Physical Adversarial Attacks on an Aerial Imagery Object Detector

Andrew Du[†] Bo Chen[†] Tat-Jun Chin[†] Yee Wei Law[‡] Michele Sasdelli[†]
Ramesh Rajasegaran[†] Dillon Campbell[†]
[†]The University of Adelaide [‡]University of South Australia

Abstract

Deep neural networks (DNNs) have become essential for processing the vast amounts of aerial imagery collected using earth-observing satellite platforms. However, DNNs are vulnerable towards adversarial examples, and it is expected that this weakness also plagues DNNs for aerial imagery. In this work, we demonstrate one of the first efforts on physical adversarial attacks on aerial imagery, whereby adversarial patches were optimised, fabricated and installed on or near target objects (cars) to significantly reduce the efficacy of an object detector applied on overhead images. Physical adversarial attacks on aerial images, particularly those captured from satellite platforms, are challenged by atmospheric factors (lighting, weather, seasons) and the distance between the observer and target. To investigate the effects of these challenges, we devised novel experiments and metrics to evaluate the efficacy of physical adversarial attacks against object detectors in aerial scenes. Our results indicate the palpable threat posed by physical adversarial attacks towards DNNs for processing satellite imagery[1].

1. Introduction

The amount of images collected from earth-observing satellite platforms is growing rapidly [69], in part fuelled by the dependency of many valuable applications on the availability of satellite imagery obtained in high cadence over large geographical areas, *e.g.*, environmental monitoring [32], urban planning [2], economic forecasting [64]. This naturally creates a need to automate the analysis of satellite or aerial imagery to cost effectively extract meaningful insights from the data. To this end, deep neural networks (DNNs) have proven effective [2, 64, 69], particularly for tasks such as image classification [2, 40], object detection [21, 47], and semantic segmentation [66].

However, the vulnerability of DNNs towards *adversarial examples*, *i.e.*, carefully crafted inputs aimed at fooling the models into making incorrect predictions, is well documented. Szegedy *et al*. [53] first showed that an input image perturbed with changes that are imperceptible to the human eye is capable of biasing convolutional neural networks (CNNs) to produce wrong labels with high confidence. Since then, numerous methods for generating adversarial examples [4, 7, 15, 25, 26, 31, 33, 34, 36, 37, 38, 51] and defending against adversarial attacks [6, 15, 16, 39, 56, 61] have been proposed. The important defence method of adversarial training [15, 22, 26, 31] requires generating adversarial examples during training. Hence, establishing effective adversarial attacks is a crucial part of defence.

Adversarial attacks can be broadly categorised into *digital attacks* and *physical attacks*. Digital attacks directly manipulate the pixel values of the input images [1, 65], which presume full access to the images. Hence the utility of digital attacks is mainly as generators of adversarial examples for adversarial training. Physical attacks insert real-world objects into the environment that, when imaged together with the targeted scene element, can bias DNN inference [3, 5, 13, 46, 55]. The real-world objects are typically image patches whose patterns are optimised in the digital domain before being printed. A representative work is [55] who optimised patches to fool a person detector [41].

Our work focusses on physical adversarial attacks on aerial imagery, particularly earth-observing images acquired from satellite platforms. Previous works on physical attacks [3, 5, 13, 17, 24, 27, 46, 55, 57, 58, 59, 60, 62] overwhelmingly focussed on ground-based settings and applications, *e.g.*, facial recognition, person detection and autonomous driving. A few exceptions include Czaja *et al*. [9], who targeted aerial image classification, and den Hollander *et al*. [10], who targeted aerial object detection. However, [9, 10] did not demonstrate their physical attacks in the real world, *i.e.*, their patches were not printed and imaged from an aerial platform. Also, although [9] mentioned the potential effects of atmospheric factors on physical attacks, no significant investigation on this issue was reported.

Contributions In this applications paper, our contributions are threefold:

- We report one of the first demonstrations of real-world physical adversarial attacks in aerial scenes, specifically

[1]See demo video at `https://youtu.be/5N6JDZf3pLQ`.

against a car detector. Adversarial patches were trained, printed and imaged from overhead using a UAV and from the balcony of a tall building; see Fig. 1. The captured images were processed using a car detector and the effectiveness of the physical attacks was then evaluated.

- We propose a novel adversarial patch design that surrounds the target object (car) with optimised intensity patterns (see Fig. 1b), which we also physically tested and evaluated. The design enables a more convenient attack since modifications to the car can be avoided, and the car can be driven directly into pattern on the ground.
- We examined the efficacy of physical attacks under different atmospheric factors (lighting, weather, seasons) that affect satellite imagery. To enable scalable evaluation, we designed an experimental protocol that performed data augmentation on aerial images and devised new, rigorous metrics to measure attack efficacy.

Our results indicate the realistic threat of physical attacks on aerial imagery object detectors. By making our code and data publicly available [12], we hope our work will spur more research into adversarial robustness of machine learning techniques for aerial imagery.

2. Related work

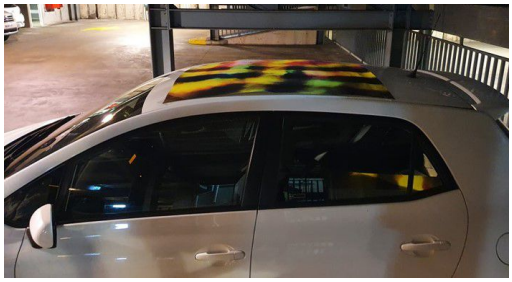
Three relevant areas of the literature on physical attacks are surveyed in this section.

2.1. Physical attacks on image classification

Kurakin *et al*. [25] generated the first physical-world attack by printing digitally perturbed images which were then captured by a smart phone and fed into an pre-trained Inception v3 [52] classifier. However, their results show that the effectiveness of the attack decreases when the images undergo the printing and photography processes. Lu *et al*. [30] confirmed this loss of effectiveness when the images were viewed at different angles and distances.

Since then, a growing number of research papers have started proposing methods to generate adversarial examples that can survive in the physical world. For instance, Sharif *et al*. [46] generated adversarial eyeglass frames to fool multiple facial recognition systems by adding a *non-printability score* (NPS) and *total variation* (TV) loss into their optimisation goal to ensure the colours used can be realised by a printer and the colour changes associated with adversarial perturbations are smooth. Komkov and Petiushko [24] added the TV loss into their optimisation goal to generate adversarial stickers (that were placed on hats) to fool a facial recognition system called ArcFace [11].

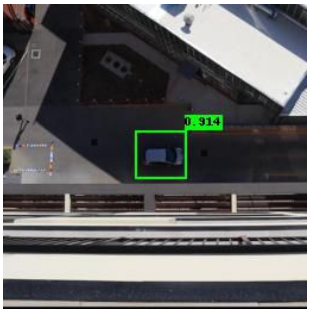
Athalye *et al*. [3] proposed a method called *Expectation Over Transformation* (EOT) which adds a transformation step into the optimisation process of generating adversarial perturbations. Their method was able to generate 3D-printed adversarial objects that were robust over a chosen distribution of synthetic transformations such as scale, ro-



(a) Adversarial patch on the roof of a car.

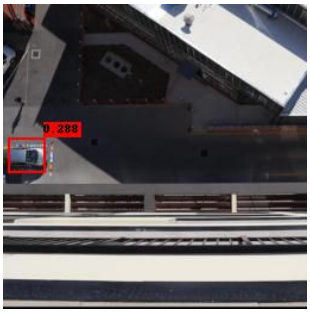
(b) Adversarial patch off-and-around a car.

(c) Cars in car park imaged without adversarial patches (objectness scores ≥ 0.9).



(d) A car in side street imaged without adversarial patch (objectness score 0.914).

(e) Two of the cars from (c) imaged with physical on-car patches; their objectness scores reduced to 0.315 and 0.396.



(f) The car in (d) imaged with a physical off-car patch; the objectness score reduced to 0.288.

Figure 1: Physical adversarial attacks on a YOLOv3 car detector in aerial imagery.

tation, translation, contrast, brightness, and random noise. These patches were shown to be universal, *i.e*., can be used to attack any scene; robust, *i.e*., effective under a variety of transformations; and targeted, *i.e*., can cause a classifier to output any target class; as demonstrated in the "rogue" traffic sign scenarios [48, 49], and even scenarios where the patch is adjacent to rather than on the targeted object [5]. Eykholt *et al*. [13] proposed a refinement of the EOT method called *Robust Physical Perturbation* (RP2), which performs sampling over a distribution of synthetic and physical transformations to generate adversarial stop signs in the form of posters or black-and-white stickers to be placed on stop signs. However, in order to use the RP2 method, the attacker would need to print out the original (clean) image and then, take multiple photos of it from different angles and distances to generate a single adversarial example. Jan *et al*. [20] proposed a transformation called D2P, to be performed prior to EOT, which uses a cGAN [19, 68] to simulate the effects produced by the printing and photography process. Their method also suffers from a feasibility problem since the attacker would need to print out hundreds of images and then capture them with a camera to build the ground truth to train the network.

2.2. Physical attacks on object detection

The building blocks of adversarial patch generation discussed in the previous subsection, ranging from NPS minimisation to cGAN, are also applicable to physical attacks on object detection, *e.g*., detection of traffic signs [14, 29, 8, 50], which is of practical importance considering the emergence of autonomous driving. These early work targeted YOLOv2/YOLO9000 [42] and Faster R-CNN [44] object detectors. Adversarial examples trained using YOLOv2 do not transfer well to (i.e., is not as effective against) Faster R-CNN, and vice versa, but adversarial examples trained using either YOLOv2 or Raster R-CNN transfer well to single-shot detectors [63].

Thys *et al*. [55] generated adversarial cardboard patches to hide people from a YOLOv2-based detector, but attacks targeting YOLOv2 do not transfer well to YOLOv3 [43] because YOLOv3 supports multi-label prediction and can detect small objects better [58]. From [55], several directions emerged: (i) off-body patches [27]; (ii) physical adversarial examples that are robust to physical-world transformations, taking into account non-rigid deformations of patches [60], viewing angle, wrinkling, stereoscopic radians, occlusion [62], material constraint, semantic constraint [17], *etc*.; (iii) targeting more advanced detectors than YOLOv2 such as YOLOv3 [27, 59], and multiple detectors simultaneously using ensemble training [59].

Cars have been the targeted objects recently [67, 57], but in these prior work, simulators based on the Unreal Engine and at most toy cars were used. Moreover, except for unpaintable surfaces like windows, the entire car was subjected to adversarial patching, whereas in our case, adversarial patches are subject to size constraints.

2.3. Physical attacks in aerial imagery

Physical attacks have mostly been demonstrated on earth-based imagery captured at relatively small distances *e.g*., within the sensing range of a camera on an autonomous car or a security system. Physical attacks against aerial or satellite imagery have not been considered or extensively studied. A few exceptions are Czaja *et al*. [9] and den Hollander *et al*. [10], who generated adversarial patches for aerial imagery classification and detection respectively. However, [9, 10] only evaluated their attacks digitally, *i.e*., they did not print their patches and deploy them in the real world. Robustifying physical attacks againts atmospheric effects, temporal variability and material properties was mentioned but no significant work was reported in [9].

3. Threat model

We first present the threat model used in our work.

Attacker's goals: The attacker aims to generate adversarial patches that can prevent cars existing in a selected locality, called the **scene of attack**, from being detected from the air by a pre-trained car detector. These patches are to be placed on the roof of cars, or off-and-around cars; see Figs. 1a and 1b. Although the attacker is only interested in hiding cars in the scene of attack, the patches should still be effective in different environmental conditions.

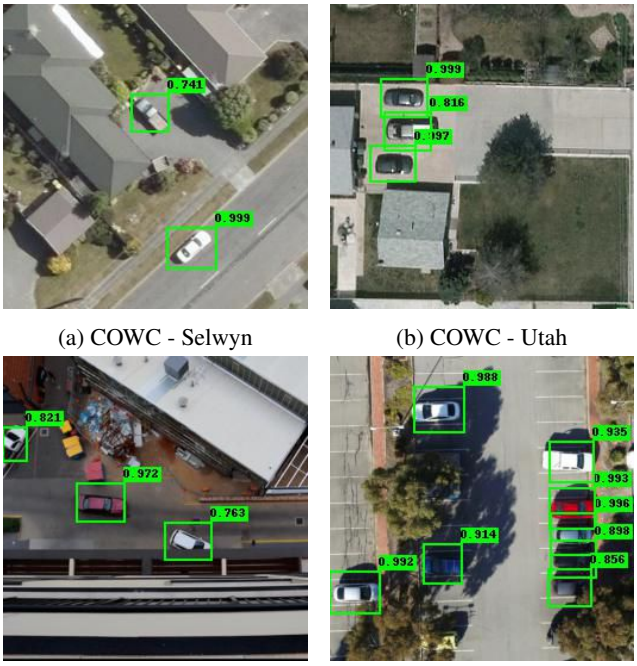
Attacker's knowledge: The attacker is assumed to have *white-box* (*i.e*., complete) knowledge of the detector model including its parameter values, architecture, loss function, optimiser, and in some cases its training data as well [54, p. 22]. Examples of scenarios where this assumption is valid include when (i) the targeted detector is known to be taken from some open-source implementation, (ii) the attacker can access and reverse-engineer a black-box detector implementation. More importantly, this assumption represents the worst-case scenario for the defender, which allows us to assess the maximum impact the attacker can cause.

Attacker's capabilities: The attacker can optimise and evaluate the adversarial patches in the digital domain, and perform physical-world attacks by printing the patches and placing them physically in the scene of attack.

Attacker's strategy: The attacker will attempt to generate adversarial patches by solving an optimisation problem that includes the maximum objectness score in the loss function. See Sec. 4.3 for more details.

4. Methods

Based on the threat model in Sec. 3, this section discusses the optimisation of physical adversarial patches to



(a) COWC - Selwyn (b) COWC - Utah

(c) Side Street (d) Car Park

Figure 2: Overhead view of cars from three datasets.

attack car detectors in aerial imagery.

4.1. Object detector model

To build a car detector for aerial imagery, we performed transfer learning on YOLOv3 [43] (pretrained on MS-COCO [28]) on the Cars Overhead with Context (COWC) dataset [35]. We used a version of COWC called COWC-M which contained $25,384$ colour images (256×256 pixels) of annotated cars in an overhead view (Figs. 2a and 2b). We separated the data into $20,306$ training and $5,078$ testing samples. The mean Average Precision (mAP) measured at IOU threshold of 0.50 of our detector was 0.50, which was on par with YOLOv3 trained on MS-COCO.

4.2. Scenes of attack and data collection

Following the threat model in Sec. 3, we selected and collected data from two scenes of attack:

Side Street Street viewed from the 10th floor (40 m height) of a building (see Fig. 2c). Data for training adversarial patches were captured with a Canon EOS M50 camera with a 15-45 mm lens. A total of 843 images were captured and divided into 780 training and 63 testing images.

Car Park Parking area with approximately 400 parking spaces (see Fig. 2d). Data for training adversarial patches were captured with a DJI Zenmuse X7 camera with a 24 mm lens on a DJI Matrice 200 Series V2 UAV at a fixed height of 60 m. A total of 565 images were captured and separated into 500 training and 65 testing images.

The data collected from the selected scenes sufficiently resembles the COWC data and the pre-trained car detector worked successfully on the collected images; see Fig. 2.

Annotation for patch optimisation Let $\mathcal{U} = \{I_i\}_{i=1}^{M}$ and $\mathcal{V} = \{J_j\}_{j=1}^{N}$ respectively be the M training and N testing images for a particular scene of attack. We executed the YOLOv3 car detector (Sec. 4.1) on $\mathcal{U}$ and $\mathcal{V}$ with an objectness threshold of 0.5 and non-max suppression IoU threshold of 0.4. This yields the tuples

$$\mathcal{T}_i^{\mathcal{U}} = \{(\mathcal{B}_{i,k}^{\mathcal{U}}, s_{i,k}^{\mathcal{U}})\}_{k=1}^{D_i}, \quad \mathcal{T}_j^{\mathcal{V}} = \{(\mathcal{B}_{j,\ell}^{\mathcal{V}}, s_{j,\ell}^{\mathcal{V}})\}_{\ell=1}^{E_j} \tag{1}$$

for each I_i and J_j, where

- D_i is the number of detections in I_i;
- $\mathcal{B}_{i,k}^{\mathcal{U}}$ is the bounding box of the k-th detection in I_i;
- $s_{i,k}^{\mathcal{U}}$ is the objectness score of the k-th detection in I_i

(similarly for E_j, $\mathcal{B}_{j,\ell}^{\mathcal{V}}$ and $s_{j,\ell}^{\mathcal{V}}$ for J_j). Note that by design $s_{i,k}^{\mathcal{U}} \geq 0.5$ and $s_{j,\ell}^{\mathcal{V}} \geq 0.5$. The sets of all detections are

$$\mathcal{T}^{\mathcal{U}} = \{\mathcal{T}_i^{\mathcal{U}}\}_{i=1}^{M}, \quad \mathcal{T}^{\mathcal{V}} = \{\mathcal{T}_j^{\mathcal{V}}\}_{j=1}^{N}, \tag{2}$$

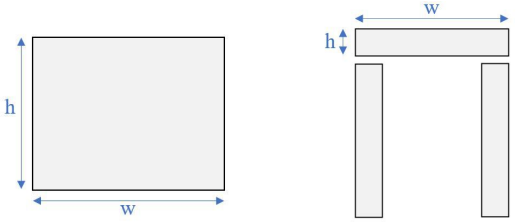
which we manually checked to remove false positives.

4.3. Optimising adversarial patches

We adapted Thys *et al.*'s [55] method for adversarial patch optimisation for aerial scenes; see Fig. 5 for our pipeline. Two patch designs were considered (see Fig. 3):

- Type ON: rectangular patch to be installed on car roof. The following dimensions were used:
  - Digital dimensions: $w = 200$ pixels, $h = 160$ pixels
  - Physical dimensions: $w = 1189$ mm, $h = 841$ mm
- Type OFF: three rectangular strips to be installed off and around car, forming a '$\sqcap$' shape, with dimensions:
  - Digital dimensions: $w = 400$ pixels, $h = 25$ pixels
  - Physical dimensions: $w = 3200$ mm, $h = 200$ mm

See Fig. 1 for the physical placement of the patches.



(a) Type ON. (b) Type OFF.

Figure 3: Patch designs.

Let P be the set of colour pixels that define a digital patch (ON or OFF). The data used to optimise P for a scene of attack are the training images $\mathcal{U}$ and annotations $\mathcal{T}^{\mathcal{U}}$ for

the scene. The values of P are initialised randomly. Given the current P, the patch is embedded into each training image I_i based on the bounding boxes $\{\mathcal{B}^{\mathcal{U}}_{i,k}\}^{D_i}_{k=1}$.

Several geometric and colour-space augmentations are applied to P to simulate its appearance when captured in the field. The geometric transformations applied are:

- Random scaling of the embedded P to a size that is roughly equivalent to the physical size of P in the scene.
- Random rotations ($\pm 20°$) on the embedded P about the centre of the bounding boxes $\{\mathcal{B}^{\mathcal{U}}_{i,k}\}^{D_i}_{k=1}$.

The above simulate placement and printing size uncertainties. The colour space transformations are

- Random brightness adjustment (± 0.1 change of pixel intensity values).
- Random contrast adjustment ($[0.8, 1.2]$ change of pixel intensity values).
- Random noise (± 0.1 change of pixel intensity values).

The above simulate (lack of) colour fidelity of the printing device and lighting conditions in the field.

Let $\tilde{I}_i$ be I_i embedded with P following the steps above. Since the patches are situated outdoors, we also considered weather and seasonal changes. The Automold tool [45] was applied on each $\tilde{I}_i$ to adjust the sun brightness and add weather and seasonal effects (snow, rain, fog, autumn leaves). See Fig. 4 for the augmentations applied. Ablation studies on the augmentations will be presented in Sec. 6.

(a) No geometric and colour-space transformations.

(b) With geometric and colour-space transformations.

(c) No weather transformation.

(d) Synthetic rain added.

Figure 4: Effects of augmentations applied in our pipeline.

The augmented image $\tilde{I}_i$ is forward propagated through the car detector (Sec. 4.1). The output consists of three grids which correspond to the three prediction scales of YOLOv3. Each cell in the output grids contains three bounding box predictions with objectness scores. Let $\tilde{\mathcal{S}}^{\mathcal{U}}_i$ be the set of all predicted objectness scores for $\tilde{I}_i$. Since we aim to prevent cars from being detected, we define the loss for $\tilde{I}_i$ as

$$L_i(P) = \max\left(\tilde{\mathcal{S}}^{\mathcal{U}}_i\right) + \delta \cdot NPS(P) + \gamma \cdot TV(P), \quad (3)$$

where $\max(\tilde{\mathcal{S}}^{\mathcal{U}}_i)$ is the maximum predicted objectness score in $\tilde{I}_i$, and δ, γ are weights for the non-printability score [46] and total variation [46] of P respectively. Specifically,

$$NPS(P) = \sum_{u,v}\left(\min_{c\in C}\|\mathbf{p}_{u,v} - \mathbf{c}\|_2\right), \quad (4)$$

where $\mathbf{p}_{u,v}$ is the pixel (RGB vector) at (u, v) in P, and $\mathbf{c}$ is a colour vector from the set of printable colours [55]. The NPS term encourages colours in P to be as close as possible to colours that can be reproduced by a printing device. The TV term

$$TV(P) = \sum_{t,u,v}\sqrt{(p_{t,u,v} - p_{t,u+1,v})^2 + (p_{t,u,v} - p_{t,u,v+1})^2}$$

encourages P's that are smooth, where $p_{t,u,v}$ is the t-channel pixel (scalar) at (u, v) in P, which also contributes to the physical realisability of P. Minimising L to optimise P is achieved using the Adam [23] stochastic optimisation algorithm. Note that the pre-trained detector is not updated.

5. Measuring efficacy of physical attacks

The efficacy of the patch P^* optimised according to Sec. 4 is to be evaluated, but previous metrics are not directly or intuitively interpretable here. For example, *average precision* [10, 55, 59] measures the classification accuracy of object detectors, whereas *attack success rate* [9, 13, 59, 60] focusses on misclassification. While these metrics can be applied in our experiments[2], we propose new metrics that directly measure the impact on objectness score, developed according to two evaluation regimes.

5.1. Evaluation in the digital domain

As a sanity test, as well as a more scalable approach to perform ablation studies, we evaluate P^* using the testing set $\mathcal{V}$ from Sec. 4.2. In addition to the detection results $\mathcal{T}^{\mathcal{V}}$ on $\mathcal{V}$, we embed P^* into each $J_j \in \mathcal{V}$ (following the same augmentation steps in Sec. 4.3) to yield $\tilde{J}_j$. Executing the car detector (Sec. 4.1) on $\tilde{J}_j$ yields the tuple

$$\tilde{\mathcal{T}}^{\mathcal{V}}_j = \{(\tilde{\mathcal{B}}^{\mathcal{V}}_{j,\ell}, \tilde{s}^{\mathcal{V}}_{j,\ell})\}^{E_j}_{\ell=1}. \quad (5)$$

A one-to-one matching between $\mathcal{T}^{\mathcal{V}}_j$ and $\tilde{\mathcal{T}}^{\mathcal{V}}_j$ is achieved by using an objectness threshold close to zero, and associating for each $\mathcal{B}^{\mathcal{V}}_{j,\ell}$ the resulting bounding box with the highest objectness score that overlaps sufficiently with $\mathcal{B}^{\mathcal{V}}_{j,\ell}$. Let

$$\tilde{\mathcal{T}}^{\mathcal{V}} = \{\tilde{\mathcal{T}}^{\mathcal{V}}_j\}^{N}_{j=1}. \quad (6)$$

[2]See supplementary material for our average precision results.

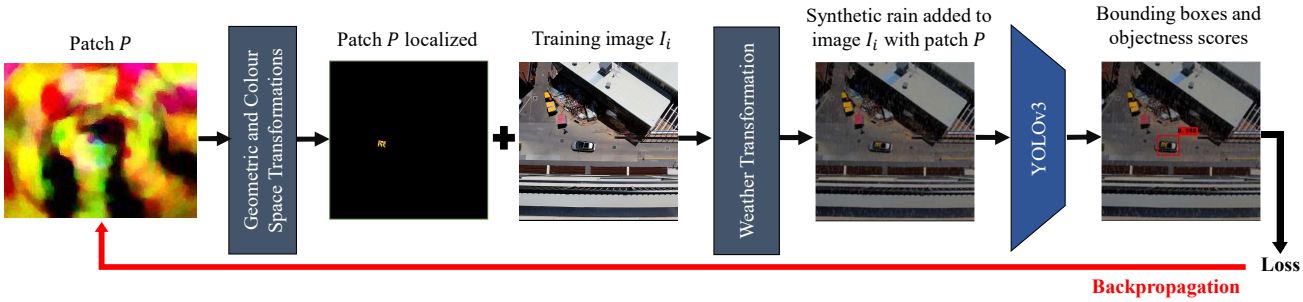


Figure 5: Optimisation process for generating adversarial patches.

Define *average objectness reduction rate (AORR)* as

$$\text{AORR}(\mathcal{T}^{\mathcal{V}}, \tilde{\mathcal{T}}^{\mathcal{V}}) \triangleq \frac{1}{NE_j} \left(\sum_{j=1}^{N} \sum_{\ell=1}^{E_j} \frac{s_{j,\ell}^{\mathcal{V}} - \tilde{s}_{j,\ell}^{\mathcal{V}}}{s_{j,\ell}^{\mathcal{V}}} \right). \quad (7)$$

Intuitively, AORR measures the average reduction in objectness score of an object due to the adversarial patch. A more effective attack will yield a higher AORR.

5.2. Evaluation in the physical domain

To evaluate P^* in the physical domain, we collect new image sets (e.g., videos) $\mathcal{F} = \{F_f\}_{f=1}^{\alpha}$ and $\mathcal{G} = \{G_g\}_{g=1}^{\beta}$ from the same scene of attack, with $\mathcal{F}$ containing the physical realisations of P^* installed on several target cars, while $\mathcal{G}$ does not contain any adversarial patches.

For each car that was targeted in $\mathcal{F}$, the car was tracked using the Computer Vision Annotation Tool (CVAT) [18] in both $\mathcal{F}$ and $\mathcal{G}$. Both sets $\mathcal{F}$ and $\mathcal{G}$ were then subject to the car detector (Sec. 4.1), and the objectness scores corresponding to the targeted object

$$\tilde{\mathcal{S}} = \{\tilde{s}_f\}_{f=1}^{\alpha}, \quad \mathcal{S} = \{s_g\}_{g=1}^{\beta} \quad (8)$$

were retrieved (similarly to Sec. 5.1). Define the *objectness score ratio (OSR)* of the targeted car as

$$\text{OSR}(\tilde{\mathcal{S}}, \mathcal{S}) \triangleq \frac{\frac{1}{\alpha}\sum_{f=1}^{\alpha} \tilde{s}_f}{\frac{1}{\beta}\sum_{g=1}^{\beta} s_g}. \quad (9)$$

Intuitively, OSR measures the ratio of the objectness scores after and before the application of physical patch P^*. A more effective attack will yield a lower OSR.

Define the *normalised detection rate (NDR)* of the targeted car as

$$\text{NDR}_{\tau}(\tilde{\mathcal{S}}, \mathcal{S}) \triangleq \frac{\frac{1}{\alpha}\sum_{f=1}^{\alpha} \mathbb{I}(\tilde{s}_f \geq \tau)}{\frac{1}{\beta}\sum_{g=1}^{\beta} \mathbb{I}(s_g \geq \tau)}, \quad (10)$$

where τ is a given objectness threshold, and $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the input predicate is true and 0 otherwise. Intuitively, NDR measures the proportion of frames where the object is detected, hence a more effective attack will yield a lower NDR.

6. Results

We first perform ablation studies on the augmentations in Sec. 4.3 by evaluating the efficacy of P^* in the digital domain (see Sec. 5.1). Then we evaluate the efficacy of P^* in the physical domain (see Sec. 5.2).

6.1. Results for digital domain evaluation

Based on the collected data (Side Street and Car Park; see Sec. 4.2), we optimised adversarial patches under different variants of the pipeline (Sec. 4.3):

- G/C+W: geometric, colour-space and weather augmentations were applied (full pipeline).
- G/C: only geometric and colour-space augmentations.
- Control: patches were random intensity patterns, *i.e.*, the optimisation pipeline was completely bypassed.

The weights for the terms in (3) was $\delta = 0.01$ and $\gamma = 2.5$. Type OFF patch was not optimised for Car Park, due to the close proximity of the cars parked in the scene which prevented the off car patch from being embedded without occluding cars. Fig. 6 depicts the resulting patches. Note that G/C+W patches appear dimmer than G/C patches, suggesting that the optimisation is accounting for changes in scene appearance due to sun brightness and weather (the practical usefulness of this will be discussed below).

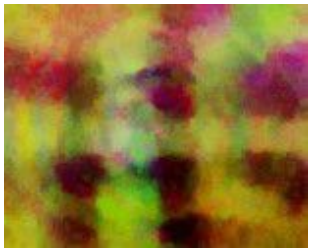
Two variants of the testing regime (Sec. 5.1) were used:

- STD: following Sec. 5.1 exactly.
- STD-W: the testing images $\mathcal{V}$ were augmented with weather effects using the same steps in Sec. 4.3.

Fig. 7 illustrates sample qualitative results from our digital domain evaluation, while Table 1 shows quantitative results.

Qualitative results suggest that while both weather effects alone and Control patches can reduce the objectness scores, optimised patches are more successful. This is confirmed by the AORR values in Table 1. While both G/C and G/C+W were significantly more effective than Control, G/C visibly outperforms G/C+W, which indicates the lack of value in performing weather augmentations during training; this finding also motivated us to ignore G/C+W for optimising Type OFF patches for Side Street. However, the

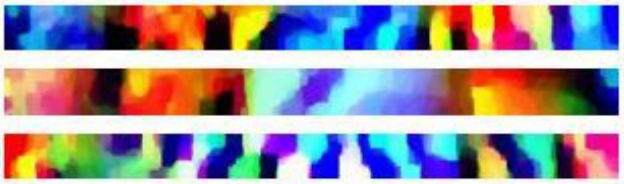
(a) Type ON G/C+W optimised on Car Park.



(b) Type ON G/C+W optimised on Side Street.

(c) Type ON G/C optimised on Car Park.

(d) Type ON G/C optimised on Side Street.



(e) Type OFF G/C optimised on Side Street.

Figure 6: Optimised patches P^* in our experiments.

Training pipeline	Patch type	AORR (STD)	AORR (STD+W)
G/C+W	ON	0.602	0.558
G/C	ON	0.712	0.615
Control	ON	0.290	0.303
G/C+W	OFF	-	-
G/C	OFF	0.814	0.706
Control	OFF	0.220	0.291

(a) Side Street.

Training pipeline	Patch type	AORR (STD)	AORR (STD+W)
G/C+W	ON	0.856	0.693
G/C	ON	0.878	0.768
Control	ON	0.513	0.294

(b) Car Park.

Table 1: Efficacy of adversarial patches on Side Street and Car Park scenes of attack under digital domain evaluation (higher AORR implies more effective attack).

results show that, for the same scene (Side Street), Type OFF G/C patches outperformed Type ON G/C patches.

(a) No patches embedded in an original testing image (STD) and the weather (snow) augmented version of the testing image (STD-W).

(b) Embedded with Control patches (random intensity patterns).

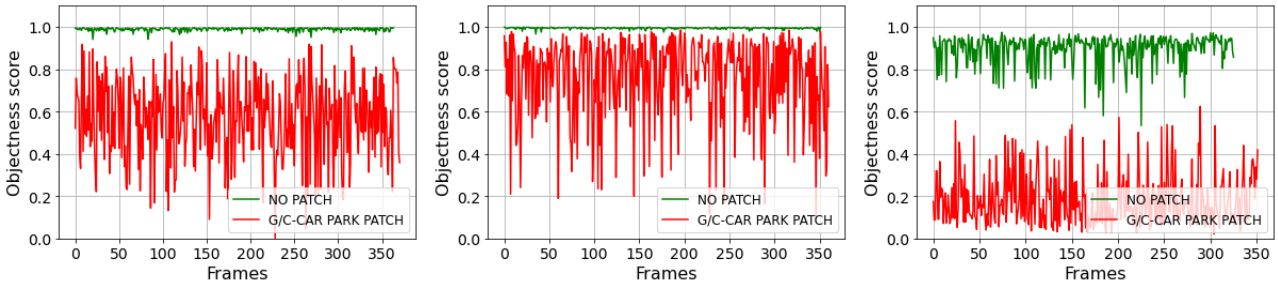
(c) Embedded with patches optimised with the G/C pipeline.

Figure 7: Qualitative results from digital domain evaluation in the Car Park scene. Green and red bounding boxes indicate objectness scores ≥ 0.5 and < 0.5 respectively.

6.2. Results for physical domain evaluation

Under the testing regime in Sec. 5.2, patches were printed on 160 gsm coated paper, 25 FPS videos were captured to form the new testing images $\mathcal{F}$ and $\mathcal{G}$ for both Car Park and Side Street. When capturing $\mathcal{F}$, three cars under our control ("Grey", "White", "Blue") were installed with the physical realisations of the optimised patches P^*. Basic statistics of the data are as follows:

- Car Park: $|\mathcal{F}| = 1,084$ frames, $|\mathcal{G}| = 1,042$.
- Side Street: $|\mathcal{F}| = 4,699$ frames, $|\mathcal{G}| = 526$.



(a) (b) (c)

Figure 8: Objectness scores for Car Park physical domain testing sets $\mathcal{F}$ (with Type ON optimised patches P^* installed in the cars) and $\mathcal{G}$, plotted according to frame index, for the targeted (a) Grey, (b) White and (c) Blue cars.

We further categorise the images into different settings:

- Lighting: whether the car was in sunlight or shade.
- Motion: whether the car was static or moving wrt ground.

Due to the cost of performing physical experiments (*e.g.*, civil aviation approval, availability of personnel) and uncontrollable environmental factors, not all the combinations above were explored. Also, only patches optimised under the G/C setting were used based on the findings in Sec. 4.2.

Qualitative results are illustrated in Fig. 1. Fig. 8 shows the objectness scores in $\mathcal{F}$ and $\mathcal{G}$ for the Car Park scene, which highlights differences in attack efficacy for the different cars. In particular, the attack was much more successful for Blue, even accounting for a visibly lower objectness score prior to the attack. Quantitative results (mean OSR; lower means more successful attack) in Table 2 illustrate the general effectiveness of the physical attack, *i.e.*, significant reductions in objectness scores (25% to 85%) were achievable (depending on the car and environmental factors). While Type OFF patches are successful in reducing objectness scores, in contrast to the digital evaluation they are less effective than Type ON patches. Fig. 9 plots mean NDR as a function of objectness threshold τ (lower NDR means higher attack effectiveness) for the three cars, which also illustrates differences in attack efficacy.

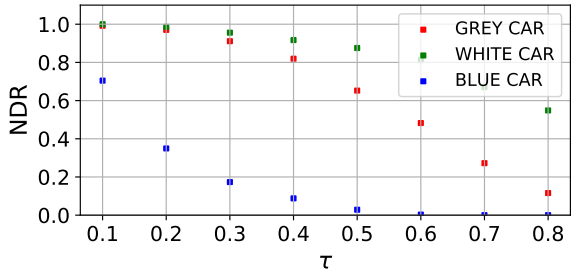


Figure 9: Mean NDR as a function of objectness threshold τ for Car Park and Side Street (Type ON patches only).

Car	Patch type	Lighting	Motion	Mean OSR
Grey	ON	Both	Moving	0.343
Grey	ON	Sun	Static	0.251
Grey	ON	Shade	Static	0.255
Grey	OFF	Sun	Static	0.429
White	ON	Both	Moving	0.286
White	ON	Sun	Static	0.285
White	ON	Shade	Static	0.197
White	OFF	Sun	Static	0.748

(a) Side Street.

Car	Patch type	Lighting	Motion	Mean OSR
Grey	ON	Sun	Static	0.509
Blue	ON	Shade	Static	0.208
White	ON	Sun	Static	0.746

(b) Car Park.

Table 2: Mean OSR in Car Park and Side Street.

7. Conclusions

We demonstrated physical adversarial attacks on a car detector in aerial scenes, and proposed a novel "off car" patch design which was shown to be effective. Our results indicate that, while the efficacy of the attack is subject to atmospheric factors (lighting, weather, seasons) and the target-observer distance, physical adversarial attacks are a realistic threat. Curiously, our ablation tests showed that augmenting patch optimisation with weather effects did not result in higher effectiveness, even in the digital domain. This will form a useful topic for future investigations.

Acknowledgements

Tat-Jun Chin is SmartSat CRC Professorial Chair of Sentient Satellites.

References

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

[2] Adrian Albert, Jasleen Kaur, and Marta C. Gonzalez. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1357–1366, New York, NY, USA, 2017. Association for Computing Machinery.

[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

[4] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations*, 2018.

[5] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[6] Nicholas Carlini and David Wagner. MagNet and "efficient defenses against adversarial attacks" are not robust to adversarial examples. *arXiv preprint arXiv:1711.08478*, 2017.

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

[8] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng (Polo) Chau. ShapeShifter: Robust physical adversarial attack on Faster R-CNN object detector. In Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley, and Georgiana Ifrim, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 52–68, Cham, 2019. Springer International Publishing.

[9] Wojciech Czaja, Neil Fendley, Michael Pekala, Christopher Ratto, and I-Jeng Wang. Adversarial examples in remote sensing. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '18, page 408–411, New York, NY, USA, 2018. Association for Computing Machinery.

[10] Richard den Hollander, Ajaya Adhikari, Ioannis Tolios, Michael van Bekkum, Anneloes Bal, Stijn Hendriks, Maarten Kruithof, Dennis Gross, Nils Jansen, Guillermo Perez, Kit Buurman, and Stephan Raaijmakers. Adversarial patch camouflage against aerial detection. In Judith Dijk, editor, *Artificial Intelligence and Machine Learning in Defense Applications II*, volume 11543, pages 77–86. International Society for Optics and Photonics, SPIE, 2020.

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.

[12] Andrew Du. Physical adversarial attacks on an aerial imagery object detector. GitHub repository at `https://github.com/andrewpatrickdu/adversarial-yolov3-cowc`, 2021. Accessed 11 Oct 2021.

[13] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on deep learning models. In *Computer Vision and Pattern Recognition*, 2018.

[14] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Dawn Song, Tadayoshi Kohno, Amir Rahmati, Atul Prakash, and Florian Tramer. Note on attacking object detectors with adversarial stickers. *arXiv preprint arXiv:1712.08062*, 2017.

[15] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[16] Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

[17] 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*, pages 720–729, 2020.

[18] Intel. Computer Vision Annotation Tool (CVAT). GitHub repository, 2021. v1.6.0, accessed 6 October 2021.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[20] 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, July 2019.

[21] Hong Ji, Zhi Gao, Tiancan Mei, and Bharath Ramesh. Vehicle detection in remote sensing images leveraging on simultaneous super-resolution. *IEEE Geoscience and Remote Sensing Letters*, 17(4):676–680, 2020.

[22] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] Stepan Komkov and Aleksandr Petiushko. AdvHat: Real-World Adversarial Attack on ArcFace Face ID System. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 819–826, 2021.

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

[26] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)*, 2017.

[27] Mark Lee and Zico Kolter. On physical adversarial patches for object detection. In *ICML 2019 Workshop on the Security and Privacy of Machine Learning*, 2019.

[28] 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*, pages 740–755. Springer, 2014.

[29] Jiajun Lu, Hussein Sibai, and Evan Fabry. Adversarial examples that fool detectors. *arXiv preprint arXiv:1712.02494*, 2017.

[30] Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsyth. No need to worry about adversarial examples in object detection in autonomous vehicles. *arXiv preprint arXiv:1707.03501*, 2017.

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

[32] Salvatore Manfreda, Matthew F. McCabe, Pauline E. Miller, Richard Lucas, Victor Pajuelo Madrigal, Giorgos Mallinis, Eyal Ben Dor, David Helman, Lyndon Estes, Giuseppe Ciraolo, Jana Müllerová, Flavia Tauro, M. Isabel De Lima, João L. M. P. De Lima, Antonino Maltese, Felix Frances, Kelly Caylor, Marko Kohv, Matthew Perks, Guiomar Ruiz-Pérez, Zhongbo Su, Giulia Vico, and Brigitta Toth. On the use of unmanned aerial systems for environmental monitoring. *Remote Sensing*, 10(4), 2018.

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

[34] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[35] T Nathan Mundhenk, Goran Konjevod, Wesam A Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *European Conference on Computer Vision*, pages 785–800. Springer, 2016.

[36] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[37] 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*, pages 506–519. ACM, 2017.

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

[39] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016.

[40] Mark Pritt and Gary Chern. Satellite image classification with deep learning. In *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7, 2017.

[41] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[42] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[43] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[44] 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*, pages 91–99, 2015.

[45] Ujjwal Saxena. Automold. Road augmentation library at https://github.com/UjjwalSaxena/Automold–Road-Augmentation-Library, 2018.

[46] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540. ACM, 2016.

[47] Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[48] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. Darts: Deceiving autonomous cars with toxic signs. *arXiv preprint arXiv:1802.06430*, 2018.

[49] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Prateek Mittal, and Mung Chiang. Rogue signs: Deceiving traffic sign recognition with malicious ads and logos. *arXiv preprint arXiv:1801.02780*, 2018.

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

[51] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 2019.

[52] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[53] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[54] Elham Tabassi, Kevin J. Burns, Michael Hadjimichael, Andres D. Molina-Markham, and Julian T. Sexton. A taxonomy

and terminology of adversarial machine learning. Draft NISTIR 8269, National Institution of Standards and Technology, 2019.

[55] Simen Thys, Wiebe Van Ranst, and Toon Goedeme. Fooling automated surveillance cameras: Adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[56] Beilun Wang, Ji Gao, and Yanjun Qi. A theoretical framework for robustness of (deep) classifiers against adversarial examples. *arXiv preprint arXiv:1612.00334*, 2016.

[57] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8565–8574, June 2021.

[58] Yajie Wang, Haoran Lv, Xiaohui Kuang, Gang Zhao, Yu-an Tan, Quanxin Zhang, and Jingjing Hu. Towards a physical-world adversarial patch for blinding object detection models. *Information Sciences*, 556:459–471, 2021.

[59] Zuxuan Wu, Ser-Nam Lim, Larry S. Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 1–17, Cham, 2020. Springer International Publishing.

[60] 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 *Computer Vision – ECCV 2020*, pages 665–681, Cham, 2020. Springer International Publishing.

[61] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Network and Distributed Systems Security Symposium (NDSS)*, 2018.

[62] Mingfu Xue, Can He, Zhiyu Wu, Jian Wang, Zhe Liu, and Weiqiang Liu. 3D invisible cloak. *arXiv preprint arXiv:2011.13705*, 2020.

[63] Mingfu Xue, Chengxiang Yuan, Can He, Jian Wang, and Weiqiang Liu. NaturalAE: Natural and robust physical adversarial examples for object detectors. *Journal of Information Security and Applications*, 57:102694, 2021.

[64] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11(1):2583, May 2020.

[65] X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2805–2824, 2019.

[66] Xiaohui Yuan, Jianfang Shi, and Lichuan Gu. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169:114417, 2021.

[67] 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.

[68] 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*, pages 2223–2232, 2017.

[69] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.