Building Towards “Invisible Cloak”: Robust Physical Adversarial Attack on YOLO Object Detector

Darren (Yu) Yang
GIX[1]
Tsinghua University, UW[2]
Beijing, China
yu-yang16@mails.tsinghua.edu.cn

Jay Xiong
GIX
University of Washangton
Bellevue,WA, U.S.
junfex@uw.edu

Xincheng Li
GIX
Tsinghua University, UW
Beijing, China
lixc16@mails.tsinghua.edu.cn

Xu Yan
GIX
Tsinghua University, UW
Beijing, China
yanx17@mails.tsinghua.edu.cn

John Raiti
Electrical Engineering, GIX
University of Washington
Seattle, WA, USA
jraiti@uw.edu

Yuntao Wang
GIX
Tsinghua University, UW
Beijing, China
yuntaowang@mail.tsinghua.edu.cn

HuaQiang Wu
DMN[3]
Tsinghua University
Beijing, China
wuhq@mail.tsinghua.edu.cn

Zhenyu Zhong[4]
X-Lab
Baidu USA
Sunnyvale, CA
edwardzhong@baidu.com

***Abstract*—Deep learning based object detection algorithms like R-CNN, SSD, YOLO have been applied to many scenarios, including video surveillance, autonomous vehicle, intelligent robotics et al. With more and more application and autonomy left to deep learning based artificial intelligence, humans want to ensure that the machine does the best for them under their control. However, deep learning algorithms are known to be vulnerable to carefully crafted input known as adversarial examples which makes it possible for an attacker to fool an AI system. In this work, we explored the mechanism behind the YOLO object detector and proposed an optimization method to craft adversarial examples to attack the YOLO model. The experiment shows that this white box attack method is effective and has a success rate of 100% in crafting digital adversarial examples to fool the YOLO model. We also proposed a robust physical adversarial sticker generation method based on an extended Expectation Over Transformation (EOT) method(a method to craft adversarial example in the physical world). We conduct experiments to find the most effective approach to generate adversarial stickers. We tested the stickers both digitally as a watermark and physically showing it on an electronic screen on the front surface of a person. Our result shows that the sticker attack as a watermark has a success rate of 90% and 45% on photos taken indoors and on random 318 pictures from ImageNet. Our physical attack also has a success rate of 72% on photos taken indoors. We shared our project source code on the Github and our work is reproducible.**

***Index Terms*—Artificial Intelligence, Security, Neural network, Adversarial Example, Physical Attack, Object Detector**

I. Introduction

Deep Learning provides tons of tools for computer vision systems based on Neural Network, which proved to be vulnerable to adversarial attacks [7]. With small perturbation to the original image, neural network based image classification system could totally malfunction. Traditional digital attack methods only focus on the generation of a picture-format adversarial attack example, but lack exploration on how to achieve them in the physical world. Physical Attack is the implementation of a digital attack in the physical world. For application scenario, physical attack turns out to be more worthy of study because it is physically more closer to real-world possible crime-scene. In addition, it also turns out to be difficult to achieve because different light conditions, angle of view and distance may affect the result of these crafted adversarial examples.

Comparing to image classification task, object detection task requires the detector to locate the possible object on a certain image first, then reports what it think the object is and the confidence of this decision. The model of object detection is more complex but has more application scenario such as self driving cars, video surveillance system et al. There are group of object detection algorithms such as different versions of R-CNN [6], SSD [10] and YOLO [13]. In general, the speed of YOLO object detector is faster than other models such as Faster R-CNN [14] and SSD so that YOLO model could be used more in real-time object detection tasks.

A. Our Contributions

In summary, we have two main contributions:

1) We explore the mechanism of the state-of-art object detector YOLO (You only look once) [13] and success-

[1] Global Innovation Exchange Institute
[2] University of Washington
[3] Department of Microelectronics and Nanoelectronics
[4] Corresponding Author

Acknowledgement: This work is supported by the National Key Research and Development Plan under Grant No. 2016YFB1001200, and the China Postdoctoral Science Foundation under grant No. 2017M610089.

fully generate the adversarial examples on it in white-box settings. We propose a novel optimization method to craft the adversarial examples for YOLO detector. The experiment shows that our optimization method is effective and has a success rate of 100% in crafting a digital adversarial examples attack the YOLO model.

2) Building upon the first contribution, we introduced a robust physical adversarial sticker generation method based on an extended Expectation Over Transformation (EOT) [1] method which is a method to craft adversarial example in the physical world. We conduct experiments to find the most effective approach to generate adversarial stickers. We tested the stickers both digitally as a watermark and physically showing it on an electronic screen on the front surface of a person. Our result shows that the sticker attack as a watermark has a success rate of 90% and 45% on photos taken indoors and on random 318 pictures from ImageNet. Our physical attack also has a success rate of 72% on photos taken indoors.

II. Related Work

There are lots of work already been done in the digital attack towards neural networks. They can be described in several dimensions of threaten model. Goodfellow et al. [7] introduced an algorithm called fast gradient sign method. Target model is a classifier with image input and possibility labels output. It assumes the parameters and training data of the target model are known, which is known as a white box model. Papernot et al. [12] extended the previous work by introducing an algorithm that can lead the misclassification toward a specific target. Though this method show the potential of adversarial attacks, its effects are limited due the assumption of white box is not practical in the real world. Papernot et al. [11] introduced an algorithm with the assumption of only having knowledge about the target models output label result, which can be used to attack the model hosted on a API such as Google Cloud and AWS. Liu et al. [9] explored the transferability over a large scale dataset and large models. They found a large part of targeted adversarial examples that are able to transfer the attack with their intended target labels even when the attacker does not know targets machine learning model, training data, training process and test labels. Carlini et al. [2] proposed the strongest white box attack method which is based on optimization over a selected loss function.

Besides that, physical attack is getting more and more attention.

Kurakin et al. [8] proposed the first physical attack by simply print out the digital attack pattern and the experiment show that it can also fool the neural networks deployed in the physical world using photos shot by cameras.

Athalye et al. [1] addressed the problem in physical world such as various angle, distance, viewpoint, and so on by proposing a framework called expectation over transformation (EOT). It is also the first paper that implement the real world 3D physical attack.

Eykholt et al. [4] introduced a general physical attack method, Robust Physical Perturbations (RP2), to synthesize adversarial perturbations under different physical conditions. They use masks when construct the stop sign adversarial examples. They did the experiment in three categories: Object-constrained Poster-Printing Attacks, Sticker Attacks and Drive-By Testing. The solid experiment verified the effectiveness of their model.

Chen et al. [3] designed a method to attack image-based objector detectors like Faster R-CNN. They showed how to leverage Expectation over Transformation technique to improve the robustness of adversarial examples in object-detection area. Their indoor experiment settings provided us a feasible method to take our photos around the physical examples we built and start our evaluation.

Sharif et al. [16] proposed an attack method based on adversarial generative network with a hand-made mask which could generate a glass on the photo of a people's face to cheat the face recognition system.

III. Problem Setup and Goal

To conduct this artificial intelligence safety related research, We try to attack Neural Network based Object Detector and figure out more about its mechanism and provide suggestions about building more robust artificial systems. In this work, we try to attack the state-of-art object detector, the YOLO model, which could detect objection task in real time. We want to make people disappear under the YOLO object detector and building towards the "invisible cloak".

IV. Attack Method

A. Background

1) Image Classifier Attack: Adversarial Attack starts from digital attacks on image classifier [7]. Image classifier takes in an image and output the classification of the object showing on the image, like a cat, dog, or any other pre-defined classes. Mathematically, if we donate $F(x)$ as the image classifier, taking a image x as input, it will output the probability of each class in the pre-defined classes categories. Machine Learning systems usually select the the class with the highest probability as the result of this input image. An adversarial attack is to construct a new image x' based on the original image x with little perturbation even not seen by human eyes and use it to fool the image classifier. If the image classifier takes the image x' as input, the output probability distribution is a lot different from the original one, especially the class with the highest probability will change, which means a perturbed cat image could be recognized as a dog.

2) Optimization based attack method: There are a lot of algorithms to craft adversarial examples from the original image, most of them are based on optimization method. A classical optimization method proposed by Carlini et al. [2] can be expressed as follows. If we donate $L_F(x, y) = L(F(x), y)$ as the loss function describing the difference between the model output and target label y, the total loss could be written as

$$L(x) = L(F(x), y) + \lambda \cdot \|x' - x\|_2^2 \tag{1}$$

where $L2-norm$ is used to measure the perturbation level, and λ is the punishment weight of the perturbation. Then the method to generate adversarial example x' is to minimize the total loss $L(x)$. In experiment, they found the attack result would be better if donating $x' = \frac{1}{2} \cdot (tanh(w)+1)$ and optimize w during the adversarial example generation.

3) Expectation over Transformation: Expectation over Transformation (EOT) proposed by [1] is a widely used method to construct adversarial examples in physical world. EOT does not optimize a single example, but samples a batch of transformation t over a distribution T, and it considers each image x as a transformation result with transformation t from the hidden model X. In the optimization process, EOT calculates the expectation of all transformed images, these images could include photos taking in different angles, distance and light conditions and those digital transformed images. Given a distance function $d(y', y)$, the expression of EOT-based model loss function could be expressed as

$$L_{F\_EOT}(x) = \mathbb{E}_{t\sim T}[d(t(x'), t(x))] \quad (2)$$

4) YOLO Object Detector: An object detector takes in an image and output the same image with detected object's bounding box around it, each bounding box will report a class in the pre-defined class categories. Traditional Object Detector such as R-CNN and its variations propose a interesting region and then use image classification network to report the class of the region. Redmon et al. [13] proposed YOLO, a method without repurposing classifiers when finish object detection tasks. They solve the object detection problem as a regression problem. A single neural network was built to predict bounding boxes and class probabilities from a image in one step, this makes it possible to detect object in real time.

B. Our Attack method

1) YOLO Object Detector Digital Attack: YOLO object detector [13] could detect 20 classes from a single image. It will split the image into $S * S$ grids, each grid will generate two bounding boxes where there is probably one of the target objects inside it. Every generated bounding box comes with a 20-dimension confidence C vector, representing the probability of each class. We specifically minimize the confidence of person C_{person} to conduct attack to each image. Our Digital Attack loss function could be written as

$$L(x) = \max_{i\in B} C_{i,person}(x) + \lambda \cdot \|tanh(x') - x\|_2 \quad (3)$$

In (3), B donates the set of all bounding boxes generated by YOLO model. We refer to C&W optimization method [2] to define our second item in equation (3) For the digital perturbations, we add a rectangular mask on each person. During our attack example generating process, we repeated update the pixels inside the mask to minimize the loss function. The steps to craft digital adversarial examples could be expressed as follows:

1) Label an image with people, drag a rectangular mask as the sticker area on a person in the image.
2) Optimization method takes the masked image as input, and output a image with perturbed mask.

2) Physical Adversarial Sticker Generation Method: We introduce an extended version of Expectation over Transformation Implementation while crafting our physical attack. Traditional Expectation over Transformation [1] implementations only transform, rotate and scale digital images in 2D space. We consider the transformation of our stickers in real 3D world while generating these digital transformations. We build up an pin hole camera model to simulate the stickers' transformation in the physical world. Our original sticker is a US-Letter sized paper, we sample our transformation by translating and rotating the photo with physical sticker each in three directions in the camera coordinate system. Then we get the projection of the photo with transformed stickers on the pin hole camera focal plane as an EOT digital transformed sample.

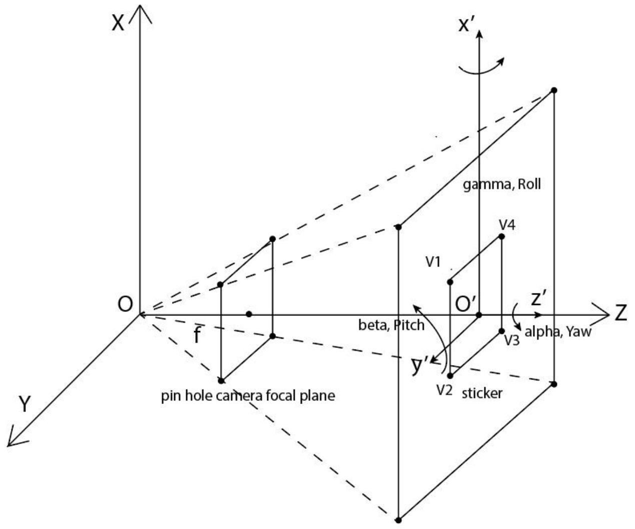


Fig. 1. Extended Expectation over Transformation Model

To make the perturbation sticker more smoothness, we add a total variation (TV) [15] between perturbed image and original image as a non-smoothness punishment, we use λ_2 as the weight of non-smoothness punishment. For an image x, TV is defined as

$$TV(x) = \sum_{i,j}((x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1}))^{1/2} \quad (4)$$

where $x_{i,j}$ donates the pixel value at the position i, j of the image x. To sum up, the loss of our physical sticker attack generation can be expressed as

$$\begin{aligned} L(x) = \mathbb{E}_{t\sim T, x\sim X} \max_{i\in B} C_{i,person}(t(x)) \\ + \lambda_1 \cdot \|tanh(x') - x\|_2) + \lambda_2 \cdot TV(x) \end{aligned} \quad (5)$$

The steps to craft physical adversarial examples could be expressed as follows:

1) Calibrate the pin hole camera model with a US-letter sized paper.

2) Update the parameter of transformation with calibrated parameter.
3) Label the image with people, drag a rectangular mask as the sticker area on/off a person in the image.
4) Optimization method takes the masked image as input, sample transformed image with pre-defined sample parameter, and output a image with perturbed mask.
5) Crop the output sticker from the generated image.
6) Use the generated sticker as watermark attack and physical “invisible cloak”.

V. Experiment

We use pre-trained *Tiny YOLO* model to conduct our experiment. We use an tensorflow-based Tiny YOLO implementation. Tiny YOLO is a small model of YOLO, it could run up to 244 FPS, much faster than original YOLO model.

A. *YOLO digital attack and evaluation*

We take 50 photos with a person standing in the middle of the camera, the scenario of these photos covers indoor and outdoor environment.The left column of Table I shows several examples of these photos, we feed these photos to the Tiny YOLO model, and we calculate our precision by

$$Precision = \frac{number\ of\ detection\ as\ people}{total\ number\ of\ image\ in\ test} \tag{6}$$

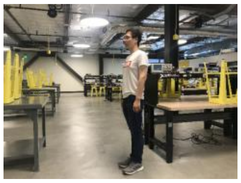
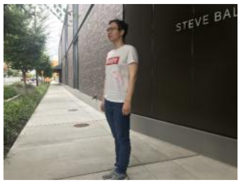
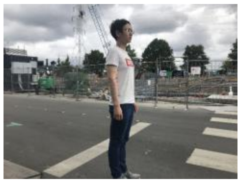
As expected, the result shows that Tiny YOLO could detect all these original photos correctly, with a precision rate 1.0. We construct our digital attack in this way. We label the person on an image, and drag a rectangular mask as the sticker area on a person in that image. Then we record the corner coordinate of the mask and feed the image with the mask coordinate to the optimization function. Using this method, we generate 50 perturbed photos and the right column of Table I shows several adversarial attack results and in this case the Tiny YOLO detection precision rate is 0.0.

B. *Physical adversarial sticker generation and evaluation*

In the physical adversarial sticker generation process, we use different mask areas: sticker area on person and sticker area not on person. These two experiment settings could help us figure out if the difference of the mask area could influence the effectiveness of adversarial sticker. To better evaluate the effectiveness of our extended Expectation over Transformation (EOT) model, We also use different settings in optimization with-EOT version stickers and without-EOT version stickers. In our EOT version sticker generation process, we donate $x' = \tanh(w_0 + \delta)$, and we minimize δ in our optimization formulation, and actual optimization process could be expressed as follows:

$$\begin{aligned} \arg\min_{\delta}(&\sum_{t \in T_0} \max_{i,j=1}^{S} \max_{k=1}^{B} C_{i,j,k,person}(t(\tanh(w_0 + \delta))) \\ &+ \lambda_1 \cdot \| \tanh(w_0 + \delta) - x_0 \|_2 \\ &+ \lambda_2 \cdot TV(\tanh(w_0 + \delta)) \end{aligned} \tag{7}$$

TABLE I
YOLO Digital Attack and Evaluation

	Original picture	**Ad attack result**
1		
2		
3		
	...	...
Precision	1.0	0.0

In the experiment settings, S in (7) means the number of the grids in YOLO model, we use $S = 7$. B in (7) means the number of bounding boxes in each grid, we use $B = 2$.

For our digital sampling method, we view our sticker as a rigid body and we rotate it around x, y and z axis, and we give roll, pitch, yaw each 21 small angles from $-\pi/60$ to $\pi/60$ with step $\pi/600$. In total we created 63 rotations of the image as our transformation set T_0. We use all these projection transformations with small angles as samples because we want to better simulate the camera’s slightly tilt when taking the photo. The later experiment will show our 3D transformation based EOT with small angles sampling method could generate more robust adversarial stickers attacking the Tiny YOLO model both digitally and physically.

With no-EOT version setting, we only use our previous YOLO digital attack method with total variation(TV) item in the loss function.

Table II shows the generation result under different experiment settings, we crop the stickers out to get Table III from the result shows in Table II. From Table III, we could notice that our EOT-version stickers have vivider color than the No-EOT-version, the reason for that is only one punishment item $\lambda = 0.01$ is used which ensures less difference between original region and the generated sticker in the No-EOT-version sticker generation procedure. Comparing to it, there is two punishment item $\lambda_1 = 0.01, \lambda_2 = 0.5$ in the EOT-version sticker generation procedure, so the smoothness factor takes more controls of the generation result.

We evaluate our generated stickers with two different ways:

TABLE II

ORIGINAL PICTURE ON THE LEFT THAT WE USED TO GENERATE AD-STICKER. THE PICTURE ON THE RIGHT IS THE OUTCOME

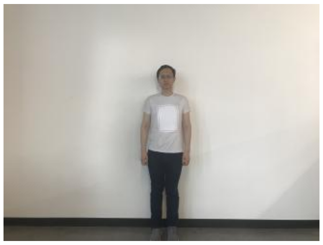
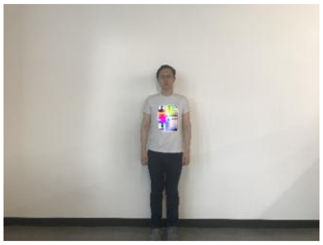
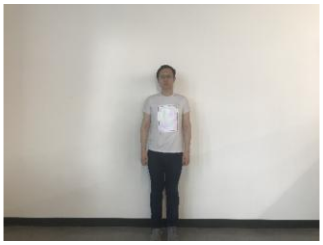
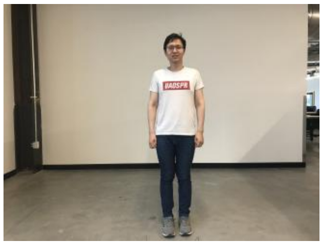
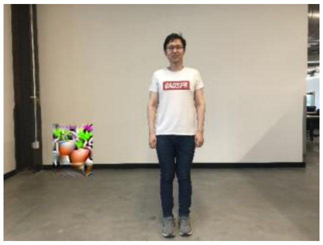
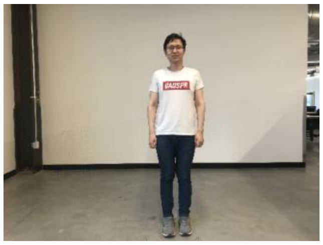
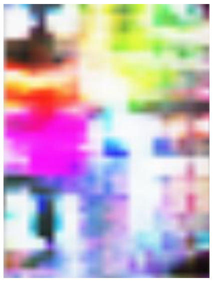
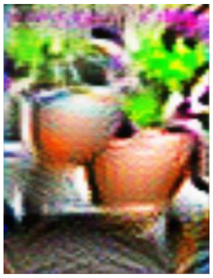
	Original	**EOT**	**No-EOT**
On Person			
Not on Person			

TABLE III

AD-STICKER GENERATED USING DIFFERENT SETTINGS

EOT on Person	**No-EOT on Person**	**EOT not on Person**	**No-EOT not on Person**
			

Watermark attack and physical displayed sticker attack.

1) Watermark attack evaluation: In the watermark attack testing, we put four stickers generated with different experiment settings each as a watermark on the 50 photos we take indoors and outdoors. We also use the original sticker and a white sticker for control group. We place the stickers in two different ways, one is on person, anther is side of the person. Table IV shows some of the watermark test examples. The result of this test can be found in Table V. It shows that Tiny YOLO object detector works quite well in control group photos as expected, but performs not very well in our photos with adversarial stickers as a watermark. We noticed that the EOT version stickers could perform better than the no-EOT version, We also noticed that even we created the sticker in the original mask area in a photo with person indoors, Tiny YOLO still could not detect the person in those photos taking outdoors(in different environment) with adversarial stickers even put in the mask area different from the generating process. If we put the sticker generated with setting "EOT not on person" on a person's front surface of the photo, Tiny YOLO model totally malfunction. The person was totolly invisible by the Tiny YOLO detector in this case. A possible reason is that if the sticker is generated with EOT but not on person, there is no mask on the person, the training process could take the whole person's attribute while generating the sticker. In comparison, the sticker generated with EOT but on person would lose some of the person's info under the mask itself while training.

To further evaluate our adversarial stickers, we randomly selected 318 photos with people from ImageNet, these photos include photo with multiple people and with only a part of a person like only a face. We put our sticker with best performance to these random photos. The result Table VI shows that the detection precision of Tiny YOLO model decreases from 1.0 to 0.55, which means our adversarial sticker generation method could be used more widely.

2) Physical displayed sticker attack: In the physical attack experiment setting, we display our sticker with best performance on the screen of a Macbook Retina 15-inch Laptop, and then let a person hold it and take photos in different scenarios.

TABLE IV
SOME DIGITAL AD WATERMARK TEST EXAMPLES

Sticker Generate Method Sticker Location	Original Picture	White Sticker	EOT on person	EOT not on person	No-EOT on person	No-EOT not on person
On person						
Not on person						
On person						
Not on person						
On person						
Not on person						
More	...	...	...	...	...	...

TABLE V
DIGITAL AD-WATERMARK EXPERIMENT WITH NON-OVERLAP WITH PERSON, EOT-BASED TRAINING STICKER, TESTED ON 50 INDOOR PERSON PHOTOS

Sticker Generate Method Sticker Location	Original pictures	White Sticker	EOT on person	EOT not on person	No-EOT on person	No-EOT not on person
On person	1.0	1.0	0.3	0.0	0.98	0.8
Not on person	1.0	1.0	0.92	0.1	0.98	0.86

Our experiment result shows that the precision of Tiny YOLO model decreases from 1.0 to 0.28.

VI. Discussion and Future Work

In this work, we focus on making a person disappeared from the Tiny YOLO object detector. Comparing to attacks on image classifier, this is a kind of untargeted attack to object detector. It would also be easy to conduct target attack on object detectors, making it report wrong classes of the detected object. In the future, we would construct more complex stickers even covering all the surface of a person as a real "invisible cloak". The problem exists in current physical attack implementation is that our model should be improved for printed version stickers. We know that if we print the sticker from RGB color space directly with a printer, the loss between the digital version and the printed version would be a lot. We will improve our model to construct more robust physical adversarial stickers in the future.

Digital adversarial watermark is generated using a single image in which the object person is fixed in the location in the middle. However, the attack is capable of generalizing to different picture background, different people and sticker locations. We think this phenomenon is due to the fact that YOLO, like all other deep learning based object detector, has shared weights in convolution networks layers. The shared weights mechanism results in adversarial characteristic propagating across the whole neural network and lowering the confidence of attacked targets. [5]

To better defense this kind of sticker attack, future object detection model could add adversarial training, which means taking original photos as well as white-box generated adversarial photos using our method purposed above as input to train more robust object detector.

TABLE VI
IMAGENET DIGITAL AD WATERMARK SAMPLES AND TEST RESULT (318 PHOTOS)

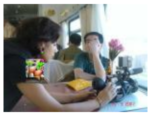
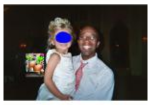
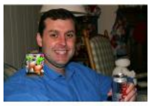
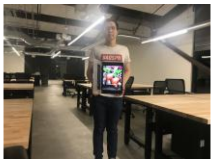
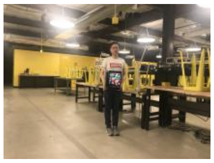
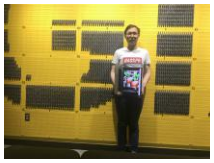
	Original ImageNet	**Add white-sticker**	**Add ad-sticker**
1			
2			
3			
More	...	...	...
Precision	1.0	0.97	0.55

TABLE VII
COMPARISON BETWEEN IMAGES WITH EOT-ATTACKED STICKERS AND IMAGES WITHOUT

	With EOT-attacked stickers	**Without stickers**
1		
2		
3		
More	...	...
Precision	0.28	1.0

ACKNOWLEDGMENT

We would like to thank Bob Tian and Annie Zhang providing support for our experiment, and other GIX faculties for their helpful feedback. This work was supported by Global Innovation Exchange (GIX), Tsinghua University, University of Washington and Baidu USA.

REFERENCES

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples.

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks.

[3] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. Robust physical adversarial attack on faster r-CNN object detector.

[4] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning models.

[5] 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.

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation.

[7] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples.

[8] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world.

[9] Jerry Liu, Fisher Yu, and Thomas Funkhouser. Interactive 3d modeling with a generative adversarial network.

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot MultiBox detector. 9905:21–37.

[11] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning.

[12] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings.

[13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-CNN: Towards real-time object detection with region proposal networks.

[15] 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 - CCS'16*, pages 1528–1540. ACM Press.

[16] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Adversarial generative nets: Neural network attacks on state-of-the-art face recognition.