文章编号：0253-374X(2022)10-1377-08
DOI：10.11908/j.issn.0253-374x.22227

针对车载环境感知系统的对抗样本生成方法

黄世泽[1]，张肇鑫[2]，董德存[1]，秦晋哲[2]
（1. 上海市轨道交通结构耐久与系统安全重点实验室，上海 201804；2. 同济大学 道路与交通工程教育部重点实验室，上海 201804）

摘要：针对车载环境感知场景中的目标检测系统，提出了一种针对目标检测器的对抗样本生成方法。该方法能够实现对目标检测器的白盒对抗攻击，包括目标隐身攻击和目标定向攻击。在Rail数据集和Cityscapes数据集中进行测试，测试结果验证了所提方法对YOLO目标检测器对抗攻击的有效性。

关键词：车载环境感知系统；对抗攻击；目标检测；深度学习；白盒攻击
中图分类号：TP389.1　　**文献标志码**：A

Adversarial Example Generation Method for Vehicle Environment Perception System

HUANG Shize[1], *ZHANG Zhaoxin*[2], *DONG Decun*[1], *QIN Jinzhe*[2]
(1. Shanghai Key Laboratory of Rail Infrastructure Durability and System Safety, Shanghai 201804, China; 2. Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China)

Abstract: A method of generating adversarial examples against object detectors was proposed for object detection system in vehicle environment perception scenarios. The method achieves white-box adversarial attacks on object detectors, i.e., object invisible attacks and object targeted mis-detectable attacks. On the Rail dataset and Cityscapes dataset, experimental results indicate that the method has good performance on the object invisible attacks and the object targeted mis-detectable attacks in the process of YOLO object detection.

Key words: vehicle environment perception system; adversarial attack; object detection; deep learning; white-box attack

随着深度学习和卷积神经网络(convolutional neural network，CNN)的不断发展，通过CNN解决诸如图像分类[1]、目标检测[2]以及故障诊断[3-4]等问题已成为共识。在智能交通领域[5]，通过车载摄像设备采集车辆行驶环境视觉信息，基于深度学习检测车辆行驶前方障碍物，这可突破司机感知的局限性，提高交通运营安全。

研究表明，对抗样本的存在对深度学习造成较大的威胁，即通过对输入图像施加人眼不可察觉的细微扰动，可以使深度神经网络以较高的置信度输出任意想要的分类，这样的输入称为对抗样本。Szegedy等[6]提出了一种有限记忆BFGS(limited BFGS，L-BFGS)算法，通过尽量找到最小的可能的攻击扰动来生成对抗样本，即使存在扰动的图像与干净的图像只有微小的差别，甚至这些扰动肉眼察觉不到，也会导致分类器分类错误。Moosavi-Dezfooli等[7]证明了深度学习网络中普遍存在一种使其错误识别的扰动。Goodfellow等[8]提出了快速梯度符号方法(fast gradient sign method，FGSM)，寻找深度学习模型的梯度变化最大方向，并按照此方向对图像添加扰动。上述几种方法需要获取网络结构，因此被称为白盒攻击。除了上述几种方法，基于雅可比矩阵的显著性图攻击(Jacobian-based saliency map attack，JSMA)[9]、Carlini&Wagner(C&W)算法[10]、迭代极小可能类法(iterative least-likely class method，ILCM)[11]、TargetedFool[12]也是白盒攻击方法。与白盒攻击相对应的是黑盒攻击，黑盒攻击不需要获取网络的详细结构。文献[13]中，通过粒子群优化(particle swarm optimization，PSO)算法寻找对抗样本，不需要获取网络结构，取得了较好的实验效果。在物理世界，通过对真实世界中的车牌进行黑盒攻击，欺骗车牌识别系统，从而

收稿日期：2022-05-10
基金项目：国家自然科学基金(61703308)
第一作者：黄世泽(1983—)，男，副教授，博士生导师，工学博士，主要研究方向为交通信息可信感知。E-mail：hsz@tongji.edu.cn
通信作者：张肇鑫(1994—)，男，博士生，主要研究方向为交通信息可信感知。E-mail：1910925@tongji.edu.cn

论文拓展介绍

验证了攻击方法的迁移性[14]。文献[15]中提出一种面向人脸活体检测的对抗样本生成方法。总体来说,上述方法都是基于分类器网络的对抗样本生成方法,对于其他类型的深度学习对抗攻击具有非常重要的借鉴意义。

随着深度学习网络应用场景的不断拓展,针对目标检测器的对抗样本生成方法[16-17]近年来陆续被提出。Xie等[18]提出了稠密对抗生成(dense adversary generation,DAG)算法,将梯度下降算法应用到对抗样本的优化问题来实现对目标检测器的攻击。2019年,Wei等[19]提出了统一有效对抗(unified and efficient adversary,UEA)算法,基于生成对抗网络(generative adversarial networks,GAN)框架来获取对抗性图像和视频。Wang等[20]将投影梯度下降(projected gradient descent,PGD)算法运用到目标检测器攻击,取得了较好的攻击效果,该算法能够应用于许多神经网络结构。Huang等[21]提出了针对Faster R-CNN的改进的PGD算法和改进的C&W算法,成功攻击了Faster R-CNN目标检测器。Xiao等[22]提出了一种针对目标检测器的对抗样本生成方法,无目标攻击效果较好,但未得到目标定向攻击效果和目标隐身攻击效果。通过寻优算法在不需要获取网络参数的条件下生成能够让目标隐身的黑盒对抗样本,但生成过程中需要查询模型输出结果的次数过多[23]。物理攻击方面,通过改进的ShapeShifter方法并利用Faster R-CNN网络,在不同的距离和角度攻击中文停车牌,取得了较好的攻击效果[24]。

基于分类器网络的对抗样本生成方法不能有效攻击目标检测器,现有攻击目标检测器的对抗样本生成方法仅针对无目标攻击,攻击方式和效果有限。因此,针对YOLO目标检测器的对抗样本生成问题,提出了目标隐身攻击和目标定向攻击2种对抗样本生成方法。

1 对抗样本生成方法

1.1 对抗样本生成流程

首先,通过目标检测网络的训练参数得到网络输出信息,包括目标所在的包围框和对应的类别置信度;然后,设计一种损失函数,用于对抗样本所需梯度信息的生成;最后,通过线性化梯度信息获取针对目标检测网络的对抗样本。对抗样本生成流程如图1所示。

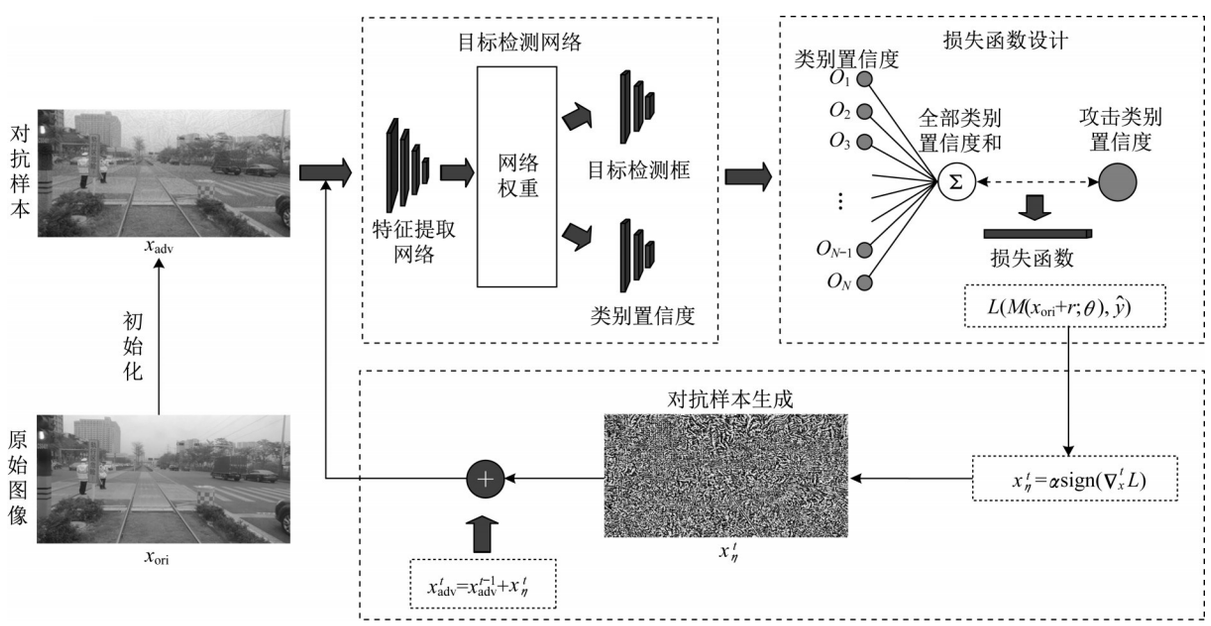

图1 对抗样本生成流程

Fig.1 Main framework of generating adversarial examples

对抗样本生成算法的具体实施过程如下:

(1)对抗样本初始化,将原始图像作为对抗样本的初始图像。

(2)将初始化的对抗样本输入目标检测器YOLO,得到目标的位置和置信度。

(3)损失函数设计。

(4)根据构建的损失函数计算对抗样本相应梯度。

(5)通过反向传播算法更新对抗样本。

(6)判断是否达到设置的迭代次数,若是则输出对抗样本,若不是则返回(2)进行下一次迭代。

1.2　目标隐身攻击对抗样本生成算法

目标隐身攻击具体表现为：YOLO目标检测器不能检测出图像中真实存在的目标，也就是目标在YOLO目标检测器下处于隐身状态。在针对目标隐身攻击的对抗样本生成方法中，需要寻找能够使目标类别的置信度最小的扰动，如下所示：

$$\min_r L(M(x_{\text{ori}}+r;\theta),\hat{y}),\text{s.t.}\hat{y}=0 \tag{1}$$

式中：L为损失函数；M为目标检测器；x_{ori}为原始图像；r为需要计算的对抗扰动；θ为模型参数；$\hat{y}$为模型对图像的预测值，在本研究中指的是目标类别的置信度。

通过损失函数的设计降低真实目标类别的置信度，生成隐身攻击所需的对抗样本梯度信息。设计的损失函数如下所示：

$$L=\sum_{i=0}^{N}(C_i(M(x+r;\theta))) \tag{2}$$

式中：N为由目标检测器计算得到的目标个数；C为目标类别的置信度。

L_∞范数约束下的对抗扰动为

$$x_\eta^t=\alpha\text{sign}(\nabla_x^t L) \tag{3}$$

L_2范数约束下的对抗扰动为

$$x_\eta^t=\alpha\frac{\nabla_x^t L}{\|\nabla_x^t L\|_2} \tag{4}$$

式(3)和式(4)中：$\|\ \|_2$为L_2范数归一化；α为学习率，$\alpha=0.02$；$\nabla_x^t L$为计算得到的相应梯度，其中t为迭代次数；sign为符号函数，即$\text{sign}(\phi)=\begin{cases}-1,\phi<0\\0,\phi=0\\1,\phi>0\end{cases}$，其中$\phi$为真实值。

对抗样本x_{adv}^t的计算式为

$$x_{\text{adv}}^t=x_{\text{adv}}^{t-1}-x_\eta^t \tag{5}$$

1.3　目标定向攻击对抗样本生成算法

目标定向攻击具体表现为：原始类别为“car”的目标，YOLO目标检测器错误地将其识别为类别“bus”。在针对目标定向攻击的对抗样本生成方法中，需要寻找能够使目标类别的置信度最小的扰动，如下所示：

$$\min_r L\left(M(x_{\text{ori}}+r;\theta),\hat{y}\right),\text{s.t.}\hat{y}=\hat{y}' \tag{6}$$

式中：$\hat{y}'$为定向识别的目标类别置信度。

在针对目标定向攻击的对抗样本生成方法中，通过损失函数的设计降低被攻击目标类别的置信度，提高攻击定向目标类别的置信度，进而生成目标定向攻击中的对抗样本梯度。

通过损失函数的设计降低真实目标类别的置信度，进而生成目标攻击所需的对抗样本梯度信息。损失函数如下所示：

$$L=\sum_{i=0}^{N}(C_i(M(x+r;\theta))) \tag{7}$$

L_∞范数约束下的对抗扰动为

$$x_\eta^t=\alpha\text{sign}(\nabla_x^t L) \tag{8}$$

L_2范数约束下的对抗扰动为

$$x_\eta^t=\alpha\frac{\nabla_x^t L}{\|\nabla_x^t L\|_2} \tag{9}$$

对抗样本x_{adv}^t的计算式为

$$x_{\text{adv}}^t=x_{\text{adv}}^{t-1}+x_\eta^t \tag{10}$$

1.4　算法伪代码

对抗样本生成算法的伪代码如图2所示。

Input:
原始干净图像x_{ori}
定义的损失函数L
迭代次数n
学习率α
Output:
对抗样本x_{adv}^n

1: $t=0$,初始化对抗样本$x_{\text{adv}}^0=x_{\text{ori}}$
2: while $t<n$ do
3:　　输入x_{adv}^{t-1}到模型 M，获取对应梯度信息$\nabla_x^t L$
4:　　if 目标隐身攻击:
5:　　根据公式(3)得到L_∞范数约束下x_η^t
6:　　根据公式(4)得到L_2范数约束下x_η^t
7:　　根据公式(5)得到x_{adv}^t
8:　　if 目标定向攻击:
9:　　根据公式(8)得到L_∞范数约束下x_η^t
10:　　根据公式(9)得到L_2范数约束下x_η^t
11:　　根据公式(10)得到x_{adv}^t
12:　　end if
13:　　$t=t+1$
14: end while
15: return x_{adv}^n

图2　对抗样本生成算法的伪代码

Fig.2　Pseudo code for adversarial example generation algorithm

2　方法验证

2.1　数据集来源介绍

为验证数据的有效性，收集了Rail数据集和Cityscapes数据集[25]。Rail数据集为根据深圳龙华有轨电车运行的真实数据制作的有轨电车数据集，数据图像的原始分辨率为1 920×1 080，实验中将分辨率调整为960×540。Rail数据集包括了1 094张有轨电车运行环境图片，选择全部图片进行目标隐身攻击，选择

115张包含类别“car”和“bus”的目标进行目标定向攻击。Cityscapes数据集为由车载相机采集的德国真实城市道路图像数据集,选取了包含类别“car”和“bus”的404张图片进行实验。为减少计算消耗,将图像分辨率由2 048×1 024调整为1 024×512。软件测试环境为TensorFlow 1.13.1和Keras 2.2.4。硬件环境为Intel(R) Core(TM) i7-7800X CPU, 3.50 GHz, 32 GB内存, NVIDIA GeForce GTX 1080Ti 11GB。

2.2 目标隐身攻击实验结果

目标隐身攻击下的原始图像和对抗样本如图3所示。由图3可知,YOLO version3(YOLOv3)目标检测器能够有效识别原始图像中的目标。本方法生成的对抗样本导致YOLOv3目标检测器不能识别出目标($L_2=0.2, L_\infty=0.05$)。

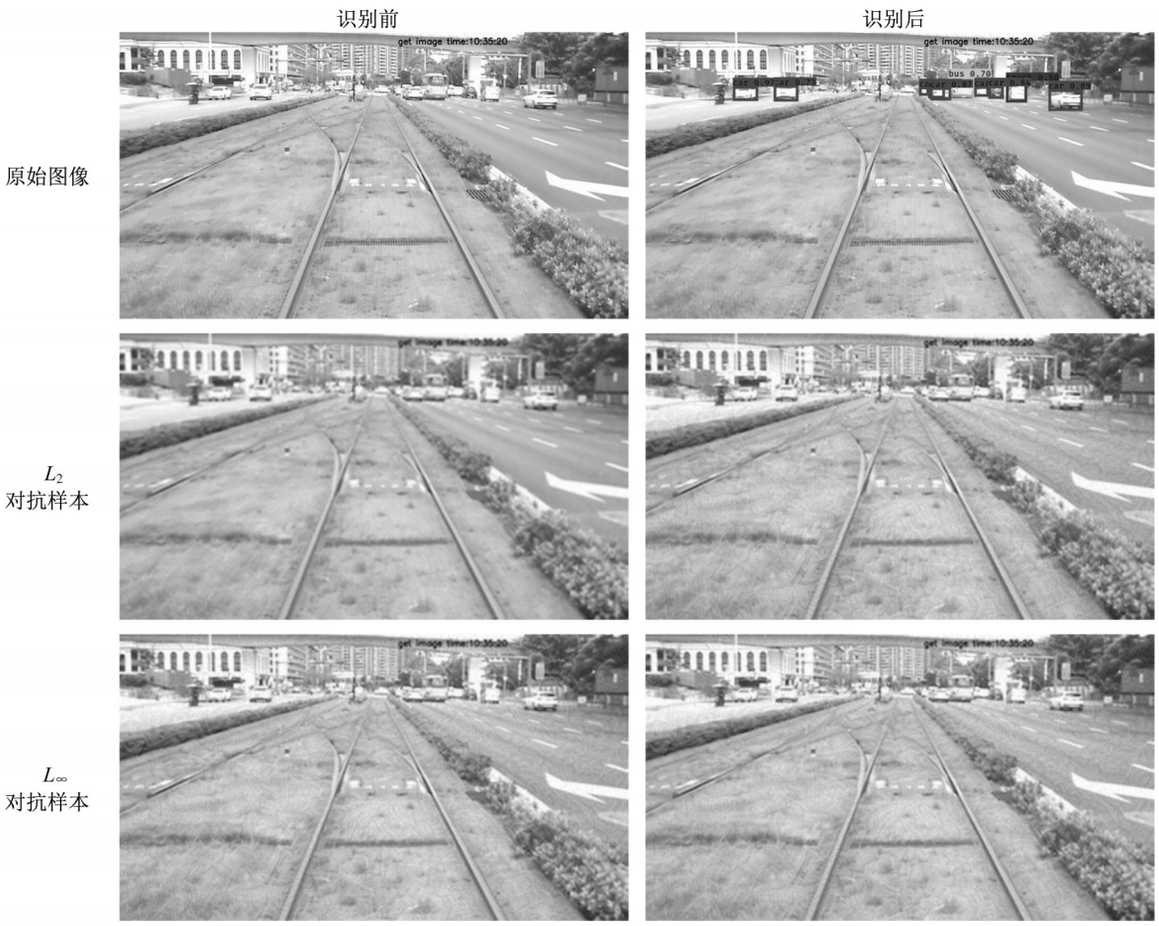

图3 原始图像和目标隐身攻击下的对抗样本

Fig.3 Original image and adversarial example under object invisible attacks

平均准确率(mean average precision, α_{mAP})指标通常被用作目标检测数据集的评价指标。为了进一步评估对抗效果,利用平均准确率指标进行2种数据集的效果验证。实验中交并比(intersection over union, IoU)阈值设置为0.5,目标置信度设置为0.5。同时,改写了攻击分类器的对抗样本算法CI-FGSM[26]和AI-FGSM[27],用于攻击YOLOv3目标检测器,并与本方法进行比较,如图4所示。实验结果表明,在YOLOv3目标检测器的目标隐身攻击中,本方法相比其他2种方法攻击效果更加明显。

为了综合评估本方法的攻击效果,利用峰值信噪比(peak signal-to-noise ratio, β_{PSNR})和结构相似性(structural similarity, γ_{SSIM})指标进行图像相似度比较,如图5和图6所示。

由图4和图5可见:本方法的峰值信噪比与CI-FGSM相近,但本方法保持了较高的攻击成功率;与AI-FGSM相比,本方法保持了较高的攻击成功率和较高的图像峰值信噪比。由图4和图6可见:L_2范数攻击下结构相似性相近,同时本方法保持了较高的攻击成功率;L_∞范数攻击下,虽然本方法的结构相似性有所降低,但是仍保持了较高的攻击成功率。综合来看,本方法更加有效。

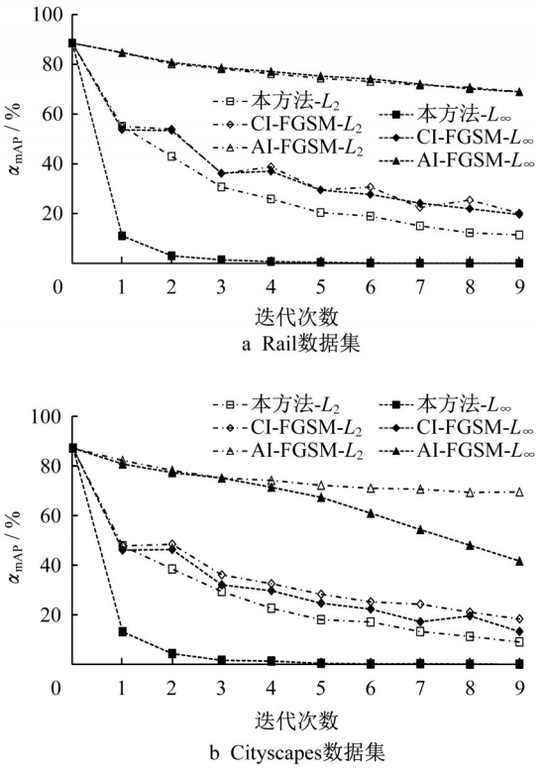

图4　基于目标隐身攻击对抗样本的平均准确率

Fig.4　Mean average precision of adversarial example under object invisible attacks

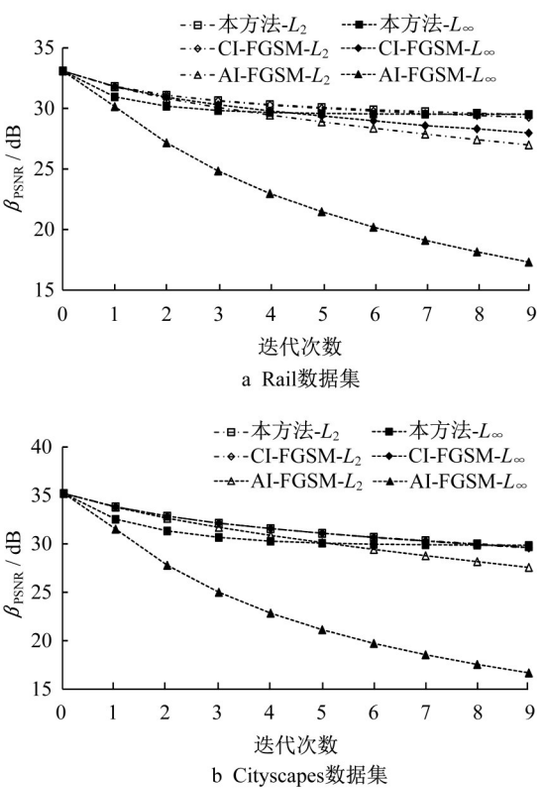

图5　原始图像与对抗样本的峰值信噪比

Fig.5　Peak signal-to-noise ratio of original image and adversarial example

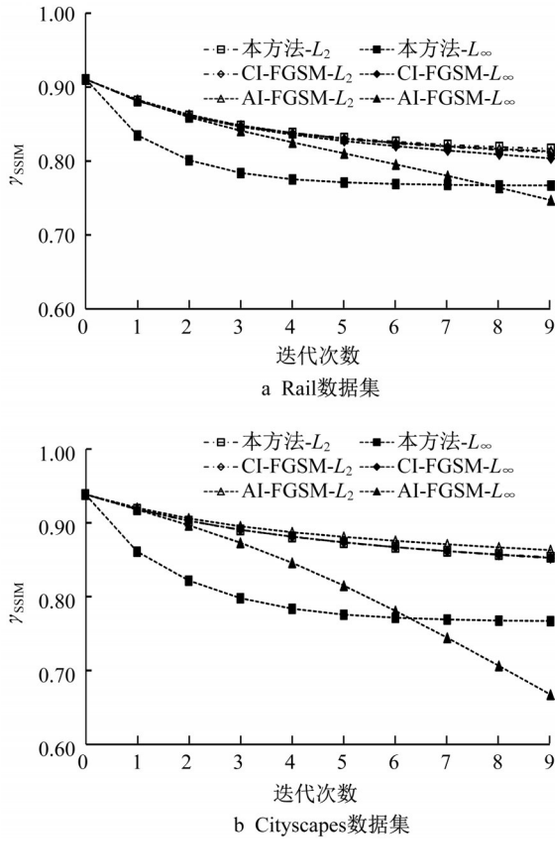

图6　原始图像与对抗样本的结构相似性

Fig.6　Structural similarity of original image and adversarial example

2.3　目标定向攻击实验结果

原始图像和目标定向攻击下对抗样本图像如图7所示。由图7可知，YOLOv3目标检测器能够有效识别原始图像中出现的目标。本方法生成的对抗样本则导致YOLOv3目标检测器将原本类别为“car”的目标识别成了“bus”（$L_2=1.0$，$L_\infty=0.05$）。

本方法对抗样本目标定向攻击是将原本类别为“car”的目标识别成了“bus”，因此利用类别“car”的识别召回率（recall rate，r_{RR}）和类别“bus”的识别准确率（precision rate，p_{PR}）指标进行对抗样本攻击效果的验证，如图8和图9所示。实验结果表明，与其他方法相比，本方法具有更好的攻击效果。

从图8～10可见，本方法与CI-FGSM和AI-FGSM相比，峰值信噪比相近，同时本方法攻击效果较好。从图8、图9和图11可见：L_2范数攻击下3种方法的结构相似性相近，同时本方法攻击效果较好；L_∞范数攻击下，虽然本方法结构相似性有所降低，但是仍保持了较高的攻击成功率。因此，本方法更加有效。

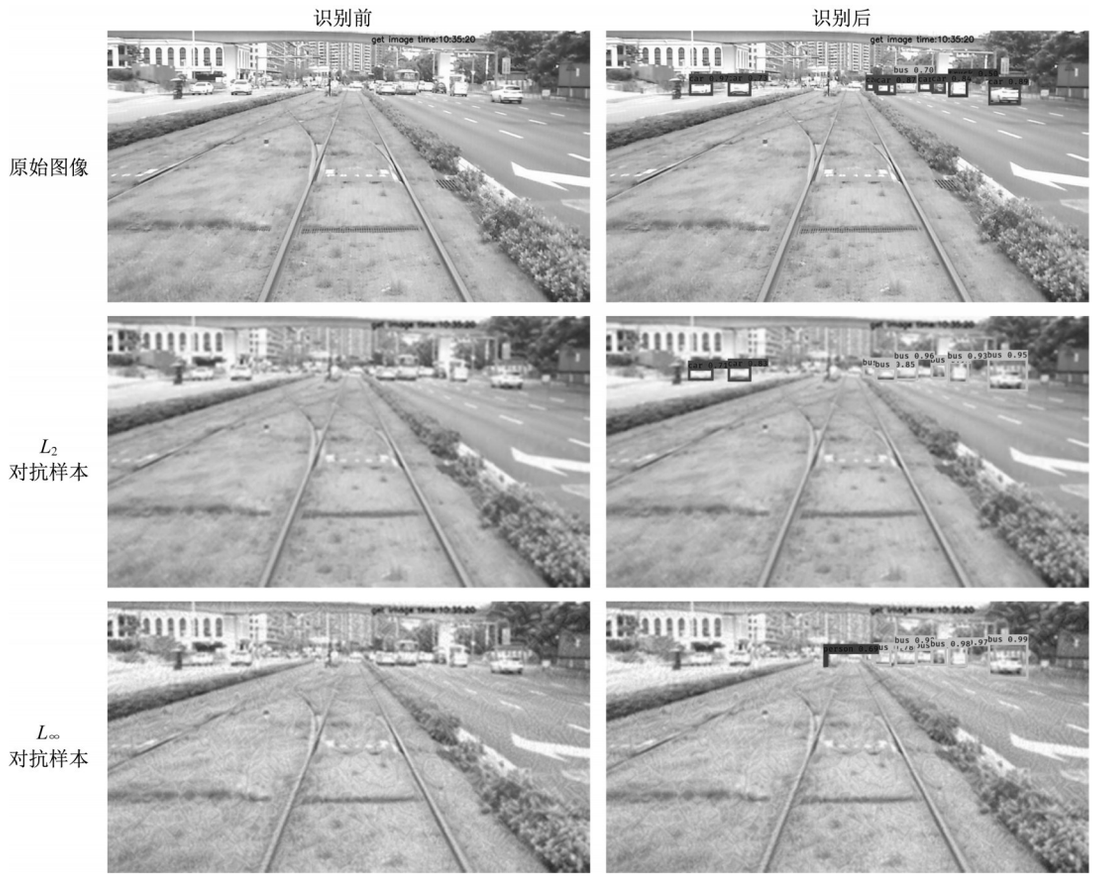

图7 原始图像和目标定向攻击下的对抗样本

Fig.7 Original image and adversarial example under object targeted mis-detectable attacks

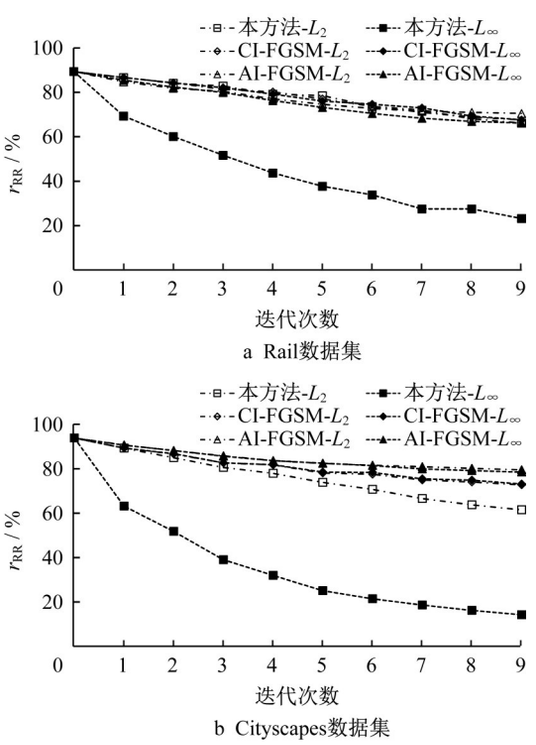

图8 基于目标定向攻击的对抗样本召回率

Fig.8 Recall rate of adversarial example under object targeted mis-detectable attacks

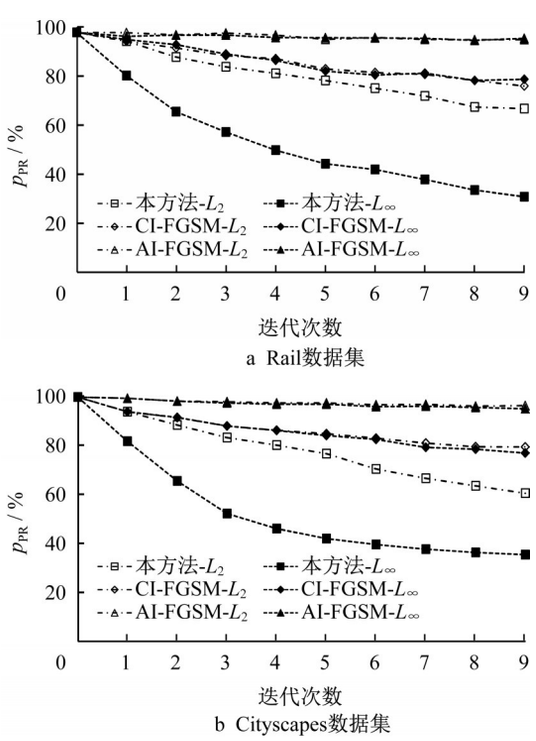

图9 基于目标定向攻击的对抗样本准确率

Fig.9 Precision rate of adversarial example under object targeted mis-detectable attacks

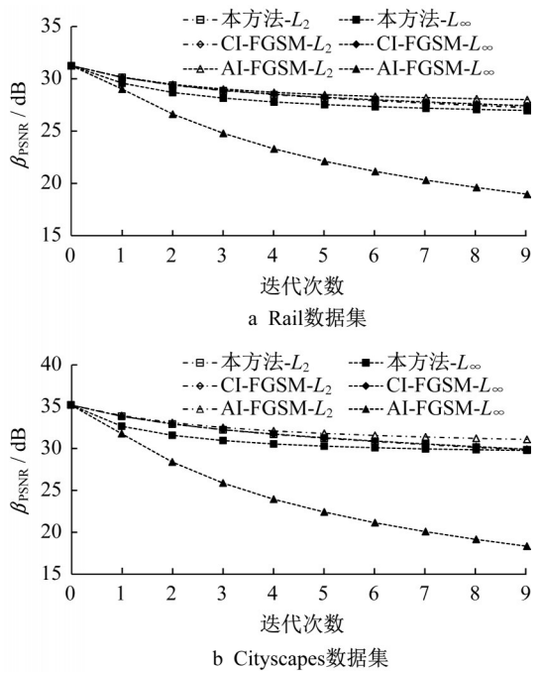

图10　原始图像与对抗样本的峰值信噪比

Fig.10　Peak signal-to-noise ratio of original image and adversarial example

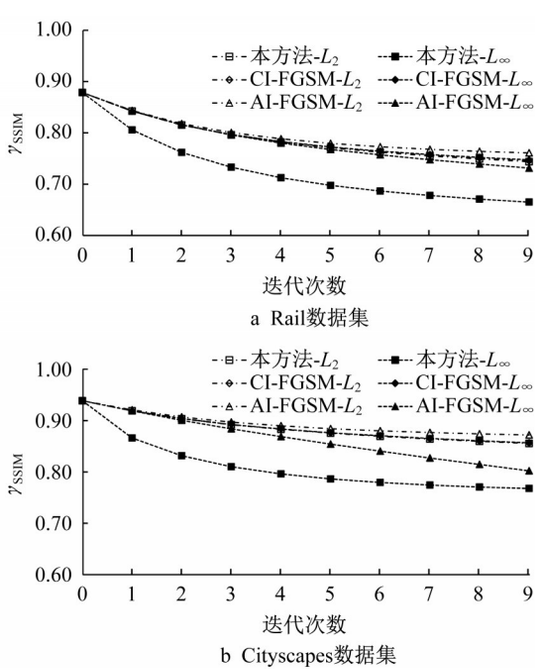

图11　原始图像与对抗样本的结构相似性

Fig.11　Structural similarity of original image and adversarial example

3　结语

针对YOLO目标检测器，提出了攻击效果更加全面的对抗样本生成方法。通过获取目标检测器的网络结构，设计对抗样本的损失函数，然后通过所提出的对抗样本生成方法获取对抗样本。在Rail数据集和Cityscapes数据集上进行了验证，表明该方法对YOLOv3目标检测器具有较高的攻击率，并且该方法能够实现目标隐身攻击和目标定向攻击。

作者贡献声明：

黄世泽：提出对抗样本生成研究方案，最终版本修订。

张肇鑫：具体程序设计实现。

董德存：基于车载环境感知的可靠性提出研究思路。

秦晋哲：算法的验证和对比。

参考文献：

[1] HUANG Shize, ZHAI Yachan, ZHANG Miaomiao, *et al*. Arc detection and recognition in pantograph-catenary system based on convolutional neural network [J]. Information Sciences, 2019, 501:363.

[2] 黄世泽，杨玲玉，陶婷，等. 基于实例分割的有轨电车障碍物入侵检测及轨道识别方法[J]. 上海公路，2021(2)：89.
HUANG Shize, YANG Lingyu, TAO Ting, *et al*. A method of tram obstacle intrusion detection and track recognition based on instance segmentation[J]. Shanghai Highways, 2021(2):89.

[3] 黄世泽，陈威，张帆，等. 基于弗雷歇距离的道岔故障诊断方法[J]. 同济大学学报(自然科学版)，2018，46(12)：1690.
HUANG Shize, CHEN Wei, ZHANG Fan, *et al*. Method of turnout fault diagnosis based on Fréchet distance[J]. Journal of Tongji University (Natural Science), 2018, 46(12):1690.

[4] TAO Ting, DONG Decun, HUANG Shize, *et al*. Gap detection of switch machines in complex environment based on object detection and image processing [J]. Journal of Transportation Engineering, Part A: Systems, 2020, 146(8): 04020083.

[5] HUANG Shize, YANG Lingyu, ZHANG Fan, *et al*. Turnout fault diagnosis based on CNNs with self-generated samples[J]. Journal of Transportation Engineering, Part A: Systems, 2020, 146(9):1.

[6] SZEGEDY C, ZAREMBAW, SUTSKEVER I, *et al*. Intriguing properties of neural networks[J/OL]. [2021-12-21]. https://arxiv.org/abs/1312.6199.

[7] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016:2574-2582.

[8] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [J/OL]. [2021-12-20]. https://arxiv.org/abs/1412.6572.

[9] PAPERNOT N, MCDANIEL P, JHA S, *et al*. The limitations of deep learning in adversarial settings [C]// Proceedings of 2016 IEEE European Symposium on Security and Privacy (EuroS&P). Los Alamitos: IEEE Computer

Society, 2016:372-387.

[10] CAELINI N, WAGNER D. Towards evaluating the robustness of neural networks [C]//Processing of the 2017 IEEE Symposium on Security and Privacy (SP). Los Alamitos: IEEE Computer Society, 2017:39-57.

[11] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[J/OL]. [2021-07-08]. https://arxiv.org/abs/1607.02533.

[12] 张华,高浩然,杨兴国,等. TargetedFool:一种实现有目标攻击的算法[J].西安电子科技大学学报,2021,48(1):149.
ZHANG Hua, GAO Haoran, YANG Xingguo, *et al*. TargetedFool: an algorithm for achieving targeted attacks[J]. Journal of Xidian University,2021,48(1):149.

[13] 陈晋音,陈治清,郑海斌,等. 基于PSO的路牌识别模型黑盒对抗攻击方法[J]. 软件学报,2020,31(9):2785.
CHEN Jinyin, CHEN Zhiqing, ZHENG Haibin, *et al*. Black-box physical attack against road sign recognition model via PSO [J]. Journal of Software,2020,31(9):2785.

[14] 陈晋音,沈诗婧,苏蒙蒙,等. 车牌识别系统的黑盒对抗攻击[J]. 自动化学报,2021,47(1):121.
CHEN Jinyin, SHEN Shijing, SU Mengmeng, *et al*. Black-box adversarial attack on license plate recognition system[J]. *Acta Automatica Sinica*,2021,47(1):121.

[15] 马玉琨,毋立芳,简萌,等. 一种面向人脸活体检测的对抗样本生成算法[J]. 软件学报, 2019,30(2):469.
MA Yukun, WU Lifang, JIAN Meng, *et al*. Approach to generate adversarial examples for face-spoofing detection[J]. Journal of Software,2019,30(2):469.

[16] 张翰韬. 面向图像目标检测的对抗攻击[D]. 合肥:中国科学技术大学,2020.
ZHANG Hantao. Adversarial attack on image object detection [D]. Hefei: University of Science and Technology of China, 2020.

[17] 刘嘉阳. 针对图像分类的对抗样本防御方法研究[D]. 合肥:中国科学技术大学,2020.
LIU Jiayang. Research on defense against adversarial examples for image classification[D]. Hefei: University of Science and Technology of China,2020.

[18] XIE Cihang, WANG Jianyu, ZHANG Zhishuai, *et al*. Adversarial examples for semantic segmentation and object detection [C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Los Alamitos: IEEE Computer Society, 2017:1378-1387.

[19] WEI Xingxing, LIANG Siyuan, CHEN Ning, *et al*. Transferable adversarial attacks for image and video object detection [J/OL]. [2021-11-30]. https://arxiv. org/abs/1811.12641.

[20] WANG Yutong, WANG Kufeng, ZHU Zhanxing, *et al*. Adversarial attacks on faster R-CNN object detector [J]. Neurocomputing,2020,382:87.

[21] HUANG Shize, LIU Xiaowen, YANG Xiaolu, *et al*. Two improved methods of generating adversarial examples against faster R-CNNs for tram environment perception systems [J]. Complexity,2020,2020:6814263.

[22] XIAO Yatie, PUN Chi-Man, LIU Bo. Fooling deep neural detection networks with adaptive object-oriented adversarial perturbation[J]. Pattern Recognition,2021,115:107903.

[23] WANG Yajie, TAN Yu-an, ZHANG Wenjiao, *et al*. An adversarial attack on DNN-based black-box object detectors[J]. Journal of Network and Computer Applications, 2020, 161: 102634.

[24] HUANG Shize, LIU Xiaowen, YANG Xiaolu, *et al*. An improved ShapeShifter method of generating adversarial examples for physical attacks on stop signs against faster R-CNNs[J]. Computers & Security,2021,104:102120.

[25] CORDTS M, OMRAN M, RAMOS S, *et al*. The Cityscapes dataset for semantic urban scene understanding [C]// Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE Computer Society, 2016:3213-3223.

[26] XIAO Yatie, PUN Chi-Man. Improving adversarial attacks on deep neural networks via constricted gradient-based perturbations[J]. Information Sciences, 2021,571:104.

[27] XIAO Yatie, PUN Chi-Man, LIU Bo. Adversarial example generation with adaptive gradient search for single and ensemble deep neural network[J]. Information Sciences, 2020,528:147.