Adversarial examples with transferred camouflage style for object detection

Xiaotong Deng 1, Zheng Fang 1, Yunfei Zheng 1,2, Yang Wang 1, Junming Huang 1, Xiaobing Wang 1,2 and Tieyong Cao 1*

1Institute of Command and Control Engineering, Army Engineering University, Nanjing, Jiangsu, 210000, China
2PLA Army Academy of Artillery and Air Defense, Nanjing, Jiangsu, 210000, China
*Corresponding author’s e-mail:1700710318@mails.guet.edu.cn

Abstract. Most of existing adversarial examples attacking methods for object detection models aim at generating subtle perturbation which is invisible to human vision. However, some perturbations with stronger intensity are equally effective and can even make the objects in adversarial examples more invisible. In this paper, an attack model based on the Generative Adversarial Network and style transfer method is designed to craft and add perturbation with camouflage style to the object area in the image. Experiments on PASCAL VOC 2012 dataset demonstrated that Adversarial examples generated from our model can attack both proposal-based detectors and regression-based object detection models effectively, and reduce the visual saliency of the objects.

1. Introduction
Object detection is one of the 3 essential tasks in computer vision. With the rapid development of adversarial examples [1] technology, the security of object detection has attracted more attention. However, most adversarial attacks and defenses [2-6] are still based on classifiers. Since object detectors need more complex networks than image classifiers to classify and locate the objects, it is more difficult to attack them, and there are few relevant algorithms at present. Xie et al. [7] proposed a white-box attacking method for proposal-based detectors and realized the attacking on Faster R-CNN [8], which was named Dense Adversary Generation (DAG). DAG regarded the object detection task as a multi-object classification task for inputs, so it achieved adversarial examples by attacking proposal regions and making the system misclassify all of them in each iteration, thus achieved an attack on Faster R-CNN. However, DAG cannot deal with regression-based detectors which do not require proposal regions. In 2018, Chen et al. [9] implemented an adversarial example attack against Faster RCNN in the physical domain with a similar method. In 2019, Wei et al. [10] proposed Unified and Efficient Adversary (UEA) algorithm, which can attack both proposal-based detectors and regression-based detectors. UEA used the Generative Adversarial Network (GAN) [11], proposed multi-scale attention feature loss and trained the generator with class loss in DAG jointly.

Some recent studies have shown that, unlike earlier adversarial example algorithms which generated subtle perturbations (invisible to human vision), some perturbations with stronger intensity are equally effective and can even make the objects in adversarial examples more invisible. Huang et al. [12] proposed Universal Physical Camouflage Attack (UPC), which can effectively attack the detectors in both digital and physical-world scenarios. This algorithm did not limited the perturbation
intensity and generated perturbation which was visible to the human vision but did not affect the visual effects. Duan et al. [13] proposed Adversarial Camouflage (AdvCam), which mainly generated the adversarial camouflage in physical world. This model camouflaged physical-world adversarial examples into customized and natural styles that appear legitimate to human vision by training with style loss.

In this paper, we propose jointing GAN and style transfer algorithm to generate stylized adversarial examples for fooling both proposal-based detectors and regression-based detectors. The proposed adversarial examples are equipped with "visible" and "stylized" perturbations. Moreover, unlike former works, we only perturb regions of foreground objects through the pixel-level annotations of images. Experimental results show that the mean Average Precision (mAP) reduced to 0.01 and 0.05 on Faster RCNN and SSD300 [14] respectively. In addition, the adversarial examples have "camouflage" effects that make the objects less visible in the image.

2. Method
Based on the GAN, this paper proposes a generation algorithm of adversarial examples on proposal-based detector. First, the generator produces perturbation from the original image, and then adds the perturbation into the object area of original image, which is limited by the annotation information of image semantic segmentation. In order to make the adversarial examples more legitimate visually, we introduce the style transfer algorithm [15] by using style loss and content loss to adjust the style of perturbation in feature extraction network, make them have similar style with the style reference image, such as texture information, and keep a small enough semantic missing comparing with the original image. Then, the discriminator will judge the facticity of the adversarial example. In addition, in order to improve the transferability of adversarial examples and realize the attack on regression-based detectors, multi-scale feature loss is introduced.

2.1 Problem Definition
The task of object detection is that given an input \( x \), then the model can detect the classes and positions of objects with high confidence scores. If the ground truth of \( x \) is defined as \( (L, B) \), in which \( L \) is the class label set of \( x \) and \( B \) is the corresponding coordinates set. If the detector classifies the objects in the image correctly, and the Intersection over Union (IoU) between the predicted bounding box and the ground truth is greater than the threshold, it will be regarded as a successful detection. Otherwise, the detection is deemed as fail. Therefore, in this paper, the detector is successfully attacked in the following cases: firstly, the object is not detected, that is, the prediction bounding box set is empty; secondly, the IoU between the predicted bounding box and the ground truth is less than the threshold value of 0.5; the third is wrong prediction of the class. The common evaluation metric for object detection is mAP, which reflects the accuracy of detection results. Therefore, when comparing the attack performance of the adversarial examples, we use mAP drop compared to that before the attack as the evaluation metric.

2.2 Network Architecture
Our model takes conditional GAN (cGAN) as the basic framework, consists of generator \( G \), discriminator \( D \) and feature extraction network \( F \), as shown in figure 1. In order to generate adversarial examples as close as possible with original images, we input them into \( D \) to compute GAN loss in equation (1).

\[
L_{\text{cGAN}}(G,D) = \mathbb{E}_x [\log D(x)] + \mathbb{E}_z [\log (1-D(G(z)))]
\]  

(1)

The generator is trained by cGAN and 4 loss functions from the F jointly, which can enhance the attack ability and transferability of the adversarial examples. In this paper, VGG16 [16] is selected for feature extraction and training. After training, the adversarial examples generated by \( G \) are used as the inputs of the detector to verify attack performance of the network.
2.3 Loss Functions
To implement adversarial examples with customized style, we need additional loss functions besides equation (1). Therefore, we propose 5 loss functions in this paper at all, as shown in equation (2).

\[ L = L_{\text{GAN}} + \alpha L_s + \beta L_c + \delta L_{\text{class}} + \varepsilon L_{\text{feats}} \]  (2)

In the proposed “stylized” perturbations, the style distance between adversarial example \( x' \) and a style reference image \( 2 \times 10^4 \) can express the stealthiness and reasonability of \( x \). A style loss is defined as follows:

\[ L_s = \sum_{l \in S} \left\| \tilde{G}(F_l(x')) - \tilde{G}(F_l(x)) \right\|_2^2 \]  (3)

Where \( F \) is the feature extraction network, \( \tilde{G} \) is the Gram matrix of deep features extracted at a set of style layers of \( F \), and \( S \) represents the set of style layers used for extracting style representations. In our work, we use two of all convolutional layers in \( F \). The style loss can craft adversarial examples in a customized style, but it may result in large differences from the original images. Therefore, a content loss is added to protect the content of original images, as shown in equation (4). \( F \) is the feature extraction network, and \( C \) represents the set of content layers used for extracting content representations. In our work, we use deeper layers in \( F \).

\[ L_c = \sum_{l \in C} \left\| F_l(x) - F_l(x') \right\|_2^2 \]  (4)

Object detection is actually a classification task of multi-bounding boxes, so we introduce in a class loss, as shown in equation (5). \( X \) is the feature map from \( F \) on input \( x \), \( t \) is the n-th proposal region from Region Proposal Network (RPN) on \( X \), while \( t \) and \( \hat{t} \) are the ground-truth and wrong label of \( t \), respectively. \( f_t(X,t) \) represents the class score of \( t \). We try to assign wrong class label to each proposal region generated by RPN through training, so as to achieve better attack effects.

\[ L_{\text{class}}(G) = \mathbb{E}_t \left[ \sum_{n=1}^N \left( f_t(X,t_n) - f_{\hat{t}}(X,t_n) \right) \right] \]  (5)

The above loss functions all aim at Faster RCNN, whose effects cannot be transferred to regression-based detectors, such as SSD. However, all detectors are based on feature extraction network, and if the feature extraction network is attacked, an effective adversarial example for all detectors can be generated. Therefore, multi-scale feature loss is proposed to enhance the transferability of adversarial examples, as shown in equation (6). \( X_m \) is feature map extracted from m-
th layer of $F$, $R_m$ is the feature map of the corresponding layer of the predefined reference image from $F$, $\odot$ is the Hadamard product between two matrices, and $A_m$ is the attention weight matrix, which is calculated from the candidate areas pre-selected by RPN, one pixel containing the object will get a larger weight. Such multi-scale feature loss makes the perturbation mainly focus on the objects to obtain better transferability.

$$L_{fus}(G) = E_x \left[ \sum_{m=1}^{M} \| A_m \odot (X_m - R_m) \|_2 \right]$$

(6)

3. Experiments

3.1 Experimental Setup

PASCAL VOC 2012 dataset is used for training and testing in our experiment. It contains 20 classes and it is one of the classic datasets commonly used in object detection tasks. In order to evaluate the attack effects of our model, we adopt mAP drop as the evaluation metric. mAP is the evaluation metric of object detection task, reflecting the model's ability to successfully detect the objects. The more the mAP decreases, the better the attack performance. We use Faster R-CNN and SSD300 to reflect the attacking effects of our model on proposal-based and regression-based detectors.

Based on the camouflaged and visual effect of the camouflage pattern, we choose Canadian Disruptive Pattern (CADPAT) of Canadian army as the style reference image in the experiment, to generate adversarial examples with camouflage style, which can reduce the significance of the object in human vision while resisting machine vision. In addition, we choose the Relu layer after conv3-3 and the Relu layer after conv4-2 in VGG16 to compute their style loss and multi-scale feature loss, and the Relu layer after conv5-2 as the content layer for content loss. To compute the total loss in equation (1), we set $\alpha = 250$, $\beta = 10^{-4}$, $\delta = 1$, and $\varepsilon$ is $1 \times 10^{-4}$ and $2 \times 10^{-4}$ for two selected layer respectively.

Figure 2. Comparisons with DAG and UEA on Faster RCNN and SSD300. (Ours refers to our model)
3.2 Results

3.2.1 Comparisons with other methods. In this section, we compare our method with the current state-of-the-art attacking methods DAG and UEA, and perform the same Faster RCNN and SSD300 on the adversarial examples from the three methods to observe the accuracy drop respectively. The results are shown in Table 1. We can see, all of 3 methods can attack Faster RCNN well, while our method achieves more drop in mAP. However, DAG cannot attack SSD because the mAP drop is only 0.05. By contrast, UEA and our method achieve 0.47 and 0.62 mAP drop respectively. The results demonstrate that our method has better attack performance and transferability. We show the adversarial examples of our model, UEA and DAG in figure 2, and compare their detection accuracy on Faster RCNN and SSD300. From the picture, we can see that both Faster RCNN and SSD300 can work well on all clean images. For DAG and UEA, (c) and (d) show that most of adversarial examples attack Faster RCNN successfully with zero bounding box or wrong class label, but DAG can hardly attack SSD300 in column (g) while UEA can attack SSD300 to some degree in column (h). However, the adversarial examples from our model get better attack effects on both Faster RCNN and SSD300. In column (e), we can see Faster RCNN doesn’t detect any objects on all five images. For column (i), the SSD300 detects zero object on one image, predicts wrong labels on three images and misses an object on the forth image.

Table 1. Comparisons of detect results.

| Attack Methods | mAP Faster RCNN | mAP SSD300 |
|----------------|-----------------|------------|
| Original Images | 0.71 | 0.67 |
| DAG | 0.06 | 0.62 |
| UEA | 0.06 | 0.20 |
| Ours | 0.01 | 0.05 |

a Ours refers to the model proposed in this paper.

3.2.2 Objective saliency in human vision. In this section, we verify the visual effects of adversarial examples with camouflage style. There has been work on eye tracking self-calibration using image saliency to estimate user gaze for calibration purposes [17], so we use some salient object detection models to simulate human vision, including Attentive Saliency Network (ASNet) [18] and High-Dimensional Color Transform (HDCT) [19], as shown in figure 3. We can see that all five adversarial examples can attack the salient object detection models well and reduce the salience of the objects. That is, we transferred the confusing effects of camouflage patterns on human vision to adversarial examples successfully, and generated examples that can reduce the visual saliency of the objects in image.

Figure 3. Comparisons of human visual effects.
4. Conclusion
In this paper, we propose an adversarial example generation method to attack object detection models. This method is based on conditional GAN, combined with camouflage style transfer, to generate adversarial examples which have better attack effects and transferability than existing methods such as UEA. Experimental results on PASCAL VOC 2012 verified the effectiveness of our model. That is, our adversarial examples can attack both proposal-based detectors and regression-based detectors. In addition, they can reduce the saliency of the objects in human vision.

Acknowledgments
This work was supported by the National Natural Science Foundation for Youth of China grant funded by the Chinese government (No. 61801512).

References
[1] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi and P. Frossard, Universal adversarial perturbations. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[2] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and Harnessing Adversarial Examples, arXiv preprint arXiv:1412.6572, 2015.
[3] S. Baluja, I. Fischer, Adversarial Transformation Networks: Learning to Generate Adversarial Examples, arXiv preprint arXiv:1703.09387, 2017.
[4] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. arXiv preprint arXiv:1801.09344, 2018.
[5] N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami, Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks, In IEEE Symposium on Security and Privacy (SP), pp. 582-597, 2016.
[6] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, Adversarial Examples for Semantic Segmentation and Object Detection, arXiv preprint arXiv:1703.08603, 2017.
[7] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In ICCV. IEEE, 2017.
[8] S. Q. Ren, K. M. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 142-158, 2017.
[9] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. Robust physical adversarial attack on faster r-cnn object detector. arXiv preprint arXiv:1804.05810, 2018.
[10] Wei X, Liang S, Chen N, et al. Transferable Adversarial Attacks for Image and Video Object Detection[J]. 2018.
[11] Mirza M, Osindero S. Conditional Generative Adversarial Nets[J]. arXiv preprint arXiv:1411.1784, 2014.
[12] Huang L, Gao C, Zhou Y, et al. Universal Physical Camouflage Attacks on Object Detectors[J]. 2019.
[13] Duan R, Ma X, Wang Y, et al. Adversarial Camouflage: Hiding Physical-World Attacks with Natural Styles[J]. 2020.
[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. SSD: Single shot multibox detector. In ECCV, 2016.
[15] Johnson J, Alahi A, Fei-Fei L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution[J]. 2016.
[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
[17] Y. Sugano, Y. Matsushita, and Y. Sato. Calibration-free gaze sensing using saliency maps. In Proc. CVPR, pages 2667-2674, 2010.
[18] Wang W, Shen J, Dong X, et al. Inferring Salient Objects from Human Fixations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019:1-1.

[19] Kim J, Han D, Tai Y W, et al. Salient Region Detection via High-Dimensional Color Transform[J]. IEEE Transactions on Image Processing, 2014, 25(1):1-1.