooling multiple object detectors. The transferability of adversarial attacks largely drops in the physical environment, thus we consider a *physical ensemble attack* against multiple object detectors. It was recently shown in [32] that the ensemble attack can be designed from the perspective of min-max optimization, and yields much higher worst-case attack success rate than the averaging strategy over multiple models. Given N object detectors associated with attack loss functions $\{f_i\}_{i=1}^N$, the physical ensemble attack is cast as

$$\underset{\boldsymbol{\delta}\in\mathcal{C}}{\text{minimize}}\ \underset{\mathbf{w}\in\mathcal{P}}{\text{maximize}} \quad \sum_{i=1}^N w_i \phi_i(\boldsymbol{\delta}) - \frac{\gamma}{2}\|\mathbf{w} - \mathbf{1}/N\|_2^2 + \lambda g(\boldsymbol{\delta}), \tag{7}$$

where $\mathbf{w}$ are known as domain weights that adjust the importance of each object detector during the attack generation, $\mathcal{P}$ is a probabilistic simplex given by $\mathcal{P} = \{\mathbf{w}|\mathbf{1}^T\mathbf{w} = 1, \mathbf{w} \geq \mathbf{0}\}$, $\gamma > 0$ is a regularization parameter, and $\phi_i(\boldsymbol{\delta}) := \frac{1}{M}\sum_{i=1}^M \mathbb{E}_{t\in\mathcal{T}, t_{\text{TPS}}\in\mathcal{T}_{\text{TPS}}}[f(\mathbf{x}_i')]$ following (6). In (7), if $\gamma = 0$, then the adversarial perturbation $\boldsymbol{\delta}$ is designed over the *maximum* attack loss (worst-case attack scenario) since $\text{maximize}_{\mathbf{w}\in\mathcal{P}} \sum_{i=1}^N w_i\phi_i(\boldsymbol{\delta}) = \phi_{i^*}(\boldsymbol{\delta})$, where $i^* = \arg\max_i \phi_i(\boldsymbol{\delta})$ at a fixed $\boldsymbol{\delta}$. Moreover, if $\gamma \to \infty$, then the inner maximization of problem (7) implies $\mathbf{w} \to \mathbf{1}/N$, namely, an averaging scheme over M attack losses. Thus, the regularization parameter γ in (7) strikes a balance between the max-strategy and the average-strategy.

4 Experimental Results

In this section, we demonstrate the effectiveness of our approach for design of adversarial T-shirt by comparing it with 2 baseline attack methods, adversarial patch to fool YOLOv2 [14] and the variant of our approach in the absence of TPS transformation, namely, $\mathcal{T}_{\text{TPS}} = \emptyset$ in (5). We examine the convergence behavior of our proposed algorithm as well as its attack success rate (ASR) in both digital and physical worlds.

4.1 Experimental Setup

Data collection. We collect two datasets for learning and testing our proposed attack algorithm in both digital and physical worlds. The training dataset contains 30 videos, each of which takes 5-10 seconds and is captured by a moving person wearing a T-shirt with printed checkerboard under 4 different scenes. The desired adversarial pattern is then learnt from the training dataset. The second dataset contains 10 videos captured in the same setting of the training dataset (but from different persons). This dataset is used to evaluate the attack performance of the learnt adversarial pattern in the digital world. In the physical world, we create an adversarial T-shirt by printing our learnt adversarial pattern on the T-shirt. The 10 test videos are then collected from a moving person wearing the physical adversarial T-shirt. All videos are taken using an iPhone 7 Plus and are resized to 416×416.

Object detectors. We use two state-of-the-art object detectors: Faster R-CNN [18] and YOLOv2 [17] to evaluate our method. These two object detectors are both pre-trained on COCO dataset [33] which contains 80 classes including 'person'. The detection minimum threshold are set as 0.6 and 0.7 for Faster R-CNN and YOLOv2 by default respectively. For the misdetection loss f in Eq. (6) and Eq. (7), we refer [13] and [14] for our attack against Faster R-CNN and YOLOv2, respectively.

Algorithmic parameter setting. When solving Eq. (6), we use Adam optimizer [34] to train 3500 epochs, and the learning rate is set to 1×10^{-4} and decay to 2×10^{-5} at 750th epochs. The regularization parameter λ for total-variation norm is set as 5. In Eq. (7), we set γ as 1, and solve the min-max problem by 5000 epochs with initial learning rate 1×10^{-4}.

4.2 Adversarial T-shirt in digital world

Convergence performance of the proposed attack algorithm. In Figure 3, we show the convergence of our proposed algorithm to solve problem (6), in terms of attack loss and attack success rate (ASR) against epoch number. Here ASR is given by the ratio of successfully attacked testing frames over the total number of testing frames. We see that the proposed attack method is well-behaved in convergence. We also note that attacking Faster R-CNN is more difficult than attacking YOLOv2.

ASR of adversarial T-shirt in various attack settings. We perform a comprehensive evaluation on our methods in digital simulation. In Table 1, we compare ASR of adversarial T-shirt with or without using TPS transformation under 4 attack settings: a) Single-detector attack refers to adversarial T-shirt designed and evaluated using the same object detector; b) Transfer attack refers to adversarial T-shirt designed and evaluated using different object detectors; and c) ensemble attack (average) and ensemble attack (min-max) refer to the design of ensemble attack using the averaged attack loss and the min-max attack loss in (7), respectively. As we can see, it is crucial to incorporate TPS transformation in the design of adversarial T-shirt: ASR drops from 0.65 to 0.36 when attacking faster R-CNN and drops from 0.79 to 0.52 when attacking YOLOv2 in the single-detector attack setting. We also note that the transferability of single-detector attack is poor, and consistent

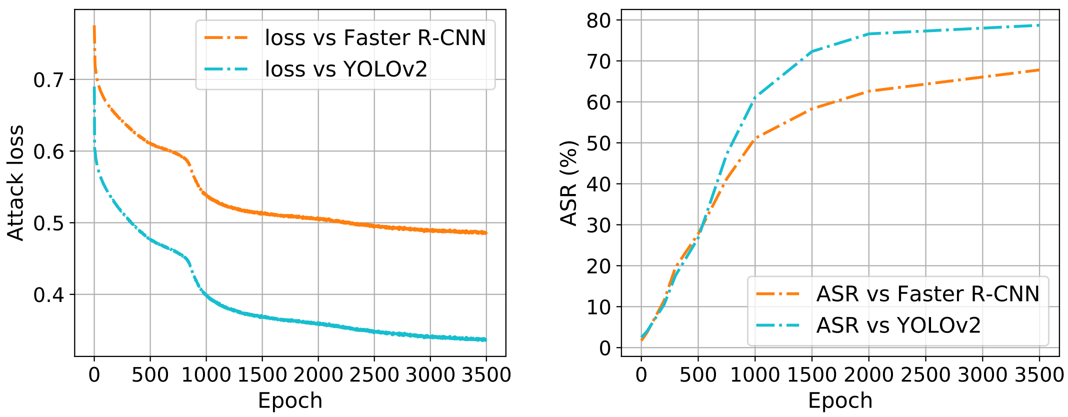


Figure 3: Left: Attack loss vs. epoch numbers when generating perturbations by solving problem (6) for Faster R-CNN and YOLOv2. Right: ASR vs. epoch numbers.

with Figure 3, faster R-CNN is more robust than YOLOv2. Furthermore, we evaluate the effectiveness of the proposed min-max ensemble attack (7). As we can see, when attacking faster R-CNN, the min-max ensemble attack significantly outperforms its counterpart using the averaging strategy, leading to 20% improvement in ASR. This improvement is at the cost of 4% degradation when attacking YOLOv2.

4.3 Adversarial T-shirt in physical world

Next, we evaluate our method in the physical world setting. We generate adversarial pattern by solving problem (6) against YOLOv2, same as Section 4.2. We then print the adversarial pattern and paste it on a white T-shirt, obtaining the adversarial T-shirt. At the testing phase, we use iPhone 5 to record videos for tracking a moving person wearing the obtained adversarial T-shirt. We compare our method with the baseline method in [14], where both methods are trained using the same dataset for design of adversarial patches with the same size. We show in Table 2 that our method achieves 63% ASR, which is much higher than 27% for baseline method when attacking YOLOv2. For comparison, we also present ASRs as attacking Faster R-CNN, which was not considered in [14].

Table 1: The ASR (%) of adversarial T-shirt with or without (w/o) using TPS transformation under four attack settings.

Model	w/o TPS	with TPS
single-detector attack		
Faster R-CNN	36%	65%
YOLOv2	52%	79%
transfer single-detector attack		
Faster-RCNN	8%	9%
YOLOv2	12%	11%
ensemble attack (average)		
Faster-RCNN	21%	41%
YOLOv2	36%	69%
ensemble attack (min-max)		
Faster-RCNN	47%	60%
YOLOv2	33%	65%

Figure 4 elaborates on our physical-world attack results in three settings: a) single moving person (row 1&2 of Figure 4), b) two moving persons wearing adversarial T-shirts generated using our method and the baseline method in [14] respectively (row 3), and c) two moving persons wearing the proposed adversarial T-shirt and the normal T-shirt respectively (row 4). As we can see, the baseline method failed in most of cases since it neglects the factor of T-shirt deformation, and was designed for generating adversarial pattern on a rigid object (cardboard). We also note even if the moving person is far from the camera, the proposed physical attack is still powerful. By contrast, the baseline method and the case of normal T-shirt can still be detected by an object detector. Compared to the digital results, ASRs in the physical world drop around 15%.

5 Conclusion

In this paper, we propose *Adversarial T-shirt*, the first successful adversarial wearable to evade detection of moving persons. Since T-shirt is a non-rigid object, its deformation induced by pose change of a moving person is taken into account when generating adversarial perturbations. We also propose a min-max ensemble attack algorithm to fool multiple object detectors simultaneously. We show that in both digital and physical worlds, our attack method can achieve 79% and 63% attack success rate (ASR), respectively. By contrast, the baseline method can only achieve 27% ASR. Based on our studies, we hope to provide some

Table 2: ASRs of our method and baseline method in [14] when attacking YOLOv2 and Faster R-CNN in the physical world.

Model \ ASR	baseline	our method
Faster R-CNN	11%	52%
YOLOv2	27%	63%

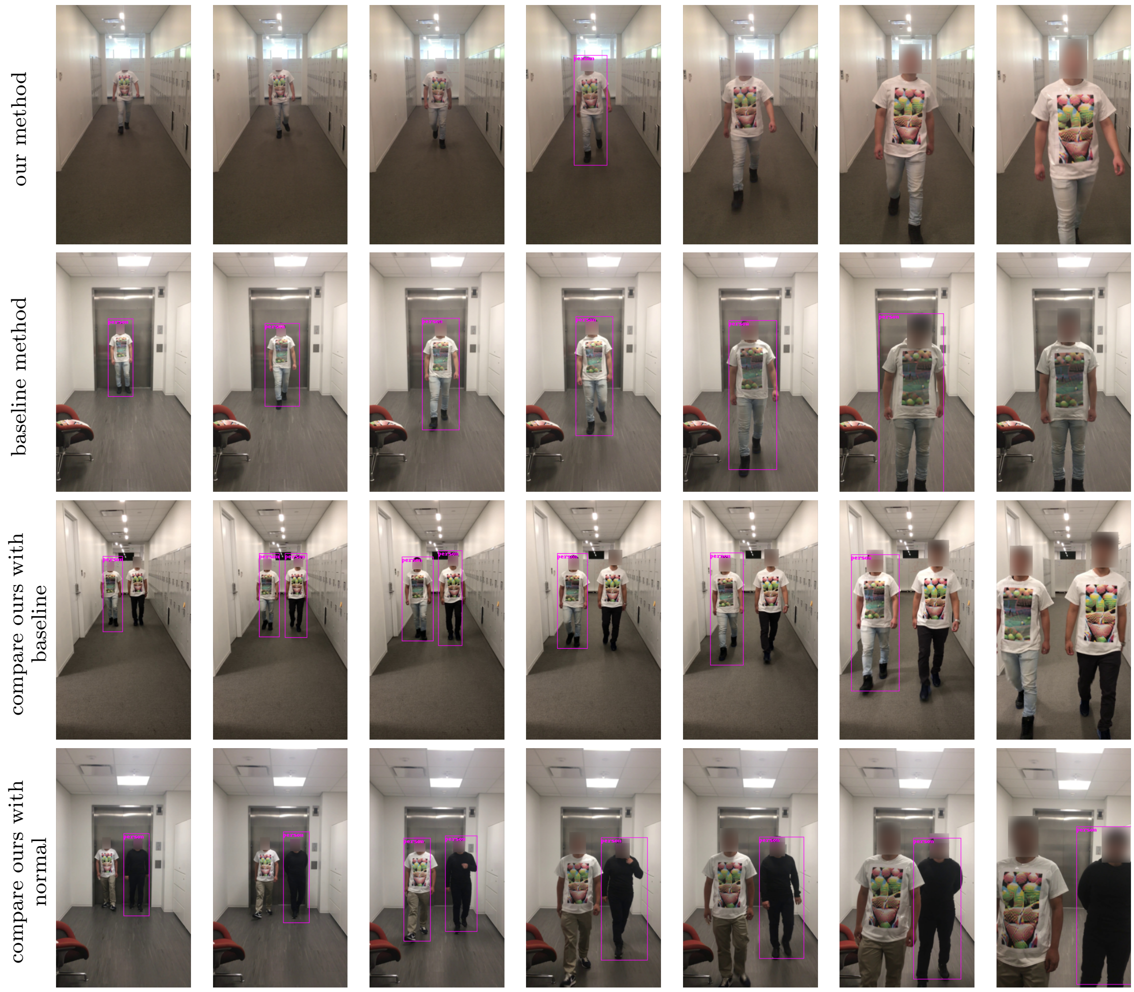


Figure 4: Some testing frames in the physical world using adversarial T-shirt against YOLOv2. First row: detecting a single person wearing the proposed adversarial T-shirt. Second row: detecting a single person detection wearing the adversarial T-shirt generated from the baseline method [14]. Third row: detecting two moving persons wearing two adversarial T-shirts generated from our method and the baseline method, respectively. Fourth row: detecting two moving persons wearing the proposed adversarial T-shirt and the normal cloth, respectively.

implications on how the adversarial perturbations can be implemented with human clothing, accessories, paint on face, and other wearables.

References

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[2] K. Xu, S. Liu, P. Zhao, P.-Y. Chen, H. Zhang, Q. Fan, D. Erdogmus, Y. Wang, and X. Lin, "Structured adversarial attack: Towards general implementation and better interpretability," in *International Conference on Learning Representations*, 2019.

[3] P. Zhao, K. Xu, S. Liu, Y. Wang, and X. Lin, "Admm attack: an enhanced adversarial attack for deep neural networks with undetectable distortions," in *Proceedings of the 24th Asia and South Pacific Design Automation Conference*, pp. 499–505, ACM, 2019.

[4] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7, IEEE, 2018.

[5] K. Xu, H. Chen, S. Liu, P.-Y. Chen, T.-W. Weng, M. Hong, and X. Lin, "Topology attack and defense for graph neural networks: An optimization perspective," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

[6] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," *arXiv preprint arXiv:1802.00420*, 2018.

[7] K. Xu, S. Liu, G. Zhang, M. Sun, P. Zhao, Q. Fan, C. Gan, and X. Lin, "Interpreting adversarial examples by activation promotion and suppression," *arXiv preprint arXiv:1904.02057*, 2019.

[8] M. Sharif, S. Bhagavatula, L. Bauer, and M. 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.

[9] 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, 2018.

[10] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80, pp. 284–293, 10–15 Jul 2018.

[11] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramer, A. Prakash, T. Kohno, and D. Song, "Physical adversarial examples for object detectors," in *12th USENIX Workshop on Offensive Technologies (WOOT 18)*, 2018.

[12] J. Lu, H. Sibai, and E. Fabry, "Adversarial examples that fool detectors," *arXiv preprint arXiv:1712.02494*, 2017.

[13] S.-T. Chen, C. Cornelius, J. Martin, and D. H. P. 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.

[14] S. Thys, W. Van Ranst, and T. 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.

[15] Y. Cao, C. Xiao, D. Yang, J. Fang, R. Yang, M. Liu, and B. Li, "Adversarial objects against lidar-based autonomous driving systems," *arXiv preprint arXiv:1907.05418*, 2019.

[16] J. Li, F. Schmidt, and Z. Kolter, "Adversarial camera stickers: A physical camera-based attack on deep learning systems," in *International Conference on Machine Learning*, pp. 3896–3904, 2019.

[17] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

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

[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[20] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[22] H.-T. D. Liu, M. Tao, C.-L. Li, D. Nowrouzezahrai, and A. Jacobson, “Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer,” in *International Conference on Learning Representations*, 2019.

[23] 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.

[24] G. W. Ding, K. Y. C. Lui, X. Jin, L. Wang, and R. Huang, “On the sensitivity of adversarial robustness to input data distributions,” in *International Conference on Learning Representations*, 2019.

[25] F. L. Bookstein, “Principal warps: Thin-plate splines and the decomposition of deformations,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 11, no. 6, pp. 567–585, 1989.

[26] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, “Spatial transformer networks,” in *Advances in neural information processing systems*, pp. 2017–2025, 2015.

[27] H. Chui, “Non-rigid point matching: algorithms, extensions and applications,” 2001.

[28] G. Donato and S. Belongie, “Approximate thin plate spline mappings,” in *European conference on computer vision*, pp. 21–31, Springer, 2002.

[29] Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, 2000.

[30] A. Geiger, F. Moosmann, Ö. Car, and B. Schuster, “Automatic camera and range sensor calibration using a single shot,” in *2012 IEEE International Conference on Robotics and Automation*, pp. 3936–3943, IEEE, 2012.

[31] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, “Synthesizing robust adversarial examples,” in *International Conference on Machine Learning*, pp. 284–293, 2018.

[32] J. Wang, T. Zhang, S. Liu, P.-Y. Chen, J. Xu, M. Fardad, and B. Li, “Beyond adversarial training: Min-max optimization in adversarial attack and defense,” *arXiv preprint arXiv:1906.03563*, 2019.

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[34] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *2015 ICLR*, vol. arXiv preprint arXiv:1412.6980, 2015.