基于压缩感知的神经网络实时综合防御策略

王 佳[1] 张扬眉[2),3)] 苏武强[1] 罗成文[1] 吴 超[1] 林秋镇[1] 李坚强[1]

[1)](深圳大学计算机与软件学院 深圳 518060)

[2)](中国科学院大学 北京 100049)

[3)](中国科学院计算技术研究所计算机体系结构国家重点实验室 北京 100190)

摘 要 近年来，基于深度神经网络的视觉识别模型因其在准确率、成本及效率等方面的优势而广泛应用于自动驾驶、工业检测及无人机导航等领域. 而深度神经网络自身易受数字域或物理域对抗样本攻击导致模型误判，因此其在无人驾驶等具有强鲁棒性、高实时性要求的场景中部署和应用可能为系统引入新的风险. 现有的防御方案在增强模型鲁棒性的同时往往造成准确率明显下降，且往往不能对像素攻击和补丁攻击均提供较强防御能力. 因此，设计一种精度高且对多类对抗攻击均具有强鲁棒性的实时综合防御策略成为深度神经网络视觉方案落地应用的关键. 本文提出一种基于压缩感知的神经网络实时综合防御策略 ComDCT，首先构建图像压缩感知压缩域与其稀疏离散余弦系数之间的映射神经网络，并将网络输出的离散余弦系数通过离散余弦逆变换恢复为去除对抗性扰动的图像作为分类器输入，以降低对抗样本攻击成功率. 其次，本文提出通过引入分类损失进一步提升防御策略的综合性能，并根据防御者是否掌握分类模型参数结构等信息分析讨论并验证了黑盒、白盒两种防御模式下引入分类损失的有效性. 相比于 ComDefend、MF、TVD、LRR 等多种防御方法，本文提出的基于压缩感知的神经网络实时综合防御策略在白盒防御模式下防御性能综合指标 PDA 在 LISA、SVHN 数据集上分别提升 11.88%、7.01%以上，黑盒防御模式下分别提升 9.25%、6.7%以上.

关键词 深度神经网络；对抗防御；压缩感知；无人驾驶

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2023.00001

Compressive Sensing Based Real-Time Comprehensive Defense Strategy for Neural Networks

WANG Jia[1)] ZHANG Yang-Mei[2),3)] SU Wu-Qiang[1)] LUO Cheng-Wen[1)]
WU Chao[1)] LIN Qiu-Zhen[1)] LI Jian-Qiang[1)]

[1)](*Colledge of Computer Science and Software Engineering, Shenzhen University,* 518060)

[2)](*University of Chinese Academy of Science, Beijing* 100049)

[3)](*State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Science, Beijing* 100190)

Abstract In recent years, Deep Neural Networks (DNN) have been widely applied in visual classification tasks in fields such as autonomous driving, industrial detection and drone navigation, mainly due to their advantages in accuracy, cost and efficiency. However, despite of these preponderance, deep neural networks are reported to be vulnerable to adversarial examples which could be generated either

收稿日期：2021-09-16；在线发布日期：2022-05-25. 本课题得到国家自然科学基金联合基金重点项目（U1713212）、国家重点研发项目（2020YFA0908700）、国家自然科学基金（61806130，6197071246，62002338）、广东省基础与应用基础研究基金项目（2021A1515011153）、“珠江人才计划”引进创新创业团队项目（2019ZT08X603）、深圳市科技创新委项目-稳定支持（面上项目20200805142159001）、深圳市重点项目（R2020A045）资助. **王 佳**，博士，助理教授，主要研究领域为机器学习鲁棒性、隐私保护、物联网安全. E-mail: jia.wang@szu.edu.cn. **张扬眉**，博士研究生，主要研究领域为人工智能. **苏武强**，硕士研究生，主要研究领域为压缩感知和网络安全. **罗成文**，博士，副教授，主要研究领域为移动端及普适计算和物联网相关信息安全. **吴 超**，博士，现从事博士后工作，主要研究领域为人工智能、嵌入式系统、存储架构、文件系统. **林秋镇**，博士，副教授，主要研究领域为优化算法等. **李坚强**（通信作者），博士，教授，主要研究领域为嵌入式系统和物联网. E-mail: lijq@szu.edu.cn.

digitally or physically. Noise images with intentionally crafted adversarial invisible or visible but inconspicuous perturbations could fool the classifier to make incorrect yet confident misclassifications. Hence, the deployment of such models in scenarios where robustness is a critical demand would introduce the system potential security risk. Existing defense strategies usually lead to a drop in test accuracy. And these algorithms are typically designed for defending against either pixel adversarial attacks or patch adversarial attacks in a dedicated manner, and their defensive capability usually does not translate to the other. Furthermore, when it's applied in real-time safety scenarios like autonomous driving, decision latency is required to be imperceptible, which makes many defensive algorithms far from a solution. Therefore, designing practical real-time comprehensive defense strategies for DNNs against a variety of adversarial attacks is of paramount of importance to its application, as well as represents a critical machine learning challenge. This paper attempts to address the problem of robustness of DNN-based visual classifiers against various adversarial examples by proposing a Compressive Sensing (CS) based defensive strategy combined with Discrete Cosine Transform (DCT), doomed ComDCT. ComDCT works in the compress-DCT-IDCT way to remove the adversarial perturbations from the input and then feed the denoised image to the classifier for inference. Specifically, to achieve this goal, ComDCT firstly train a neural network to learn the mapping from the measurements of the image to its sparse discrete cosine coefficients. And through inverse discrete cosine transform, cleaned images could be conveniently restored from the obtained DCT coefficients. To further improve the comprehensive performance of the defense scheme, we also suggest the introduction of the classification loss to optimize the compress-restore network. Finally, to demonstrate the efficacy of the proposed defensive strategy, intensive experiments have been conducted on two commonly used datasets, LISA and SVHN. For the purpose of achieving a comprehensive assessment, the adversarial examples used are generated using multiple attacks, including Fast Gradient Sign Method (FGSM), Carlini Wagner (CW-L2), Localized and Visible Adversarial Noise (LaVAN) as well as sticker attacks. And considering the difference of the adversaries' knowledge on the classification models, we give the performance evaluation, comparison as well as analysis in both white-box and black-box settings. Empirical results showed that, compared with other state-of-the-art defensive strategies in terms of ComDefend, MF, TVD, LRR and so on. Specifically, under white-box setting, ComDCT obtains at least 11.88% superiority in comprehensive performance indicator for LISA and 7.01% for SVHN. With the introduction of the classification loss in optimization, even under the black-box setting, the proposed scheme ComDCT still achieves at least 9.25% higher on the LISA dataset and 6.7% higher on the SVHN, which further confirms its advantages in alleviate the adversarial effects and improving the robustness of visual classifiers constructed on neural networks.

Keywords deep neural networks; adversarial defense; compressive sensing; autonomous driving

1 引 言

基于深度神经网络(Deep Neural Networks，DNN)的视觉识别算法为无人驾驶系统提供了一种可精确感知、充分理解车辆周围环境以及时响应的有效方式. 与传统的基于人工特征的识别算法相比，基于深度神经网络的识别分类模型在准确率和效率上所具有的优势使其在智能交通设计中得到了广泛的应用.

然而，最近的研究表明，基于深度神经网络的识别分类模型容易受到攻击从而导致模型失效或性能大幅度地降低. 因而，基于深度神经网络的识别分类模型在关键安全场景中的部署和应用可能会给系统带来新的潜在风险. 例如，在无人驾驶汽车系统中，交通标志识别分类模型失效或误判将会使行人和车辆处于危险之中，并可能会导致严重事故. 典型的攻击策略包括在样本中添加不可见或可见但不显眼的扰动来制作对抗样本以欺骗模型[1-2]. 前者

称为像素攻击，主要通过利用不同优化策略极小化地改变图像各个位置的像素值以向原始图像注入难以察觉的对抗噪声来实现；后者被称为补丁攻击，是通过添加“自然”的可见补丁来覆盖图像的突出特征[3]或设计可见仅限于图像局部位置的对抗性补丁来覆盖图像[4]而达到攻击目的.

这些攻击策略通常针对一个攻击者完全了解的深度神经网络模型进行攻击以制作对抗样本，并假设攻击者可直接将这些对抗样本作为底层机器学习识别器的输入. Eykholt 等人[5]在物理世界中实现了针对基于深度神经网络道路标志识别模型的贴纸攻击(Sticker Attack). Sticker 攻击属于补丁攻击，通过在物理世界真实的交通标志的表面上粘贴黑白贴纸，可以成功地误导模型将停车标志识别为限速 45 标志. 通过固定位置拍照、在行驶的汽车上拍摄获得视频帧等方式都可以从 Sticker 攻击生成的对抗性交通标志得到对抗性图像，Eykholt 等人的研究[5]表明在实验室环境下利用这些对抗性图像对识别模型进行攻击的成功率分别高达 100%和 84.8%. 显然，当对抗性交通标志被视为交通参与者时，这些攻击策略对可靠的无人驾驶汽车系统的设计提出了挑战.

为增强模型对于攻击策略的鲁棒性，研究者针对基于深度神经网络的视觉识别分类模型提出了多种有效的对抗防御技术[6-10]，包括对抗性训练、输入或特征随机化、去噪和可证明性防御策略[10]等. 然而，现有的针对像素攻击的大多数防御方法都缺乏针对补丁攻击的防御能力，例如 JPEG[6]、Median Filter (MF)[7]和 Total Variation Denoising (TVD)[6]. 反之亦然，针对补丁攻击的大多数防御策略，例如 Local Region Reconstruction(LRR)[9] 和 Local Gradients Smoothing (LGS)[8]，则缺乏对像素攻击的防御能力. 此外，相对于未加防护的深度神经网络模型，防御策略的引入往往需要额外的运行时间，考虑到无人驾驶汽车等应用场景的实时性要求[11-12]，对其设计实现提出了更高的挑战. 综上，设计一种对补丁攻击和像素攻击都具有较强防御能力且满足实时性的综合防御策略对于推动深度神经网络在关键领域落地应用具有重要意义.

为此，本文提出了一种基于压缩感知的神经网络实时综合防御策略 ComDCT，首先基于神经网络构建图像压缩感知压缩域与其稀疏离散余弦系数之间的映射关系，并将网络输出的离散余弦系数通过离散余弦逆变换恢复待测图像以去除对抗性扰动，以该图像为分类器输入进行降低对抗样本攻击成功率. 其次，本文提出通过引入分类损失进一步提升防御策略的综合性能，并评估了黑盒、白盒两种防御模式下 ComDCT 对像素攻击和补丁攻击的综合防御性能，主要贡献包含以下两个方面：

（1）提出图像压缩感知域与离散余弦变换域双域映射去噪方式，与 ComDefend、MF、TVD、LRR 等现有防御方案相比，可有效提升分类模型对像素攻击、补丁攻击等多种类对抗样本综合防御成功率，增强模型鲁棒性；

（2）提出引入分类器分类损失优化双域映射去噪网络，分析并通过实验验证了所提方案在黑盒、白盒两种防御模式下均可大幅度地提高防御策略综合性能，提升模型可用性.

本文的章节分布主要如下. 第二章介绍了相关的背景知识及最新相关工作. 第三章对本文所提出的基于压缩感知的神经网络实时综合防御策略进行了系统化阐述. 第四章通过大量对比实验验证并分析了本文所提出的神经网络综合防御策略在提升模型鲁棒性、可用性方面的优势. 最后第五章对本文进行总结及展望.

2　相关工作

自从 Szegedy 等人[13]发现深度神经网络模型易受对抗样本攻击而失效以来，研究者在如何生成对抗样本[1-5]以及如何增强对于这些对抗样本的鲁棒性[6-10]等方面均展开了一系列研究. 在本节中，本文对相关攻击策略以及防御策略进行总结并对本文使用的压缩感知进行简单介绍.

2.1　攻击策略

本文将现有的对抗攻击策略划分为两类：数字域的攻击策略和物理域的攻击策略.

2.1.1　数字域的攻击策略

Szegedy 等人[13]率先开展了通过生成数字域的对抗样本以实现对深度神经网络模型的攻击的研究. Szegedy 等人[13]通过使用 L-BFGS 方法在数字图像上引入像素级对抗性扰动，成功地导致深度神经网络模型的识别错误. Goodfellow 等人[2]提出了一种非目标攻击策略 FGSM，该策略沿着梯度的最陡方向对图像所有像素执行一次更新以有效地将分类结果导向错误类别. Kurakin 等人[14]提出了 Basic Iterative Attack (BIM)攻击策略，通过使用迭代更新取代一次更新，进一步提高 FGSM 的攻击性能. 受到 BIM 的启发，Dong 等人[15]将动量记忆引入 BIM 的迭代更新过程并以此提出一种新的攻击策略 MI-FGSM. Carlini 和 Wagner 在其工作[1]中提出一种基于优化的像素攻击方式 CW，同时兼顾攻击成功率

的提高和扰动的降低，当使用 L2 距离度量时，本文中记为 CW-L2. Localized and Visible Adversarial Noise (LaVAN) 攻击由 Karmon 等人在工作[4]中提出，其基本攻击思路为通过只在图像局部位置引入可见的扰动来欺骗深度神经网络模型，从而实现攻击. 此外，典型数字域的攻击策略还包括 DeepFool[16]、Distributionally Adversarial Attack (DAA)[17]等.

2.1.2 物理域的攻击策略

与数字域的攻击策略不同，物理域的攻击策略旨在通过粘贴打印得到的对抗性扰动等手段为现实世界的对象创建鲁棒的对抗样本. Parkhi 等人[18]打印了一副带有对抗性补丁的眼镜，并成功地误导了进行面部识别的深度神经网络模型. Athalye 等人[19]则通过创建带有对抗性扰动物体的 3D 打印复制品，实现了物理域的对抗攻击. 而 Kong 等人[20]提出一种新的攻击策略 PhysGAN，该策略可以生成任意路边交通/广告标志相对应的物理世界对抗样本，并不断误导无人驾驶车辆转向. 另外，Duan 等人[21]提出了一种基于激光束的新型物理域攻击方式，通过利用激光束扭曲照明可以使拍摄得到的图像具有对抗性.

2.2 防御策略

现有的防御策略大致可以分为以下四类：检测对抗样本、修改训练过程、基于随机化的策略和预处理及去噪.

2.2.1 检测对抗样本

对抗样本检测是一种直接简单的对抗样本防御方法. Metzen 等人[22]提出通过训练一个进行二分类的深度神经网络模型作为检测器可以将良性样本和对抗样本区分开来. Meng 等人[23]则提出了一种独立于攻击策略的防御策略，该防御策略包括一个或多个独立的检测器网络和一个重建器网络，其中检测器网络通过学习良性样本的流形来区分良性样本和对抗样本，重建器网络则将难以检测到的对抗样本重建以靠近良性样本的流形，减轻对抗性扰动的影响. Xu 等人[7]提出通过比较受保护分类模型对原始图像和对应特征压缩后的图像的预测来实现对抗样本的检测.

2.2.2 修改训练过程

对抗性训练(Adversarial Training)也是防御对抗性攻击的一种直观思路. 对抗性训练的基本思想是在深度神经网络模型的训练过程中包含对抗样本，以提高其可靠性和鲁棒性. Goodfellow 等人[2]提出了基于 FGSM 对抗性训练的防御策略，通过使用训练数据和 FGSM 生成的对抗样本一起训练神经网络来提高鲁棒性. 类似地，Madry 等人[24]额外使用 PGD 生成的对抗样本训练神经网络. Tramèr 等人[25]提出集成对抗性训练(Ensemble Adversarial Training, EAT)，其用于训练的对抗样本来源于针对不同分类模型的攻击，以提高对由黑盒攻击策略生成的对抗样本的鲁棒性.

2.2.3 基于随机化的策略

基于深度神经网络模型通常对随机扰动具有鲁棒性这一事实，研究者提出采用随机化方案来减轻对抗性扰动对其影响. Xie 等人[26]提出随机调整图像大小及填充以提高神经网络的鲁棒性. Liu 等人[27]则提出随机自集成(Random Self-Ensemble, RSE)的随机加噪方案，在输入到神经网络之前向对抗性图像的各个像素添加随机噪声. Dhillon 等人[28]提出随机激活修剪防御策略(Stochastic Activation Pruning, SAP)以提高深度神经网络模型对于对抗样本的鲁棒性.

2.2.4 预处理及去噪

在此类防御策略中，主要通过分类前对输入图像进行预处理、去噪以减少对抗性扰动的影响. 基于中值滤波器(Median Filter, MF)的方法[7]在图像上使用滑动窗口将像素与其相邻像素的平均值取代其自身取值以平滑图像和滤除对抗性扰动. 该技术已用于针对对抗样本的多种防御机制. 考虑到具有较高对抗性扰动的图像往往在相邻像素值之间具有更大的差异等事实，去噪机制能够有效地去除这种对抗性扰动，因而也成为一类去除对抗性扰动的主流方法，如 Total Variation Denoising [6]. Liao 等人[29]提出了高层特征引导的去噪器(HGD)，使用良性样本的分类模型高层特征(如 logits)和去噪后的对抗样本的高层特征之间的差异作为损失函数训练去噪网络以避免误差放大效应. Xie 等人[30]提出了一种新的网络结构，可以对特征图进行去噪以提高分类模型的鲁棒性. Jia 等人[31]提出了一个基于端到端图像压缩模型的防御策略 ComDefend. ComDefend 由一个压缩神经网络和一个重建神经网络组成，其中压缩神经网络通过将 RGB 图像三通道的 24 位图压缩成 12 位图(每个通道被分配 4 位)以实现图像去噪，而重建神经网络则用于完成干净图像的高质量重建.

2.3 压缩感知

压缩感知(Compressive Sensing, CS)提供了一种简洁的信号采集机制，在满足一定条件的情况下，能够以远小于奈奎斯特采样定律要求的采样率对信

号进行采样. 从数学的角度来看，压缩感知试图基于线性观测值 $\boldsymbol{y}\in\mathbb{R}^m$ 来重建恢复信号 $\boldsymbol{x}\in\mathbb{R}^n(n>m)$：

$$\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}+\boldsymbol{\eta} \tag{1}$$

其中 $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ 是测量矩阵，也被称为压缩矩阵；$\boldsymbol{\eta}\in\mathbb{R}^m$ 代表测量噪声. 当信号 $\boldsymbol{x}$ 满足稀疏性且测量矩阵 $\boldsymbol{A}$ 满足一定条件如限定等距性(Restricted Isometry Property, RIP)时，传统的压缩感知恢复算法通过将信号重建过程转化为迭代地求解以下优化问题：

$$\min\|\hat{\boldsymbol{x}}\|_1 \quad \text{s.t.} \quad \boldsymbol{y}=\boldsymbol{A}\hat{\boldsymbol{x}} \tag{2}$$

可以从上述欠定方程组中近乎无损地得到恢复信号 $\hat{\boldsymbol{x}}$. 由于传统压缩感知恢复算法的迭代求解过程复杂耗时，Mousavi 等人[32]提出了一种基于神经网络的压缩感知恢复算法 DeepInverse 以快速地得到恢复信号 $\hat{\boldsymbol{x}}$. 然而，在许多应用中，信号 $\boldsymbol{x}$ 自身并不满足稀疏性，需要经过变换（即矩阵乘法）才能满足稀疏性：

$$\boldsymbol{s}=\boldsymbol{\Psi}\boldsymbol{x} \tag{3}$$

其中 $\boldsymbol{\Psi}\in\mathbb{R}^{n\times n}$ 代表使信号 $\boldsymbol{x}$ 满足稀疏性的变换矩阵，$\boldsymbol{s}\in\mathbb{R}^n$ 代表信号 $\boldsymbol{x}$ 经过变换得到的系数. 例如自然图像自身是非稀疏的，但是在经过离散余弦变换(Discrete Cosine Transform, DCT)、离散小波变换(Discrete Wavelet Transform, DWT)等处理以后就可以满足稀疏性. 在这种情况下，式(1)可以写为：

$$\boldsymbol{y}=\boldsymbol{A}\boldsymbol{\Psi}^{-1}\boldsymbol{s}+\boldsymbol{\eta} \quad \text{s.t.} \quad \boldsymbol{x}=\boldsymbol{\Psi}^{-1}\boldsymbol{s} \tag{4}$$

其中 $\boldsymbol{\Psi}^{-1}\in\mathbb{R}^{n\times n}$ 代表从系数 $\boldsymbol{s}$ 到信号 $\boldsymbol{x}$ 的逆变换矩阵. 类似地，当经过变换得到的 $\boldsymbol{s}$ 满足稀疏性且 $\boldsymbol{A}\boldsymbol{\Psi}^{-1}$ 满足 RIP 条件时，$\boldsymbol{s}$（也即 $\boldsymbol{x}$）可以被近乎无损地恢复.

3　基于压缩感知的实时综合防御策略

为了将基于深度神经网络的视觉识别算法安全地应用于无人驾驶等强鲁棒性、高实时性的场景中，本文提出基于稀疏离散余弦系数的综合防御策略 ComDCT，其既满足实时性又有强大的防御性能，在引入分类损失的情况下还可以更进一步提高防御策略的综合性能. ComDCT 的框架图如图 1 所示.

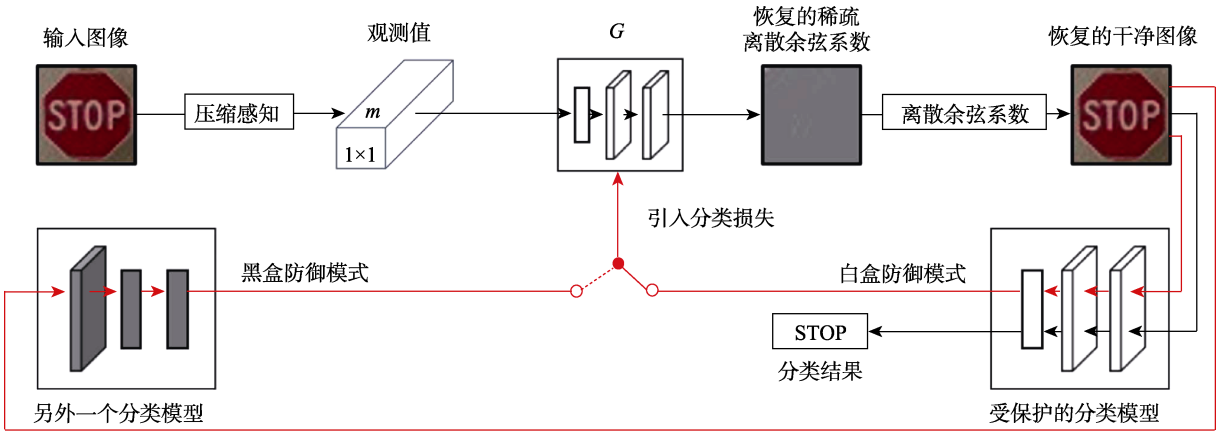

图 1　ComDCT 框架图（ComDCT 首先通过压缩感知将输入图像转换成观测值，再通过生成网络 G 从观测值得到恢复的稀疏离散余弦系数，最后通过离散余弦逆变换得到恢复的干净图像. 此外，根据对受保护分类模型的了解程度，在白盒、黑盒两种防御模式下各自通过引入分类损失优化生成网络 G，大幅度提高防御性能）

3.1　基于稀疏离散余弦系数的实时防御策略

近年来，基于压缩感知技术所设计的防御策略[33-34]得到了研究者的广泛关注，其基本思想是利用传统压缩感知恢复算法在使用对抗样本的观测值进行恢复的过程中去除对抗性扰动，从而实现对分类器模型的保护. 而传统的压缩感知恢复算法都需要迭代优化求解，如 Bafna 等人在其工作[33]中使用的 Iterative Hard Thresholding (IHT)方法、Dhaliwal 等人在其工作[34]中使用的 Basis Pursuit (BP)方法等，这使得此类方法计算量大且耗时，因而不适用于无人驾驶等高实时性要求的应用场景.

为了解决压缩感知恢复耗时的问题，Mousavi 等人[32]基于深度神经网络模型提出了一种新的压缩感知恢复算法——DeepInverse. DeepInverse 以原始图像 $\boldsymbol{x}$ 的观测值 $\boldsymbol{y}$ 为输入，重建图像 $\hat{\boldsymbol{x}}$ 为输出，通过训练一个特殊网络结构的深度神经网络来得到从观测值 $\boldsymbol{y}$ 到原始图像 $\boldsymbol{x}$ 的映射，训练损失为重建损失：

$$\begin{aligned}&\text{loss}=\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|_2^2\\&\text{s.t.}\quad \hat{\boldsymbol{x}}=G(\boldsymbol{y})=G(\boldsymbol{A}\boldsymbol{x})\end{aligned} \tag{5}$$

其中 G 代表神经网络. 训练过程耗时，但考虑为一次性成本而不影响实际使用时间，基于 DeepInverse 的防御策略满足无人驾驶汽车系统等应用的实时性要求.

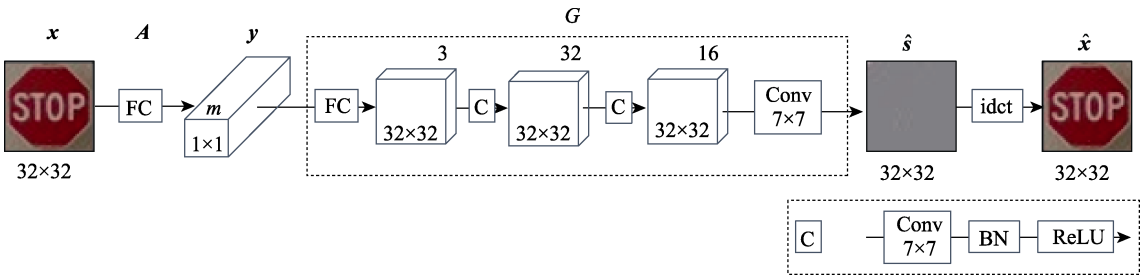

图 2 ComDCT 对图像的预处理及去噪流程图（ComDCT 首先通过测量矩阵 $\boldsymbol{A}$ 将输入图像 $\boldsymbol{x}$ 压缩成观测值 $\boldsymbol{y}$，再通过一个全连接层将观测值 $\boldsymbol{y}$ 的维度 m 扩充到输入图像 $\boldsymbol{x}$ 的维度 n，然后通过若干个卷积层得到恢复的稀疏离散余弦系数 $\hat{\boldsymbol{s}}$，最后通过离散余弦逆变换得到恢复的干净图像 $\hat{\boldsymbol{x}}$ ）

然而，本文发现在直接使用 DeepInverse 压缩感知恢复算法来取代传统耗时的恢复算法作为防御策略时，其防御性能偏低. 为了提升防御能力及提高综合性能，本文提出基于深度神经网络和稀疏系数的压缩感知恢复算法，训练神经网络学习从观测值 $\boldsymbol{y}$ 到原始图像变换得到的稀疏系数 $\boldsymbol{s}$ 的映射，在得到恢复的稀疏系数 $\hat{\boldsymbol{s}}$ 后再通过逆变换得到恢复的近似图像 $\hat{\boldsymbol{x}}$，而非如 DeepInverse 直接训练学习从观测值 $\boldsymbol{y}$ 到原始图像 $\boldsymbol{x}$ 的映射. 离散余弦变换在压缩感知领域被广泛使用[33-34]，因此本文选用离散余弦变换作为图像稀疏变换基. 其他图像稀疏变换基也适用于本文提出的框架，如上面提到的离散小波变换和离散傅里叶变换. 此时的重建损失为

$$\begin{aligned} &\mathcal{L}_{mse} = \| \hat{\boldsymbol{x}} - \boldsymbol{x} \|_2^2 = \| \mathrm{idct}(\hat{\boldsymbol{s}}) - \boldsymbol{x} \|_2^2 \\ &\text{s.t.} \quad \hat{\boldsymbol{s}} = G(\boldsymbol{y}) = G(\boldsymbol{A}\boldsymbol{x}) \end{aligned} \tag{6}$$

其中 idct 代表离散余弦逆变换. 引入自然图像的离散余弦系数具有稀疏性的先验知识作为正则项，此时的训练损失为

$$\begin{aligned} &loss = \mathcal{L}_{mse} + \lambda_r \mathcal{L}_{reg} \\ &\text{s.t.} \quad \mathcal{L}_{\mathrm{reg}} = \| \hat{\boldsymbol{s}} \|_1 \end{aligned} \tag{7}$$

其中 λ_r 代表相对于重建损失 $\mathcal{L}_{mse}$ 正则项损失 $\mathcal{L}_{reg}$ 的重要性. G 训练完成以后，对图像的预处理及去噪流程如图 2 所示. 相比于 DeepInverse 从神经网络输出直接得到恢复图像 $\hat{\boldsymbol{x}}$，虽然防御策略使用本文提出的压缩感知恢复算法需要额外通过离散余弦逆变换从神经网络输出的离散余弦系数 $\hat{\boldsymbol{s}}$ 得到恢复图像 $\hat{\boldsymbol{x}}$，但是离散余弦逆变换所需的时间相比于神经网络从输入到输出所需的时间可忽略不计，因此基于本文提出的压缩感知恢复算法的防御策略仍然满足无人驾驶汽车系统等应用的实时性要求. 实验结果表明相比于直接学习从观测值 $\boldsymbol{y}$ 到图像 $\boldsymbol{x}$ 的映射，训练神经网络学习从观测值 $\boldsymbol{y}$ 到图像的稀疏离散余弦系数 $\boldsymbol{s}$ 的映射能在对良性样本的分类性能影响小的情况下大幅度提高防御性能，从而提高综合性能.

近年来，为了提高信号恢复质量，一些工作[35-36]提出利用深度神经网络模型训练一个更加鲁棒的测量矩阵 $\boldsymbol{A}$. 受到这些工作的启发，本文提出在网络结构前面加一个没有偏置没有激活函数的全连接层来代表测量矩阵 $\boldsymbol{A}$，不固定全连接层的参数，以原始图像 x 为输入，使用训练损失(6)来训练一个更加鲁棒的测量矩阵 $\boldsymbol{A}$. 训练完成以后，使用训练得到的测量矩阵的防御策略对图像的预处理及去噪的流程跟使用通用的测量矩阵的一样，如图 2 所示，因此所需的运行时间一样，满足实时性. 相比于使用通用的测量矩阵，使用训练得到的测量矩阵能得到更好地恢复图像 $\hat{\boldsymbol{x}}$，因此对良性样本的分类性能的影响更小，但同时其去噪能力也被减弱，导致防御性能降低. 实验结果表明在使用训练损失(6)的情况下，针对本文使用的两个数据集，使用通用的测量矩阵和使用训练得到的测量矩阵具有相当的综合性能.

3.2 引入分类损失

本文提出的上述基于稀疏离散余弦系数的实时防御策略独立于保护的深度神经网络分类模型，因此对包括不采用梯度下降方式更新权重参数的分类模型在内的多种分类模型具有鲁棒的综合性能. 然而，正因为这种通用性，对于已知网络结构和参数且只含有可导操作的特定分类模型，它们通常不是最好的防御方法. 为了解决这个问题，本文提出在训练生成神经网络 G 的过程中引入分类损失（在本文中指分类模型对重建的去噪图像进行分类的损失），可使重建图像中含有更多的、受保护分类模型可识别的分类特征，从而实现对所保护的特定分类模型的专门优化，即分类损失的引入可将输入图像重建成更容易被分类模型正确分类的图像.

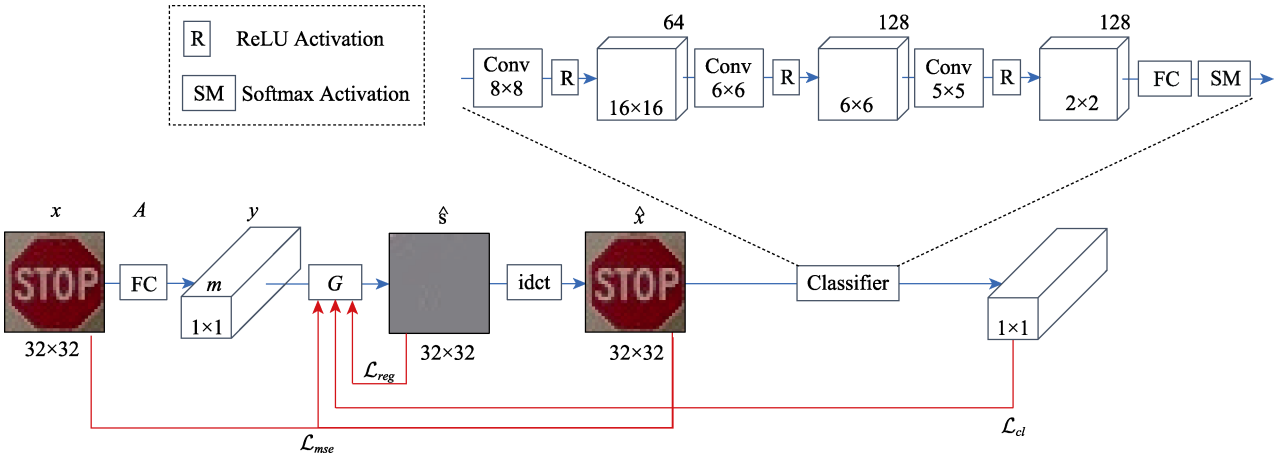

图 3　引入分类损失后 ComDCT 的训练框架（神经网络 G 的训练损失由三部分组成，分别是输入图像 $\boldsymbol{x}$ 与恢复的干净图像 $\hat{\boldsymbol{x}}$ 之间的重建误差 $\mathcal{L}_{mse}$、恢复的离散余弦系数 $\hat{\boldsymbol{s}}$ 的稀疏性先验知识带来的正则项损失 $\mathcal{L}_{reg}$ 和分类器对恢复的干净图像 $\hat{\boldsymbol{x}}$ 的分类损失 $\mathcal{L}_{cl}$）

引入分类损失对于 ComDCT 综合防御能力的提升可从两个方面进行分析. 首先，防御策略的引入可能无意中破坏原有良性样本的分类特征，因此用户通常需要在防御性能和分类精度之间进行权衡. 而通过引入分类损失，可使重建图像保留越多的分类器可识别的分类特征，从而提升分类模型对重建的良性样本的分类性能. 其次，引入分类损失还可以在重建图像上强化一些分类特征，在去噪的基础上将对抗样本转换为更易于正确分类的重建图像，进一步减小对抗性干扰的影响，提高防御性能.

基于稀疏离散余弦系数的实时防御策略引入分类损失后，训练损失对应地修改为

$$\begin{aligned} &\text{loss} = \mathcal{L}_{mse} + \lambda_r \mathcal{L}_{reg} + \lambda_c \mathcal{L}_{cl} \\ &\text{s.t.}\ \ \mathcal{L}_{cl} = -\log p_\ell(\hat{\boldsymbol{x}}) \end{aligned} \tag{8}$$

其中 λ_r、λ_c 分别代表正则项损失 $\mathcal{L}_{reg}$、分类损失 $\mathcal{L}_{cl}$ 相对于重建损失 $\mathcal{L}_{mse}$ 的重要性，$p_\ell(\hat{\boldsymbol{x}})$ 代表分类模型对重建图像 $\hat{\boldsymbol{x}}$ 正确分类的概率.

引入分类损失后，防御策略的训练框架如图 3 所示. 值得注意的是，由于分类损失的引入只对训练过程有影响，而模型实际应用阶段防御策略对图像的预处理及去噪的流程与没有引入分类损失情况下一致. 因此防御策略设计中引入分类损失并不增加模型应用处理时间，不影响模型实时性.

然而，当受保护分类模型含有不可导操作时，受保护分类模型对重建图像的分类损失的引入无法通过梯度下降方式更新生成神经网络 G 的参数，从而无法使重建图像含有更多的分类特征，即无法提高防御策略的综合性能. 受到黑盒攻击策略制作的对抗样本具有迁移性（攻击者针对一个分类模型制作的对抗样本仍然可以成功地误导其他网络结构和参数都不相同的分类模型）这一事实的启发，本文提出通过训练一个只含有可导操作的影子分类模型并引入该分类模型对重建图像的分类损失以使重建图像含有更多的分类特征和提高防御策略的综合性能.

类似地，当防御者不了解受保护分类模型和无法得到受保护分类模型对重建图像的分类损失时，防御策略也可以通过额外训练一个分类模型来引入分类损失以提高其综合性能. 本文根据防御者对其保护的分类模型的了解程度划分了引入分类损失的两种情况：白盒防御模式——完全了解受保护分类模型的网络结构及参数，可以直接引入该分类模型对重建图像的分类损失来优化训练神经网络 G；黑盒防御模式——不了解受保护分类模型的网络结构及参数，只能通过引入其他分类模型对重建图像的分类损失来优化训练神经网络 G. 实验结果表明：黑盒防御模式下引入与受保护分类模型的网络结构及参数都不相同的分类模型的分类损失后，重建图像$\hat{x}$上更多的分类特征具有迁移性，可以被受保护分类模型识别，提高防御策略的综合性能.

分类损失的引入具有鲁棒性，即便是在黑盒防御模式下，无论是使用通用的测量矩阵还是使用训练得到的测量矩阵，均可有效提升基于稀疏离散余弦系数的实时防御策略的综合性能，还可以提高其他基于深度神经网络的预处理和去噪防御策略的综合性能，如 Jia 等人[31]提出的 ComDefend. 基于稀疏离散余弦系数的实时防御策略在黑盒防御模式下，相比于使用通用的测量矩阵，使用训练得到的测量矩阵可以在更小地影响重建图像$\hat{x}$质量和对良性样本的分类性能的情况下，使重建图像$\hat{x}$含有更多的

分类特征，以得到更加高的防御性能及综合性能.

4 实验结果与分析

本章节对本文所用的数据集、评价指标以及实验设置进行介绍，并对实验结果进行分析.

4.1 数据集介绍及评价指标

本章节对本文所用的数据集和评价指标进行介绍.

4.1.1 数据集介绍

本文使用两个数据集来评价防御策略的综合性能：LISA 交通标志数据集[37]、街景门牌号码数据集(Street View House Numbers, SVHN)[38].

LISA 交通标志数据集由 7855 张大小为 32×32 的 RGB 图像组成，包含有 47 种美国交通标志. 本文对其随机地划分出 5499 张作为训练集，785 张作为验证集，1571 张作为测试集. 本文使用 Zhao 等人[9]使用的 LISA 分类模型作为受保护的分类模型，将测试集图像中分类正确的 1287 张图像作为良性图像集以评价防御策略对良性样本分类性能的影响. 为了公平客观地评价防御策略的防御性能，本文首先选用 Zhao 等人[9]评估其提出的防御策略 LRR 时使用的 Sticker 攻击策略，针对 LISA 数据集制作的、能够成功误导分类模型的 90 张图像作为 Sticker 对抗图像集，再使用 FGSM 攻击策略($\epsilon = 0.07$)和 CW-L2 攻击策略($\kappa = 0$)对测试集图像添加对抗性扰动，然后将成功误导分类模型的对抗性图像收集起来得到含有 1064 张图像的 FGSM 对抗图像集和含有 1239 张图像的 CW-L2 对抗图像集.

SVHN 数据集由大小为 32×32 的 RGB 图像组成，包含从 0 到 9 的 10 种数字. 本文从原先的 73257 张训练集图像划出 10000 张作为验证集，剩下的 63257 张作为训练集，使用原先的 26032 张测试集图像作为测试集. 本文首先使用训练集图像和验证集图像训练一个 SVHN 分类模型作为受保护的分类模型，再将测试集图像中分类正确的 23160 张图像作为良性图像集，然后使用 FGSM 攻击策略($\epsilon = 0.03$)和 CW-L2 攻击策略($\kappa = 0$)对测试集图像添加对抗性扰动，最后将成功误导分类模型的对抗性图像收集起来，得到含有 19222 张图像的 FGSM 对抗图像集和含有 17894 张图像的 CW-L2 对抗图像集. 由于缺乏公开的、针对 SVHN 数据集制作的补丁类攻击对抗图像集，且 Sticker 攻击属于物理域的攻击策略，制作 Sticker 对抗图像集难度大，因此，受限于实验场地及条件，本文使用另一种较易实现的补丁攻击 LaVAN(对抗性补丁大小为 6×6)制作对抗图像集来评估不同防御策略在 SVHN 数据集上的防御性能，最终得到含有 9332 张图像的 LaVAN 对抗图像集.

4.1.2 评价指标

在评价防御策略的性能时，需要同时考虑其对对抗样本的防御性能的提高和其对良性样本的分类性能的影响. 因此，Zhao 等人[9]提出一个新指标 *PDA* (Post-Defense Accuracy)来评价一个防御策略的综合性能. 当攻击策略对分类模型的误导成功率为 100%时，*PDA* 的定义为

$$PDA = (1-\pi_A)(CA-CD)+\pi_A DR \tag{9}$$

其中 π_A 代表分类模型的输入图像含有对抗性扰动的先验概率，*CA* 代表良性图像集没有经过防御策略处理时的分类准确率，*CD* 代表良性图像集经过防御策略处理后的分类准确率的降幅，而 *DR* 是防御成功率，代表成功误导分类模型的对抗性图像经过防御策略处理后的分类准确率.

然而 Zhao 等人[9]只考虑了一种攻击策略，但是本文对两个数据集都各自使用了三种攻击策略制作对应的对抗图像集，因此本文可以在 LISA 数据集上得到 DR_{FGSM}、DR_{CW-L2} 和 $DR_{Sticker}$ 以及在 SVHN 数据集上得到 DR_{FGSM}、DR_{CW-L2} 和 DR_{LaVAN}. 本文假设对抗性图像来自三种攻击策略的概率相同，重新定义针对 LISA 数据集的 *DR* 为

$$DR = \frac{DR_{FGSM}+DR_{CW-L2}+DR_{\text{Sticker}}}{3} \tag{10}$$

重新定义针对 SVHN 数据集的 *DR* 为

$$DR = \frac{DR_{FGSM}+DR_{CW-L2}+DR_{LaVAN}}{3} \tag{11}$$

由于 $CA-CD$ 代表良性图像集经过防御策略处理后的分类准确率，本文定义指标 *DBA* (Defensed Benign Accuracy)来取代 $CA-CD$ 以简化公式(9)：

$$DBA = CA-CD \tag{12}$$

通过结合公式(9)、公式(10)和公式(12)，同时考虑防御性能和分类精度，在 LISA 数据集上进行评估时，防御策略的综合性能指标 *PDA* 被定义为

$$PDA = (1-\pi_A)DBA + \pi_A \frac{DR_{FGSM}+DR_{CW-L2}+DR_{\text{Sticker}}}{3} \tag{13}$$

类似地，通过结合公式(9)、公式(11)和公式(12)，在 SVHN 数据集上进行评估时，防御策略的综合性能指标 *PDA* 被定义为

$$PDA = (1-\pi_A)DBA + \pi_A \frac{DR_{FGSM}+DR_{CW-L2}+DR_{LaVAN}}{3} \tag{14}$$

与 Zhao 等人的工作[9]一样，在本文所有的实验评价中π_A被设置为 0.5.

此外，为了将防御策略应用到无人驾驶汽车等高实时性场景中，还需要考虑到其运行时间以对其实时性进行评估. 为了将交通标志识别技术应用于高实时性的无人驾驶汽车场景，Yang 等人[11]提出了一种由检测和分类模块组成的实时交通标志识别系统. 该系统先用检测模块从 1360×800 大小的原始图像中检测出交通标志，再将检测到的交通标志分割出来并进行放缩，最后用分类模块对放缩后大小为 32×32 的交通标志图像进行分类以识别出是哪种交通标志. 该系统处理一张原始图像的平均时间约为 165 ms，其中检测耗时约为 162 ms，分类耗时约为 3 ms. 类似地，Shao 等人[12]也提出了一种由检测和分类模块组成的实时交通标志识别系统，此系统检测一张 1360×800 大小的图像的平均耗时约为 154 ms，分类一张 32×32 大小的图像的耗时约为 5 ms，处理一张原始图像的总时间约为 159 ms. 虽然采用防御策略对上述交通标志识别系统的分类模块进行保护往往需要额外的运行时间，但是当防御策略所需的时间为 10 ms 时，对于上述实时交通标志识别系统处理一张原始图像的总时间的影响很小(约从 159~165 ms 增加到 169~175 ms). 因此，在本文中防御策略满足实时性定义为该防御策略所需的时间不超过 10 ms.

4.2　实验设置

本文提出的基于压缩感知的实时综合防御策略(Compressive sensing Discrete Cosine Transform, ComDCT)的实现基于 Python2.7、TensorFlow 和开源代码[39]. ComDCT 的网络结构如表 1 所示. ComDCT-U 代表防御策略使用通用的测量矩阵，此时后缀加上“-U”，其中通用的测量矩阵选用随机高斯矩阵. 类似地，ComDCT-T 代表防御策略使用训练得到的测量矩阵，此时后缀加上“-T”. 当数据集的类别数为 c 时，本文所用的分类模型的网络结构如表 2 所示，两个数据集使用的分类模型的网络结构一致. ComDCT-W 代表防御策略在白盒防御模式下引入分类损失，此时后缀加上“-W”；ComDCT-B 代表防御策略在黑盒防御模式下引入分类损失，此时后缀加上“-B”.

本文中的所有涉及到训练深度神经网络模型的实验均采用梯度下降方式更新网络参数，且均使用 Adam 优化器[40]. 本文中的所有实验均基于 Ubuntu 18.04.2，使用 Quadro P5000 GPU 和主频为 2.40GHz 的 Intel(R) Xeon(R) Gold 6148 CPU.

表 1　基于压缩感知的实时防御策略 ComDCT 的网络结构

$\text{ComDCT}_{\text{LISA}}$	$\text{ComDCT}_{\text{SVHN}}$
3072×m 全连接层	3072×m 全连接层
m×307 全连接层	m×307 全连接层
3×7×7×32 卷积层	3×11×11×128 卷积层
32×7×7×16 卷积层	128×11×11×64 卷积层
16×7×7×3 卷积层	64×11×11×32 卷积层
	32×11×11×3 卷积层

表 2　分类模型的网络结构

白盒防御模式	黑盒防御模式
3×8×8×64 卷积层	3×5×5×32 卷积层
64×6×6×128 卷积层	2×2 最大池化层
128×5×5×128 卷积层	32×5×5×64 卷积层
512×c 全连接层	2×2 最大池化层
	4096×1024 全连接层
	随机失活层
	1024×c 全连接层

4.3　基于稀疏离散余弦系数的实时防御策略

本文在不同的λ_r及测量率(Measurement Rate, $MR=\frac{m}{n}$)下，使用 DeepInverse 恢复算法训练神经网络学习从观测值 y 到原始图像 x 的映射的防御策略与训练神经网络学习从观测值 y 到原始图像的稀疏离散余弦系数s的映射的 ComDCT 防御策略的性能对比如表 3 和表 4 所示，其中 DeepInverse 跟 ComDCT 使用一样的网络结构以公平比较. 无论是使用通用的测量矩阵还是使用训练得到的测量矩阵，在各个测量率，随着 λ_r 的不断增大，ComDCT 对良性图像分类能力的影响都越来越大，DBA越来越低. 但是在 λ_r 没有大到一定程度之前，随着 λ_r 的不断增大，ComDCT 去除对抗性扰动的能力和防御能力越来越强，DR越来越高，而且防御性能的增幅大于良性图像分类性能的降幅，从而导致 ComDCT 的综合性能越来越强，PDA越来越高. 值得注意的是，过大的 λ_r 会导致图像的重建质量低下，随着过大的 λ_r 的继续增大，ComDCT 一方面其防御性能的增幅逐渐减小，另一方面其良性图像分类性能的降幅逐渐增大，从而导致 ComDCT 的综合性能逐渐变弱. 与 DeepInverse 相比，在各个测量率，当 ComDCT 选取最优 λ_r 时，ComDCT 的综合性能PDA在 LISA 数据集上至少高 2.45%，在 SVHN 数据集上至少高 2.91%. 当 ComDCT 选取最优 λ_r 和最优测量率时，

– Guadarrama S, Silberman N. TensorFlow-Slim: A lightweight library for defining, training and evaluating complex models in TensorFlow, https://github.com/google-research/tf-slim 2019,6,29

表 3 不同 *MR* 和λ_r下防御策略 DeepInverse 与 ComDCT 在 LISA 数据集上的性能对比

防御策略	λ_r	*MR*=0.05			*MR*=0.10			*MR*=0.25		
		DBA	*DR*	*PDA*	*DBA*	*DR*	*PDA*	*DBA*	*DR*	*PDA*
DeepInverse-U	–	0.8866	0.4072	0.6469	0.9316	0.3750	0.6533	0.9479	0.2935	0.6207
ComDCT-U	0.0005	0.8376	0.4732	0.6554	0.8411	0.4851	0.6631	0.8500	0.4884	0.6692
	0.001	0.8446	0.4918	0.6682	0.8834	0.4872	0.6853	**0.8702**	**0.5203**	**0.6953**
	0.002	0.8411	0.5017	0.6714	0.8500	0.4996	0.6748	0.8664	0.5017	0.6840
	0.004	0.8419	0.4984	0.6702	0.8500	0.4984	0.6742	0.8718	0.4915	0.6816
DeepInverse-T	–	0.9829	0.3116	0.6473	0.9845	0.2273	0.6059	0.9876	0.2414	0.6145
ComDCT-T	0.1	**0.9798**	**0.4143**	**0.6970**	0.9899	0.3261	0.6580	0.9829	0.2916	0.6372
	0.2	0.9518	0.4369	0.6944	0.9744	0.3768	0.6756	0.9697	0.3173	0.6435
	0.4	0.8866	0.5036	0.6951	0.9448	0.4150	0.6799	0.9549	0.4185	0.6867
	0.8	0.6706	0.4063	0.5384	0.7560	0.4357	0.5959	0.7918	0.4268	0.6093

表 4 不同 *MR* 和λ_r下防御策略 DeepInverse 与 ComDCT 在 SVHN 数据集上的性能对比

防御策略	λ_r	*MR*=0.05			*MR*=0.10			*MR*=0.25		
		DBA	*DR*	*PDA*	*DBA*	*DR*	*PDA*	*DBA*	*DR*	*PDA*
DeepInverse-U	–	0.9313	0.3766	0.6539	0.9736	0.3331	0.6533	0.9863	0.2903	0.6383
ComDCT-U	0.1	0.9386	0.4042	0.6714	0.9686	0.3800	0.6743	0.9789	0.3716	0.6752
	0.2	0.9277	0.4337	0.6807	0.9589	0.4176	0.6882	0.9675	0.4091	0.6883
	0.4	0.8845	0.4815	0.6830	0.9187	0.4627	0.6907	**0.9332**	**0.4678**	**0.7005**
	0.8	0.7149	0.4649	0.5899	0.7712	0.4752	0.6232	0.7628	0.4741	0.6185
DeepInverse-T	–	0.9902	0.2687	0.6294	0.9921	0.2271	0.6096	0.9919	0.2595	0.6257
ComDCT-T	0.1	0.9874	0.3263	0.6568	0.9886	0.3105	0.6496	0.9893	0.3078	0.6486
	0.2	0.9812	0.3631	0.6721	0.9804	0.3473	0.6638	0.9831	0.3545	0.6688
	0.4	0.9560	0.4340	0.6950	0.9579	0.4325	0.6952	**0.9582**	**0.4329**	**0.6955**
	0.8	0.8188	0.5050	0.6619	0.8269	0.5072	0.6670	0.8247	0.5058	0.6652

ComDCT 的综合性能 *PDA* 比 DeepInverse 在 LISA 数据集上高 4.2%，在 SVHN 数据集上高 4.16%. 这些实验结果说明了本文提出的学习从观测值 y 到原始图像的稀疏离散余弦系数 s 的映射的 ComDCT 防御策略相对于使用 DeepInverse 恢复算法学习从观测值 y 直接到原始图像 x 的映射的防御策略的优越性.

相比于使用通用测量矩阵的 DeepInverse-U 和 ComDCT-U，使用训练得到的测量矩阵的 DeepInverse-T 和 ComDCT-T 都对良性图像分类能力的影响更小，*DBA* 更高，但是相对地其去除对抗性扰动的能力和防御能力更低，*DR* 更低. 最终针对本文使用的两个数据集，使用通用测量矩阵的 ComDCT-U 和使用训练得到的测量矩阵的 ComDCT-T 具有相当的综合性能.

ComDCT 跟其他防御策略的性能对比如表 5 和表 6 所示. 即使在不引入分类损失的情况下，ComDCT 的综合性能 *PDA* 在 LISA 数据集上相比于其他防御策略仍至少高 3.81%，在 SVHN 数据集上 ComDCT 的综合性能 *PDA* 也只比 MF 低，这再一次说明了本文提出的 ComDCT 防御策略的优越性.

表 7 中实验数据为通过对 1287 张大小为 32×32 的 LISA 良性图像集图像进行多次预处理及去噪得到的不同防御策略处理每张图像的平均运行时间. 实验结果表明，传统的基于压缩感知技术的对抗样本防御方案 IHT[33]及 BP[34]的平均处理时间分别为 76.36 ms、62174.19 ms，这是由于传统的压缩感知恢复算法都需要迭代优化求解，因而此类方法计算量大且耗时，不适用于无人驾驶等高实时性要求的应用场景. 而本文提出使用基于深度神经网络的压缩感知恢复算法作为防御策略，无论是使用本文提出的基于稀疏系数恢复算法的 ComDCT 还是直接使用 DeepInverse 恢复算法的防御策略，均提升了防御处理效率，解决了传统压缩感知防御策略中不满足实时性的问题. 此外，测量率 的变化对 ComDCT 和 DeepInverse 的处理时间几乎没有影响. 相比于 DeepInverse，尽管 ComDCT 中离散余弦逆变换需要

表 5　各个防御策略在 LISA 数据集上的性能对比

防御策略	*DBA*	DR_{FGSM}	DR_{CW-L2}	$DR_{Sticker}$	*DR*	*PDA*
ComDCT-T [$MR=0.05,\lambda_r=0.1$]	0.9798	0.1222	0.8095	0.3111	0.4143	0.6970
ComDCT-T-W [$MR=0.05,\lambda_r=0.1,\lambda_c=0.4$]	**0.9946**	**0.4323**	**0.9128**	**0.5556**	**0.6336**	**0.8141**
ComDCT-T-B [$MR=0.05,\lambda_r=0.1,\lambda_c=0.1$]	0.9930	0.3167	0.8644	0.5667	0.5826	0.7878
ComDCT-U [$MR=0.25,\lambda_r=0.001$]	0.8702	0.3069	0.7264	0.5278	0.5203	0.6953
ComDCT-U-W [$MR=0.25,\lambda_r=0.001,\lambda_c=0.2$]	**0.9782**	**0.4648**	**0.8906**	**0.6111**	**0.6555**	**0.8169**
ComDCT-U-B [$MR=0.25,\lambda_r=0.001,\lambda_c=0.008$]	0.9029	0.3534	0.7712	0.5389	0.5545	0.7287
ComDefend[31]	0.9951	0.0201	0.4609	0.0481	0.1764	0.5857
ComDefend-W [$\lambda_c=0.4$]	0.9953	0.3158	0.8846	0.4556	0.5520	0.7737
ComDefend-B [$\lambda_c=0.8$]	0.9946	0.0536	0.6312	0.1000	0.2616	0.6281
JPEG[6] [quality=10]	0.9891	0.0207	0.4754	0.0889	0.1950	0.5921
TVD[6] [weight=3]	0.9215	0.0921	0.6852	0.2333	0.3369	0.6292
MF[7] [window=5]	0.9806	0.0968	0.7934	0.1111	0.3338	0.6572
LGS[8]	0.8314	0.2481	0.4988	0.2889	0.3453	0.5883
LRR[9]	0.9394	0.1278	0.5634	0.4000	0.3637	0.6516
DeepInverse	0.9316	0.1410	0.6061	0.3778	0.3750	0.6533
IHT[33]	1.0000	0.0047	0.1340	0.0111	0.0499	0.5250
BP[34]	0.6830	0.1118	0.3430	0.2000	0.2183	0.4506

表 6　各个防御策略在 SVHN 数据集上的性能对比

防御策略	*DBA*	DR_{FGSM}	DR_{CW-L2}	$DR_{Sticker}$	*DR*	*PDA*
ComDCT-T [$MR=0.25,\lambda_r=0.4$]	0.9582	0.1772	0.8056	0.3158	0.4329	0.6955
ComDCT-T-W [$MR=0.25,\lambda_r=0.4,\lambda_c=0.1$]	**0.9238**	**0.5181**	**0.8498**	**0.6721**	**0.6800**	**0.8019**
ComDCT-T-B [$MR=0.25,\lambda_r=0.4,\lambda_c=0.1$]	0.9336	0.4619	0.8524	0.6778	0.6640	0.7988
ComDCT-U [$MR=0.25,\lambda_r=0.4$]	0.9332	0.2587	0.7941	0.3506	0.4678	0.7005
ComDCT-U-W [$MR=0.25,\lambda_r=0.4,\lambda_c=0.2$]	0.9377	0.3537	0.8170	0.4765	0.5491	0.7434
ComDCT-U-B [$MR=0.25,\lambda_r=0.4,\lambda_c=0.2$]	0.9048	0.3741	0.7800	0.4404	0.5315	0.7182
ComDefend[31]	0.9951	0.0179	0.5963	0.2179	0.2774	0.6362
ComDefend-W [$\lambda_c=0.8$]	0.9744	0.1001	0.6668	0.6056	0.4575	0.7159
ComDefend-B [$\lambda_c=0.4$]	0.9731	0.0981	0.6985	0.5881	0.4616	0.7174
JPEG[6] [quality=10]	0.9449	0.1007	0.7054	0.1577	0.3213	0.6331
TVD[6] [weight=9]	0.9630	0.1425	0.8348	0.0542	0.3439	0.6534
MF[7] [window=5]	0.9502	0.1917	0.8155	0.5330	0.5134	0.7318
LGS[8]	0.7858	0.2245	0.5549	0.6099	0.4631	0.6245
LRR[9]	0.8519	0.1453	0.5855	0.2635	0.3314	0.5917
DeepInverse	0.9313	0.1624	0.7292	0.2380	0.3766	0.6539
IHT[33]	0.9996	0.0023	0.2177	0.0521	0.0907	0.5451

表 7　各个防御策略平均处理一张大小为 32×32 的 RGB 图像的时间对比

防御策略	耗时/ms	防御策略	耗时/ms	防御策略	耗时/ms
$\text{ComDCT}_{\text{LISA}}$ [$MR=0.05$]	2.68	$\text{ComDCT}_{\text{LISA}}$ [$MR=0.10$]	2.67	$\text{ComDCT}_{\text{LISA}}$ [$MR=0.25$]	2.69
$\text{ComDCT}_{\text{SVHN}}$ [$MR=0.05$]	4.23	$\text{ComDCT}_{\text{SVHN}}$ [$MR=0.10$]	4.24	$\text{ComDCT}_{\text{SVHN}}$ [$MR=0.25$]	4.25
$\text{DeepInverse}_{\text{LISA}}$ [$MR=0.05$]	2.45	$\text{DeepInverse}_{\text{LISA}}$ [$MR=0.10$]	2.46	$\text{DeepInverse}_{\text{LISA}}$ [$MR=0.25$]	2.57
$\text{DeepInverse}_{\text{SVHN}}$ [$MR=0.05$]	4.07	$\text{DeepInverse}_{\text{SVHN}}$ [$MR=0.10$]	4.05	$\text{DeepInverse}_{\text{SVHN}}$ [$MR=0.25$]	4.10
JPEG[6] [quality=10]	0.32	TVD[6] [weight=3]	0.66	TVD[6] [weight=9]	0.54
MF[7] [window=5]	3.81	LGS[8]	0.27	LRR[9]	90.91
ComDefend[31]	4.12	IHT[33]	76.36	BP[34]	62174.19

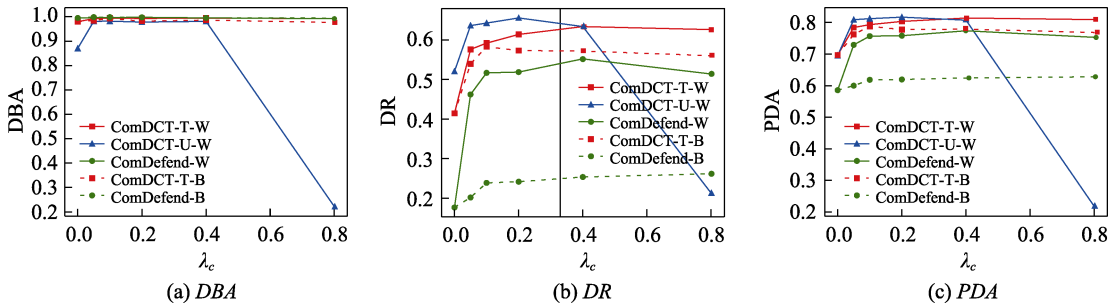

图 4 引入分类损失后不同λ_c下防御策略 ComDCT 和 ComDefend 在 LISA 数据集上的性能对比

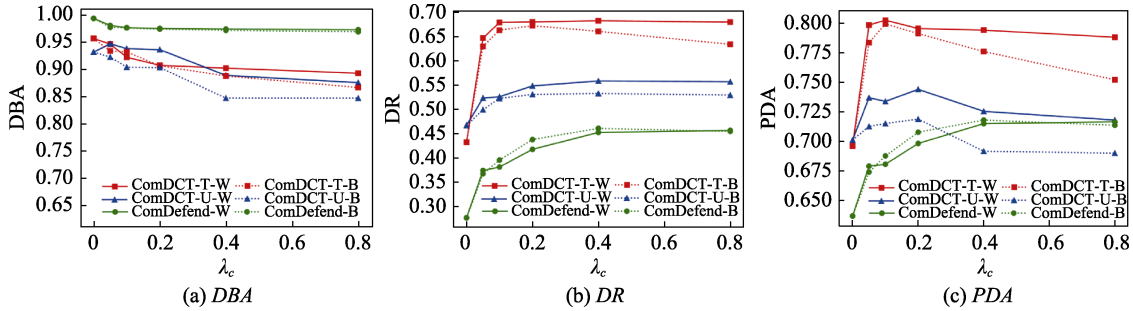

图 5 引入分类损失后不同λ_c下防御策略 ComDCT 和 ComDefend 在 SVHN 数据集上的性能对比

额外的运行处理时间，但该时间与 DeepInverse 处理时间相比可忽略不计，即 ComDCT 能够在获得与 DeepInverse 相当的实时性水平的情况下大幅度提高防御策略的综合性能，再一次说明了本文提出的基于稀疏离散余弦系数的 ComDCT 防御策略相比于直接使用 DeepInverse 的优越性.

另外，虽然 JPEG 等防御策略的处理时间少于 ComDCT，不仅满足本文提出的实时性要求（不超过 10ms）还可以应用于一些更高实时性要求的场景，但由表 5 及表 6 所示，这些运行更快的防御策略在 LISA 及 SVHN 数据集上防御综合性能指标*PDA*最高分别为 62.92%、65.34%，相较于本文所提出的 ComDCT 防御策略（综合性能指标 *PDA* 在未引入分类损失的情况下为 69.70%、70.05%，而且在引入分类损失的情况下高达 81.69%、80.19%）并不理想. 因此在同样满足实时性要求的情况下，使用综合性能更高的 ComDCT 是一个更优的选择.

4.4 引入分类损失

由于不同数据集图像的离散余弦系数的稀疏性往往不一致，因此针对不同数据集，使得 ComDCT 综合性能最优的正则项损失 λ_r 和测量率 *MR* 的取值也往往不一致. 本文把在未引入分类损失的情况下使得 ComDCT 综合性能最优的 λ_r 和 *MR* 的取值作为数据集的先验知识，并在 ComDCT 引入分类损失的实验中保留此 λ_r 和 *MR* 取值来进一步选取 λ_c.

引入分类损失后，本文提出的 ComDCT 防御策略和基于神经网络的 ComDefend 防御策略在不同 λ_c 下的性能对比如图 4 和图 5 所示. 实验结果表明：不管分类损失的引入是在白盒防御模式下还是在黑盒防御模式下，随着 λ_c 越来越大，使用训练得到的矩阵的 ComDCT-T、使用通用测量矩阵的 ComDCT-U 和 ComDefend 对良性图像的分类性能、对对抗性图像的防御性能以及其综合性能大致上都呈现出先增强再减弱的趋势. 这是因为越大的 λ_c 可以使防御策略预处理及去噪后的干净图像携带有越多的分类特征，而这些分类特征一方面可以进一步减小防御策略对良性样本分类性能的影响，另一方面可以进一步减轻对抗性扰动对分类模型的影响以提高防御性能，所以越大的 λ_c 将学习到越多的分类特征，最终可以成功地导致防御策略对良性图像的分类性能、对对抗性图像的防御性能以及其综合性能的共同提高. 然而，过大的 λ_c 会导致防御策略预处理及去噪后的干净图像跟原始图像的差异过大，表现出过拟合现象，最终导致防御策略对良性图像的分类性能、防御性能以及其综合性能的共同减弱，尤其是在黑盒防御模式下此问题更加突出.

在选取最优参数的情况下，ComDCT-T 防御策略对图像进行预处理及去噪的效果图如图 6 和图 7 所示. 图 6 和图 7 表明相比于不引入分类损失，ComDCT-T 引入分类损失以后，其预处理及去噪得到的干净图像与原始图像的差异更大，这进一步暗示了过大的 λ_c 会减弱防御策略的综合性能. 当 λ_c 的取值恰当

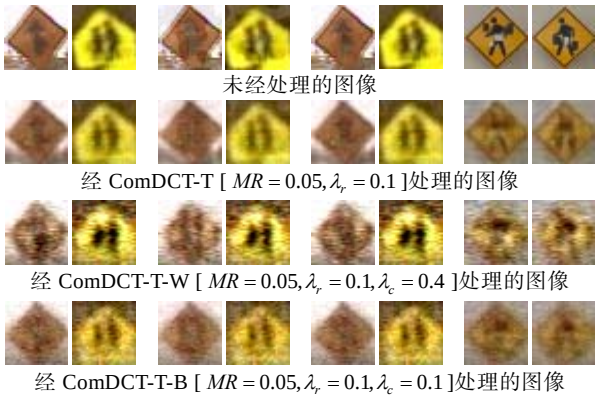

图 6　ComDCT-T 防御策略对 LISA 数据集图像进行处理的效果图（第 1-2、3-4、5-6、7-8 列依次针对良性图像、FGSM 对抗图像、CW-L2 对抗图像、Sticker 对抗图像）

图 7　ComDCT-T 防御策略对 SVHN 数据集图像进行处理的效果图（第 1-2、3-4、5-6、7-8 列依次针对良性图像、FGSM 对抗图像、CW-L2 对抗图像、LaVAN 对抗图像）

时，不管是在白盒防御模式下还是在黑盒防御模式下，分类损失的引入都可以提高防御策略 ComDCT 和 ComDefend 的综合性能，这充分表明了本文提出的引入分类损失的优越性和鲁棒性.

相比于使用通用测量矩阵的 ComDCT-U，由于使用训练得到的更加鲁棒的测量矩阵的 ComDCT-T 的重建质量更好，所以在重建图像跟原始图像的差异度相似的情况下，ComDCT-T 可以使重建图像携带有更多的分类特征，从而 ComDCT-T 更难以出现过拟合现象和拥有更加鲁棒的综合性能. 在黑盒防御模式下，ComDCT-T 的优越性更加明显，而且 ComDCT-U 对 λ_c 的取值更加敏感. 不同 λ_c 下 ComDCT-U-B 防御策略在 LISA 数据集上的性能对比如图 8 所示. 当λ_c=0.01 时，ComDCT-U-B 就已经出现了过拟合现象. 在选取最优λ_c 的情况下，相比于 ComDCT-U-B，ComDCT-T-B 的综合性能*PDA*在 LISA 数据集上高 5.91%，在 SVHN 数据集上高 8.06%.

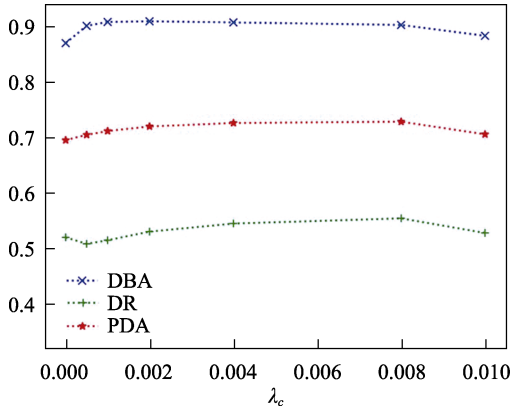

图 8　不同λ_c下 ComDCT-U-B 防御策略在 LISA 数据集上的性能对比

与 ComDefend 相比，不管是在白盒防御模式下还是在黑盒防御模式下，当选取最优λ_c 时，即使是使用通用测量矩阵的 ComDCT-U 的综合性能 PDA 也更高；而使用训练得到的测量矩阵的 ComDCT-T 的综合性能 *PDA* 与 ComDefend 的相比，在白盒防御模式下 LISA 数据集高 4.04%、SVHN 数据集高 15.97%，在黑盒防御模式下 LISA 数据集高 8.6%、SVHN 数据集高 8.14%. 这表明了本文提出的 ComDCT 防御策略的优越性.

在选取各自的最优参数的情况下，各个防御策略的性能对比如表 5 和表 6 所示. 在白盒防御模式下，引入分类损失的 ComDCT-T-W 与其他不引入分类损失的防御策略相比，其综合性能 *PDA* 在 LISA 数据集上至少高 11.88%，在 SVHN 数据集上至少高 7.01%；在黑盒防御模式下，引入分类损失的 ComDCT-T-B 与其他不引入分类损失的防御策略相比，其综合性能 *PDA* 在 LISA 数据集上至少高 9.25%，在 SVHN 数据集上至少高 6.7%. 这再一次说明了本文提出的 ComDCT 防御策略的优越性以及引入分类损失的优越性.

5　总结及展望

为了提高深度神经网络分类模型对于对抗样本的鲁棒性以将其安全地应用于无人驾驶等领域，本文提出了基于压缩感知的神经网络实时综合防御策略 ComDCT. 实验结果表明 ComDCT 与 ComDefend、MF 等现有防御策略相比，可有效提升分类模型对像素攻击和补丁攻击的防御性能，拥有更强综合鲁棒性. 本文还提出在是否了解受保护分类模型的白

盒、黑盒两种防御模式下均可通过引入分类损失大幅度提高 ComDCT 的综合性能，增强模型可用性.

然而,与 IHT[33]及 BP[34]等基于压缩感知技术的防御方案类似，受限于压缩感知大图像处理中矩阵存储及计算方面的难度，ComDCT 在大图像分类时需先将大图像分割成多个小图，然后分别对小图像进行压缩域与其稀疏离散余弦系数映射去噪处理及图像还原，最后将处理过的小图像拼接回大图像进行分类处理. 作为未来的工作，本文希望在大图像分类任务上探索更多更高效的防御算法.

参 考 文 献

[1] Carlini N, Wagner D. Towards evaluating the robustness of neural networks//Proceedings of the IEEE Symposium on Security and Privacy. San Jose, USA, 2017: 39-57

[2] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA, 2015: 1-11

[3] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition// Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Vienna, Austria, 2016: 1528-1540

[4] Karmon D, Zoran D, Goldberg Y. Lavan: localized and visible adversarial noise//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 2507-2515

[5] Eykholt K, Evtimov I, Fernandes E, et al. Robust physical-world attacks on deep learning visual classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1625-1634

[6] Guo C, Rana M, Cisse M, et al. Countering adversarial images using input transformations//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-12

[7] Xu W, Evans D, Qi Y. Feature squeezing: detecting adversarial examples in deep neural networks//Proceedings of the 25th Annual Network and Distributed System Security Symposium. San Diego, USA, 2018: 1-15

[8] Naseer M, Khan S, Porikli F. Local gradients smoothing: defense against localized adversarial attacks//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Waikoloa Village, USA, 2019: 1300-1307

[9] Zhao X, Stamm M C. Defenses against multi-sticker physical domain attacks on classifiers//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 202-219

[10] Ren K, Zheng T, Qin Z, et al. Adversarial attacks and defenses in deep learning. Engineering, 2020, 6(3): 346-360

[11] Yang Y, Luo H, Xu H, et al. Towards real-time traffic sign detection and classification. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(7): 2022-2031

[12] Shao F, Wang X, Meng F, et al. Real-time traffic sign detection and recognition method based on simplified Gabor wavelets and CNNs. Sensors, 2018, 18(10): 3192

[13] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada, 2014: 1-10

[14] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world//Proceedings of the 5th International Conference on Learning Representations. Toulon, France, 2017: 1-14

[15] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 9185-9193

[16] Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2574-2582

[17] Zheng T, Chen C, Ren K. Distributionally adversarial attack// Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 2253-2260

[18] Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition// Proceedings of the British Machine Vision Conference. Swansea, UK, 2015: 41.1-41.12

[19] Athalye A, Engstrom L, Ilyas A, et al. Synthesizing robust adversarial examples//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 284-293

[20] Kong Z, Guo J, Li A, et al. Physgan: generating physical-world-resilient adversarial examples for autonomous driving// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 14254- 14263

[21] Duan R, Mao X, Qin A K, et al. Adversarial laser beam: effective physical-world attack to DNNs in a blink//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 16062-16071

[22] Metzen J H, Genewein T, Fischer V, et al. On detecting adversarial perturbations//Proceedings of the 5th International Conference on Learning Representations. Toulon, France, 2017: 1-12

[23] Meng D, Chen H. Magnet: a two-pronged defense against adversarial examples//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Dallas, USA, 2017: 135-147

[24] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-23

[25] Tramèr F, Boneh D, Kurakin A, et al. Ensemble adversarial training: Attacks and defenses//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-20

[26] Xie C, Wang J, Zhang Z, et al. Mitigating adversarial effects through randomization//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-16

[27] Liu X, Cheng M, Zhang H, et al. Towards robust neural networks via random self-ensemble//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 369-385

[28] Dhillon G S, Azizzadenesheli K, Lipton Z C, et al. Stochastic activation pruning for robust adversarial defense//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-13

[29] Liao F, Liang M, Dong Y, et al. Defense against adversarial

attacks using high-level representation guided denoiser// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1778-1787

[30] Xie C, Wu Y, Maaten L, et al. Feature denoising for improving adversarial robustness//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 501-509

[31] Jia X, Wei X, Cao X, et al. Comdefend: an efficient image compression model to defend adversarial examples//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 6084-6092

[32] Mousavi A, Baraniuk R G. Learning to invert: signal recovery via deep convolutional networks//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, USA, 2017: 2272-2276

[33] Bafna M, Murtagh J, Vyas N. Thwarting adversarial examples: an l_0-robust sparse fourier transform//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada, 2018: 10096-10106

[34] Dhaliwal J, Hambrook K. Compressive recovery defense: defending neural networks against ℓ_2, ℓ_∞, and ℓ_0 norm attacks// Proceedings of the International Joint Conference on Neural Networks. Glasgow, UK, 2020: 1-8

[35] Wu S, Dimakis A, Sanghavi S, et al. Learning a compressed sensing measurement matrix via gradient unrolling//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 6828-6839

[36] Shi W, Jiang F, Liu S, et al. Scalable convolutional neural network for image compressed sensing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 12290-12299

[37] Mogelmose A, Trivedi M M, Moeslund T B. Vision-based traffic sign detection and analysis for intelligent driver assistance systems: perspectives and survey. IEEE Transactions on Intelligent Transportation Systems, 2012, 13 (4): 1484-1497

[38] Netzer Y, Wang T, Coates A, et al. Reading digits in natural images with unsupervised feature learning//Proceedings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning. Granada, Spain, 2011: 1-9

[39] Kabkab M, Samangouei P, Chellappa R. Task-aware compressed sensing with generative adversarial networks//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 2297-2304

[40] Kingma D P, Ba J. Adam: a method for stochastic optimization// Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA, 2015: 1-15

WANG Jia, Ph.D., assistant professor. Her current research interests include lightweight security design for IoT and data security and privacy-preserving aspects of machine learning.

ZHANG Yang-Mei, Ph.D candidate. Her current major research interest is artificial intelligence.

SU Wu-Qiang, master candidate. His current research interest is robustness of machine learning.

LUO Cheng-Wen, Ph.D., associate professor. His current major research interest is mobile and pervasive computing and security aspects of Internet of Things.

WU Chao, Ph.D., post-doctor. His current major research interest is artificial intelligence, embedded systems, file system, storage architecture.

LIN Qiu-Zhen, Ph.D., associate professor. His current major research interest is optimization algorithms.

LI Jian-Qiang, Ph.D., professor. His current major research interest is embedded systems, and Internet of Things

Background

Recent years have witnessed the rapid development of DNN-based algorithms and applications. However, research works have also shown the vulnerability of deep neural networks to intentionally crafted adversarial examples, which raises the security concerns of its deployment in security-critical scenarios like autonomous driving.

Researchers have made a lot of effort on improving the robustness of DNN models against adversarial attacks. According to the way of how the perturbations are organized, adversarial attacks could be broadly categorized into two groups, pixel attacks and patch attacks. Most well-known attacks such as FGSM, I-FGSM and CW-L2 seem to belong to the first category, which are often achieved by slightly manipulating pixel values to inject imperceptible changes into the images. And patch attacks intend to add visible noises that “naturally” covers prominent features of the images or design visible adversarial noises that confined to a localized patch of an image to fool the classifier.

Since DNN-based classifiers could be attacked by both kinds of attacks in its usage, defense strategies should be effective against both two kinds of attacks. However, existing solutions are typically designed for one class than the other. For example, Local Region Reconstruction algorithm could mitigate adversarial effects of patches but its defense performance against pixel attacks is not satisfying. Furthermore, existing defense algorithms usually lead to a drop in test accuracy, which also affect their practicability. Additionally, when it’s applied in scenarios like autonomous driving, the defense algorithm should be designed to meet the real-time requirement. Hence, designing practical real-

time comprehensive defensive strategies against a variety of adversarial attacks still remains a challenging task in machine learning.

In this paper, we propose a real-time comprehensive defense strategy ComDCT based on Compressive Sensing (CS) and neural network, which improves the existing CS recovery algorithm for the purpose of real-time comprehensive defense. Specifically, after training the neural network to learn the mapping from the measurements of the image to the sparse discrete cosine coefficient of the image, ComDCT can remove adversarial perturbations and obtain the restored clean image through the inverse discrete cosine transform from the restored discrete cosine coefficients. We also propose the introduction of classification loss to optimize the neural network for better defensive performance. Experiments are conducted on two public datasets under both the black-box and white-box settings. Empirical results showed its superiority compared with the state-of-the-art defensive methods including ComDefend, MF, TVD, LRR and so on.

The proposed ComDCT scheme provides real-time comprehensive defense performance against a variety of attacks, which could be used to improve the robustness of DNN-based visual classifier and promote its deployment in security-critical scenarios like autonomous driving.

This work was supported in part by the Joint Funds of the National Natural Science Foundation of China (Key Program Grant No. U1713212), the National Key R&D Program of China (Grant No. 2020YFA0908700), the National Natural Science Foundation of China (Grant Nos. 61806130, 6197071246, 62002338), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2021A1515011153), the Guangdong “Pearl River Talent Recruitment Program” (Grant No. 2019ZT08X603), the Shenzhen Science and Technology Innovation Commission-Stable Support Program (General Program Grant No. 20200805142159001) and the Shenzhen Science and Technology Innovation Commission (Grant No. R2020A045).