中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2024)07-1861-14

论文引用格式:Shi C, Liu Y, Zhao M H, Miao Q G and Pun C M. 2024. Contrastive semi-supervised adversarial training method for hyperspectral image classification networks. Journal of Image and Graphics, 29(07):1861-1874(石程, 刘莹, 赵明华, 苗启广, 潘治文. 2024. 面向高光谱图像分类网络的对比半监督对抗训练方法. 中国图象图形学报, 29(07):1861-1874)[DOI:10.11834/jig.230462]

面向高光谱图像分类网络的对比半监督对抗训练方法

石程[1],刘莹[1],赵明华[1*],苗启广[2],潘治文[3]

1. 西安理工大学计算机科学与工程学院,西安　710048; 2. 西安电子科技大学计算机科学与技术学院,西安　710071;
3. 澳门大学科技学院电脑及资讯科学系,澳门　999078

摘　要:**目的**　深度神经网络在高光谱图像分类任务中表现出明显的优越性,但是对抗样本的出现使其鲁棒性受到严重威胁,对抗训练方法为深度神经网络提供了一种有效的保护策略,但是在有限标记样本下提高目标网络的鲁棒性和泛化能力仍然需要进一步研究。为此,本文提出了一种面向高光谱图像分类网络的对比半监督对抗训练方法。**方法**　首先,根据少量标记样本预训练目标模型,并同时利用少量标记样本和大量无标记样本构建训练样本集合;然后,通过最大化训练样本集合中干净样本和对抗样本在目标模型上的特征差异生成高迁移性对抗样本;最后,为了减少对抗训练过程对样本标签的依赖以及提高目标模型对困难对抗样本的学习和泛化能力,充分利用目标模型和预训练模型的输出层及中间层特征,构建对比对抗损失函数对目标模型进行优化,提高目标模型的对抗鲁棒性。对抗样本生成和目标网络优化过程交替进行,并且不需要样本标签的参与。**结果**　在PaviaU和Indian Pines两组高光谱图像数据集上与主流的5种对抗训练方法进行了比较,本文方法在防御已知攻击和多种未知攻击上均表现出明显的优越性。面对6种未知攻击,相比于监督对抗训练方法AT(adversarial training)和TRADES(trade-off between robustness and accuracy),本文方法分类精度在两个数据集上平均提高了13.3%和16%,相比于半监督对抗训练方法SRT(semi-supervised robust training)、RST(robust self-training)和MART(misclassification aware adversarial risk training),本文方法分类精度再两个数据集上平均提高了5.6%和4.4%。实验结果表明了提出模型的有效性。**结论**　本文方法能够在少量标记样本下提高高光谱图像分类网络的防御性能。

关键词:对抗防御;高光谱图像分类;半监督学习;深度神经网络;对抗攻击

Contrastive semi-supervised adversarial training method for hyperspectral image classification networks

Shi Cheng[1], Liu Ying[1], Zhao Minghua[1*], Miao Qiguang[2], Pun Chi-Man[3]

1. School of Computer Science and Engineering, Xi'an University of Technology, Xi'an　710048, China;
2. School of Computer Science and Technology, Xidian University, Xi'an　710071, China;
3. Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau　999078, China

Abstract: **Objective** Deep neural networks have demonstrated significant superiority in hyperspectral image classification tasks. However, the emergence of adversarial examples poses a serious threat to their robustness. Research on adversarial training methods provides an effective defense strategy for protecting deep neural networks. However, existing adversarial

收稿日期:2023-07-16;**修回日期**:2023-10-31;**预印本日期**:2023-11-06

*通信作者:赵明华　zhaominghua@xaut.edu.cn

基金项目:国家自然科学基金项目(61902313, 62002272);陕西省重点研发计划项目(2024GH-ZDXM-47)

Supported by: National Natural Science Foundation of China (61902313, 62002272); Provincial Key Research and Development Program of Shaanxi (2024GH-ZDXM-47)

training methods often require a large number of labeled examples to enhance the robustness of deep neural networks, which increases the difficulty of labeling hyperspectral image examples. In addition, a critical limitation of current adversarial training approaches is that they usually do not capture intermediate layer features in the target network and pay less attention to challenging adversarial samples. This oversight can lead to the reduced generalization ability of the defense model. To further enhance the adversarial robustness of hyperspectral image classification networks with limited labeled examples, this paper proposes a contrastive semi-supervised adversarial training method. **Method** First, the target model is pre-trained using a small number of labeled examples. Second, for a large number of unlabeled examples, the corresponding adversarial examples are generated by maximizing the feature difference between clean unlabeled examples and adversarial examples on the target model. Adversarial samples generated using intermediate layer features of the network exhibit higher transferability compared with those generated only using output layer features. In contrast, feature-based adversarial sample generation methods do not rely on example labels. Therefore, we generate adversarial examples based on the intermediate layer features of the network. Third, the generated adversarial examples are used to enhance the robustness of the target model. The defense capabilities of the target model for the challenging adversarial samples are enhanced by defining the robust upper bound and robust lower bound of the target network based on the pre-trained target model, and a contrastive adversarial loss is designed on both intermediate feature layer and output layer to optimize the model based on the defined robust upper bound and robust lower bound. The defined contrastive loss function consists of three terms: classification loss, output contrastive loss, and feature contrastive loss. The classification loss is designed to maintain the classification accuracy of the target model for clean examples. The output contrastive loss encourages the output layer of the adversarial examples to move closer to the pre-defined output layer robust upper bound and away from the pre-defined output layer robust lower bound. The feature contrastive loss pushes the intermediate layer feature of the adversarial example closer to the pre-defined intermediate robust upper bound and away from the pre-defined intermediate robust lower bound. The proposed output contrastive adversarial loss and feature contrastive loss help improve the classification accuracy and generalization ability of the target network against challenging adversarial examples. The training process of adversarial example generation and target network optimization is performed iteratively, and example labels are not required in the training process. By incorporating a limited number of labeled examples in model training, both the output layer and intermediate feature layer are used to enhance the defense ability of the target model against known and unknown attack methods. **Result** We compared the proposed method with five mainstream adversarial training methods, two supervised adversarial training methods and three semi-supervised adversarial training methods, on the PaviaU and Indian Pines hyperspectral image datasets. Compared with the mainstream adversarial training methods, the proposed method demonstrates significant superiority in defending against both known and various unknown attacks. Faced with six unknown attacks, compared with the supervised adversarial training methods AT and TRADES, our method showed an average improvement in classification accuracy of 13.3% and 16%, respectively. Compared with the semi-supervised adversarial training methods SRT, RST, and MART, our method achieved an average improvement in classification accuracy of 5.6% and 4.4%, respectively. Compared with the target model without defense method, for example on the Inception_V3, the defense performance of the proposed method in the face of different attacks improved by 34.63%−92.78%. **Conclusion** The proposed contrastive semi-supervised adversarial training method can improve the defense performance of hyperspectral image classification networks with limited labeled examples. By maximizing the feature distance between clean examples and adversarial examples on the target model, we can generate highly transferable adversarial examples. To address the limitation of defense generalization ability imposed by the number of labeled examples, we define the concept of robust upper bound and robust lower bound based on the pre-trained target model and design an optimization model according to a contrastive semi-supervised loss function. By extensively leveraging the feature information provided by a few labeled examples and incorporating a large number of unlabeled examples, we can further enhance the generalization ability of the target model. The defense performance of the proposed method is superior to that of the supervised adversarial training methods.

Key words: adversarial defense; hyperspectral image classification; semi-supervised learning; deep neural network; adversarial attack

0 引 言

深度神经网络已广泛应用于图像分类(Zhang等，2022b)、目标检测(Kang等，2018)(袁珑 等，2022)以及自动驾驶(Eykholt等，2018)等领域。然而，众多学者发现深度神经网络极易受到对抗样本的攻击，对抗鲁棒性不足。在原始干净样本上添加人眼无法察觉的扰动，就使得深度神经网络以较高的置信度产生错误的预测(Szegedy等，2014)。目前大量针对对抗攻击和防御的研究集中于自然图像处理领域，但是遥感图像处理领域中也同样存在网络对抗鲁棒性不足的问题，如在高光谱图像分类任务中也发现了对抗样本的存在(Shi等，2022)。高光谱图像具有冗余的光谱波段和丰富的纹理信息，使得扰动更易被隐藏，可以生成扰动强度更大而不被人眼觉察的对抗样本，对高光谱图像分类网络的鲁棒性产生极大的威胁。高光谱图像在军事上具有重要地位，因此提高高光谱图像分类网络的对抗鲁棒性具有重要的应用价值。

为了抵御对抗样本对深度神经网络的威胁，现有对抗防御方法主要分为4类(Machado等，2021)：1)梯度掩蔽，其中典型的梯度掩蔽方法包括对抗训练(Madry等，2019)和防御蒸馏(Papernot等，2016)；2)对抗样本检测(Li和Li，2017)，通过检测分类器的输入图像是否为对抗样本，从而防止对抗样本对分类器的攻击；3)预处理技术，一种方式是通过对输入图像应用空间变换、JPEG(joint photographic experts group)压缩、填充等操作破坏扰动(Xie等，2018)，另一种方式是基于预处理网络去除扰动，如Jin等人(2017)利用生成对抗网络从对抗样本中重建干净样本，Zhou等人(2021)利用变分自编码器去除对抗样本中的扰动；4)改变目标模型的结构，如Tramèr等人(2020)构建集成分类器来防御对抗样本。目前为止还不存在一种防御方法能够有效抵御所有对抗攻击，但广泛认为对抗训练是目前最有效的防御方法。

对抗训练方法的研究主要分为监督对抗训练和半监督对抗训练。监督对抗训练利用对抗样本及其标签对目标网络重训练，以提高目标网络防御对抗样本的能力(Madry等，2019)。目前监督对抗训练取得了较好的防御效果，但是也存在一些挑战性的问题：1)防御性能依赖于对抗样本类型。监督对抗训练对对抗样本类型具有较高的依赖性，如果对抗样本具有较强的可迁移性，防御模型就能够有效抵御未知攻击，否则，防御模型的泛化能力急剧下降；2)防御泛化能力受限于标记样本数量。较少的标记样本会影响防御模型对未知攻击的泛化能力。问题1)和2)的联立，进一步加剧了监督对抗训练的过拟合。为此，一方面，学者研究如何生成高迁移性对抗样本，期望解决防御性能依赖于对抗样本类型的问题(Huang等，2019)；另一方面，学者进一步研究了半监督对抗训练方法，期望解决防御泛化能力受限于标记样本数量这一问题(Carmon等，2019；Miyato等，2019)。半监督对抗训练方法同时利用大量未标记样本和少量标记样本生成的对抗样本对目标模型重训练，在少量标记样本条件下提高目标模型抵御未知攻击的泛化能力。

高光谱图像存在成像场景大、标签获取困难的问题(谭琨 等，2019)，因此在少量标签样本下提高高光谱图像分类网络的对抗鲁棒性尤为重要。现有半监督对抗训练方法存在以下两个问题：一方面，目标模型的优化仅约束对抗样本和干净样本在目标网络输出层相近，忽略了中间特征层所提供的语义信息。事实上，对抗样本对目标网络的改变不仅反映在输出层，特征层也同样受到扰动的干扰。另一方面，现有半监督对抗损失函数的构建对容易和困难的对抗样本赋予相同的关注，导致对抗训练后的目标模型对困难对抗样本的防御能力不足。针对以上问题，本文提出了一种面向高光谱图像分类网络的对比半监督对抗训练方法(contrastive semi-supervised adversarial training，CSAT)，通过充分挖掘未标记样本的本质特征，实现强泛化的半监督对抗训练，解决防御性能依赖于对抗样本类型和数量的问题。

现有对抗攻击方法的研究发现，相比于根据目标模型输出层的预测结果生成对抗样本，增加对抗样本和干净样本在特征空间上的距离能够获得更高的可迁移性(Huang等，2019；Yang等，2022；Naseer等，2019；Wu等，2018；Ren等，2023)。基于此，本文基于Naseer等人(2019)的工作，通过最大化对抗样本和干净样本的特征距离生成高迁移性对抗样本；然后定义了目标网络的鲁棒上界和鲁棒下界，并依据定义在目标网络的输出层和中间特征层构建了对

比对抗损失函数，使得模型优化更关注于困难对抗样本的学习，提升对抗训练后目标模型的泛化能力。

本文的主要贡献如下：1）针对防御性能依赖于对抗样本类型的问题，引入特征层扰动生成模型，生成自适应、高迁移的对抗样本，提高对抗训练的泛化能力；2）针对防御泛化能力受限于标记样本数量的问题，提出对比对抗损失函数优化模型，增加模型优化过程中对困难对抗样本的关注，在少量标记样本情况下提高目标网络防御未知攻击的能力；3）在两组高光谱图像数据集上的实验结果表明，本文方法在少量标记样本情况下对已知和未知攻击的防御性能甚至优于监督对抗训练方法。

1 相关工作

1.1 对抗攻击

Szegedy等人（2014）首先发现了对抗样本的存在，并将对抗样本的求解问题转化为盒约束的优化问题，即

$$\arg\min_{\boldsymbol{r}} \|\boldsymbol{r}\|_2 \quad \text{s.t. } f_\theta(\boldsymbol{x}+\boldsymbol{r})\neq \boldsymbol{y} \tag{1}$$

式中，$\boldsymbol{x}$表示原始干净样本，$\boldsymbol{y}$表示对应的真值标签，f_θ表示目标模型，如VGG（Visual Geometry Group）、ResNet（residual network）等深度神经网络，$\boldsymbol{r}$表示对抗扰动。该优化问题表示攻击者想要找到能够让分类器错误分类的最小扰动。

由于式（1）无法求解，众多学者研究了多种方法求解使分类器错分的最小扰动。Goodfellow等人（2015）提出了快速梯度符号法（fast gradient sign method，FGSM），根据梯度符号生成对抗样本。由于FGSM是一种单步攻击方法，具有较快的攻击速度，但是攻击成功率有待提高。为了进一步提高攻击成功率，多步迭代攻击方法成为研究的主要方向，代表性方法有：Madry等人（2018）提出的投影梯度下降法（projected gradient descent，PGD），Kurakin等人（2017）提出的迭代快速梯度符号法（iterative-FGSM，I-FGSM），Moosavi-Dezfooli等人（2016）提出的DeepFool攻击法，Carlini和Wagner（2017）提出的C&W（Carlini and Wagner）攻击法。以上攻击方法在白盒设置下具有较高的攻击成功率，但是在黑盒设置下攻击成功率大幅下降。

为了实现黑盒攻击，在提高攻击成功率的同时学者们也关注提高对抗样本的可迁移性，代表性方法有：Dong等人（2018）提出的动量迭代的快速梯度符号法（momentum iterative-FGSM，MI-FGSM），在迭代过程中引入动量项以稳定梯度的更新方向，Lin等人（2020）利用Nesterov加速梯度（nesterov accelerated gradient，NAG）提高训练效率，Xie等人（2018）通过对输入进行变换提高对抗样本的可迁移性（diverse input iterative-FGSM，DI-FGSM），Dong等人（2019）通过平滑梯度减少对抗样本对不同目标模型的敏感度，Wang等人（2021a）通过同时考虑前向和后向梯度方差调整当前梯度，稳定扰动更新方向（variance based momentum iterative-FGSM，VMI-FGSM），Zhu等人（2023）通过考虑时间维度上的邻域梯度，提出自适应点选择迭代梯度符号法（adaptive points selecting iterative-FGSM，AI- FGSM），Croce和Hein（2020）提出了无参数的PGD方法（auto-PGD）和集成攻击方法（auto attack）。

以上对抗攻击的实现依赖于待攻击目标网络的输出层分布。现有研究表明，基于目标网络中间层特征的对抗攻击方法能够获得可迁移性更高的对抗样本。Huang等人（2019）提出在中间层的特征图上增加扰动的强度，通过最大化中间特征图中扰动的范数提高对抗样本的可迁移性；Yang等人（2022）通过破坏网络中间层特征中的高注意力特征提高对抗样本的可迁移性；Wang等人（2021b）和Zhang等人（2022a）分别通过聚合梯度和神经元属性估计中间层特征图的重要性，并对重要特征进行攻击；Ren等人（2023）提出一个联合的特征增强变换模块估计特征的重要性，并同时利用获取的正—负显著性图实现特征层攻击；Naseer等人（2019）通过增加对抗样本和干净样本在中间特征层的距离获得高迁移性对抗样本，不同于以上工作，该工作在生成对抗样本过程中不需要样本标签的参与。由于高光谱图像标签样本缺乏，本文采用Naseer等人（2019）的工作生成高迁移性对抗样本。

1.2 对抗训练

目前，对抗训练是增强目标网络对抗鲁棒性的重要方式，也是最有效的方式之一（Goodfellow等，2015）。对抗训练过程可以表示为最大—最小优化问题（Madry等，2019），即

$$\min_{\theta}\max_{\|\boldsymbol{r}\|\leqslant\varepsilon} L(f_\theta(\boldsymbol{x}+\boldsymbol{r}),\boldsymbol{y}) \tag{2}$$

式中，L表示损失函数。式(2)中，内层最大化表示通过最大化损失函数产生对抗扰动，在固定扰动的情况下，通过外层最小化优化模型参数。

基于式(2)，众多改进的对抗训练方法相继提出，目的是同时提高目标模型的对抗鲁棒性和干净样本的分类精度。Zhang等人(2019)提出了替代损失函数TRADES(trade-off between robustness and accuracy)，以平衡目标模型上对抗样本和干净样本的分类性能；Lamb等人(2022)提出了插值对抗训练方法IAT(interpolated adversarial training)，通过在对抗训练过程中对对抗样本进行插值，提高目标网络对未知攻击的泛化能力。

为了克服标记样本数量对防御模型泛化能力的制约(Hendrycks等，2019)，半监督对抗训练方法被提出，利用大量的未标记样本提高目标模型的对抗鲁棒性。同时Carmon等人(2019)提出了鲁棒性对抗训练方法RST(robust self-training)，利用目标模型对无标记样本预测伪标签，并同时利用标记样本和无标记样本对目标网络进行重训练，该工作也证明了无标记样本对提高目标网络对抗鲁棒性的重要性；北京大学王奕森团队(Wang等，2020)重新定义了对抗样本的概念，指出对抗样本不仅包括原本被正确分类、但在添加扰动后被错分的样本，还包括原本就被目标模型所错分的样本，并基于该定义，提出MART方法对原本分类错误的样本添加额外的正则化约束，以进一步提高目标模型的对抗鲁棒性提出MART(misclassification aware adversarial risk training)；Li等人(2022)提出了一种半监督对抗训练方法SRT(semi-supervised robust training)，通过约束邻域样本和干净样本具有相同预测来提高分类网络的鲁棒性。

不同于以上工作，本文提出了一种面向高光谱图像分类网络的对比半监督对抗训练方法，通过充分挖掘干净样本和对抗样本在目标网络输出层和中间特征层的特征，提高对抗样本的可迁移性并减少高光谱图像分类网络鲁棒性提升对样本标签的依赖；通过构建对比对抗损失函数，使得对抗训练过程更关注目标网络对困难对抗样本的防御能力，提升分类网络防御的泛化性。在本文提出的对比半监督对抗训练方法中，标记样本仅用于预训练目标模型，用来对未标记样本生成高质量的伪标签和构造更加准确的对比对抗损失函数，而后续对抗样本生成过程和对抗训练损失函数的构建均不需要样本标签的参与。Uesato等人(2019)通过实验验证了无标记样本可以提高防御模型的鲁棒性，并且采用的无标记样本数量越多，防御模型的泛化能力越强。因此在本文方法中，可以充分利用大量无标记样本提高高光谱图像分类网络的防御性能。本文在两个高光谱图像数据集上进行实验，实验结果证明，面对高维图像分类网络，提出方法训练的目标模型能够抵御多种对抗攻击方法，防御性能甚至优于监督对抗训练方法。

2 对比半监督对抗训练

本文提出了一种面向高光谱图像分类网络的对比半监督对抗训练方法(contrastive semi-supervised adversarial training，CSAT)，以突破目标网络对抗鲁棒性的提高依赖于对抗样本类型及标记样本数量的限制，CSAT方法的整体框架如图1所示。本节将分别对提出方法进行详细介绍。在2.1节中，根据少量标记样本预训练目标模型，并利用标记样本和无标记样本构建了训练样本集合，作为对抗训练的输入；2.2节介绍了基于特征空间的对抗样本生成方法；在2.3节中，依据预训练模型定义了鲁棒上界和鲁棒下界，并基于此定义和对抗样本在目标模型输出层和特征层的分布构建了对比对抗损失函数，以对目标网络重训练；2.4节介绍了提出方法的优化过程，其中，预训练模型过程需要少量标记样本的参与。2.2节和2.3节所述的对抗样本生成过程和目标网络重训练过程均不依赖样本标签。

2.1 准备工作

假设$\tilde{\boldsymbol{x}}=\{\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_{n1}\}\in\boldsymbol{D}_L$表示标记样本，$\hat{\boldsymbol{x}}=\{\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_{n2}\}\in\boldsymbol{D}_U$表示无标记样本，$\tilde{\boldsymbol{y}}=\{\tilde{y}_1,\tilde{y}_2,\cdots,\tilde{y}_{n1}\}$表示标记样本的真实标签。首先利用标记样本及真实标签对目标模型f_θ进行预训练。对目标模型预训练的目的是为了对大量无标记样本生成高质量的伪标签。根据预训练模型f_θ对无标记样本的标签进行预测，得到伪标签$\hat{\boldsymbol{y}}=\{\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_{n1}\}$，$\hat{y}_i=\arg\max f_\theta(\hat{y}_i)$。

同时利用标记样本和无标记样本对目标模型进行对抗训练，定义全部训练样本集合为$\{\boldsymbol{x},\boldsymbol{y}\}=\{\{\tilde{\boldsymbol{x}},\hat{\boldsymbol{x}}\}|\{\tilde{\boldsymbol{y}},\hat{\boldsymbol{y}}\}\}\in\{\boldsymbol{D}_L\cup\boldsymbol{D}_U\}$。

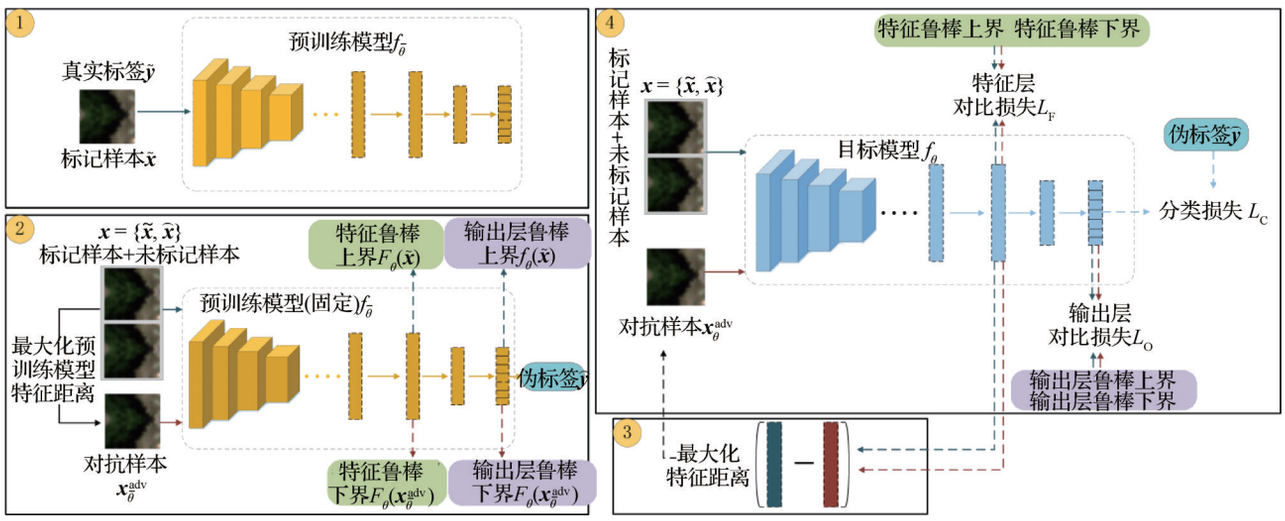

图1　对比半监督对抗训练方法整体框架图

Fig. 1　Overall framework of contrastive semi-supervised adversarial training method

2.2　基于特征空间的对抗样本生成

以往的攻击者大多利用输出层预测结果和样本标签生成对抗样本，攻击方法对样本标签和目标模型具有较强的依赖性，导致对抗样本可迁移性降低。Naseer等人(2019)指出破坏目标模型的中间层特征可以生成高迁移性对抗样本。基于此，本文通过最大化对抗样本和干净样本在目标模型中间层的特征生成对抗样本，优化函数为

$$L_f(\boldsymbol{x},\boldsymbol{x}_{\theta}^{\mathrm{adv}})=\left\|\boldsymbol{F}_{\theta}(\boldsymbol{x})-\boldsymbol{F}_{\theta}(\boldsymbol{x}_{\theta}^{\mathrm{adv}})\right\|_2^2$$
$$\text{s.t.}\ \left\|\boldsymbol{x}-\boldsymbol{x}_{\theta}^{\mathrm{adv}}\right\|_2\leqslant r \tag{3}$$

式中，$\boldsymbol{F}_{\theta}(\boldsymbol{x})$表示干净样本$\boldsymbol{x}$在目标模型$f_{\theta}$上的中间层特征，$\boldsymbol{x}_{\theta}^{\mathrm{adv}}$表示在目标模型$f_{\theta}$上生成的对抗样本，对抗样本优化过程为

$$\boldsymbol{x}_{\theta}^{\mathrm{adv}}(t+1)=\boldsymbol{x}_{\theta}^{\mathrm{adv}}(t)+\varepsilon\cdot\mathrm{sign}(\nabla_x L_f(\boldsymbol{x},\boldsymbol{x}_{\theta}^{\mathrm{adv}})) \tag{4}$$

式中，$\boldsymbol{x}_{\theta}^{\mathrm{adv}}(0)=\boldsymbol{x}+\mu$，$\mu$为随机噪声，$\varepsilon$是扰动强度，sign(·)是符号函数，$\nabla$表示梯度。式(4)中，通过迭代方式不断优化对抗样本，使得生成的对抗样本与干净样本距离最大化。

2.3　对比对抗损失约束的模型优化

为了提高网络对对抗样本、尤其是困难对抗样本的防御能力，本文定义了对比对抗损失函数对目标模型重训练。对比对抗损失函数由3部分组成，包括分类损失、输出层对比损失以及中间特征层对比损失，其中分类损失保证目标模型对干净样本的分类精度；输出层对比损失和中间特征层对比损失促使目标模型可以正确分类对抗样本。

本节首先定义了鲁棒上界和鲁棒下界的概念，用于在输出层对比损失以及中间特征层对比损失中约束目标模型f_{θ}对对抗样本的学习。

定义1：鲁棒上界。预训练模型$f_{\tilde{\theta}}$是利用干净样本训练的模型，也就是对抗样本期望得到的预测结果。因此，训练样本集合中干净样本$\boldsymbol{x}$在预训练模型$f_{\tilde{\theta}}$上的预测输出$f_{\tilde{\theta}}(\boldsymbol{x})$以及中间层特征$\boldsymbol{F}_{\tilde{\theta}}(\boldsymbol{x})$被定义为对抗样本期望达到的鲁棒上界。

定义2：鲁棒下界。没有防御保护的预训练模型$f_{\tilde{\theta}}$面对攻击时鲁棒性急剧降低，所以在预训练模型$f_{\tilde{\theta}}$上生成的对抗样本有极强的威胁性。因此，在预训练模型$f_{\tilde{\theta}}$上利用2.2节所述的特征层对抗样本生成方法生成干净样本$\boldsymbol{x}$对应的对抗样本$\boldsymbol{x}_{\tilde{\theta}}^{\mathrm{adv}}$，对抗样本$\boldsymbol{x}_{\tilde{\theta}}^{\mathrm{adv}}$在预训练模型$f_{\tilde{\theta}}$上的预测输出$f_{\tilde{\theta}}(\boldsymbol{x}_{\tilde{\theta}}^{\mathrm{adv}})$以及中间层特征$\boldsymbol{F}_{\tilde{\theta}}(\boldsymbol{x}_{\tilde{\theta}}^{\mathrm{adv}})$被定义为对抗样本的鲁棒下界。

下面详细介绍分类损失、输出层对比损失以及中间特征层对比损失。

1)分类损失。对抗训练中需要保证干净样本被正确分类，因此分类损失为

$$L_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})=L_{\mathrm{CE}}(f_{\theta}(\boldsymbol{x}),\boldsymbol{y}) \tag{5}$$

式中，L_{CE}表示交叉熵损失。

2)输出层对比损失。现有半监督对抗训练方法大多通过最小化对抗样本和干净样本在输出层的分布构建损失函数，以提高目标网络对对抗样本的分类精度。但是对抗样本也存在难易之分，与干净样本分布相近的对抗样本更容易被学习。因此对抗损失的构建方式使得网络在优化过程中对容易和困难

学习的对抗样本赋予相同的关注度，使得更新后的目标模型难以正确分类困难对抗样本。为了更关注困难对抗样本的学习。输出层对比损失为

$$L_{\mathrm{O}}(\boldsymbol{x}, f_{\tilde{\theta}}) = -\log \frac{\exp \dfrac{f_\theta(\boldsymbol{x}_\theta^{\mathrm{adv}}) \cdot f_{\tilde{\theta}}(\boldsymbol{x})}{\tau}}{\exp \dfrac{f_\theta(\boldsymbol{x}_\theta^{\mathrm{adv}}) \cdot f_{\tilde{\theta}}(\boldsymbol{x})}{\tau} + \exp \dfrac{f_\theta(\boldsymbol{x}_\theta^{\mathrm{adv}}) \cdot f_{\tilde{\theta}}(\boldsymbol{x}_{\tilde{\theta}}^{\mathrm{adv}})}{\tau}} \tag{6}$$

式中，$f_{\tilde{\theta}}$表示预训练模型，f_θ表示目标模型，$f_\theta(\boldsymbol{x}_\theta^{\mathrm{adv}})$表示对抗样本$\boldsymbol{x}_\theta^{\mathrm{adv}}$在目标模型$f_\theta$的输出层预测分布；$f_{\tilde{\theta}}(\boldsymbol{x})$表示鲁棒上界，即干净样本$\boldsymbol{x}$在预训练模型$f_{\tilde{\theta}}$的输出层分布；$f_{\tilde{\theta}}(\boldsymbol{x}_\theta^{\mathrm{adv}})$表示鲁棒下界，即对抗样本$\boldsymbol{x}_{\tilde{\theta}}^{\mathrm{adv}}$在预训练模型$f_{\tilde{\theta}}$的输出层预测分布，其中对抗样本$\boldsymbol{x}_\theta^{\mathrm{adv}}$是在预训练模型$f_{\tilde{\theta}}$上根据式(3)生成的对抗样本，$\tau$表示温度，调节对困难对抗样本的关注度。

式(6)促使目标模型上生成的对抗样本的输出更靠近预训练模型上干净样本的分布，远离预训练模型上对抗样本的分布，也就是更靠近鲁棒上界，远离鲁棒下界，从而在输出层提高目标模型防御攻击的能力。

3)特征层对比损失。对抗样本不仅造成网络预测错误，同样会破坏网络中间层特征，因此本文进一步定义了特征层对比损失，对网络的中间层特征进行约束，提高对抗训练的泛化能力。特征层对比损失为

$$L_{\mathrm{F}}(\boldsymbol{x}, \boldsymbol{F}_{\tilde{\theta}}) = -\log \frac{\exp \dfrac{\boldsymbol{F}_\theta(\boldsymbol{x}_\theta^{\mathrm{adv}}) \cdot \boldsymbol{F}_{\tilde{\theta}}(\boldsymbol{x})}{\tau}}{\exp \dfrac{\boldsymbol{F}_\theta(\boldsymbol{x}_\theta^{\mathrm{adv}}) \cdot \boldsymbol{F}_{\tilde{\theta}}(\boldsymbol{x})}{\tau} + \exp \dfrac{\boldsymbol{F}_\theta(\boldsymbol{x}_\theta^{\mathrm{adv}}) \cdot \boldsymbol{F}_{\tilde{\theta}}(\boldsymbol{x}_{\tilde{\theta}}^{\mathrm{adv}})}{\tau}} \tag{7}$$

式中，$\boldsymbol{F}_\theta(\boldsymbol{x}_\theta^{\mathrm{adv}})$表示对抗样本$\boldsymbol{x}_\theta^{\mathrm{adv}}$在目标模型$f_\theta$的中间层特征分布；$\boldsymbol{F}_{\tilde{\theta}}(\boldsymbol{x})$表示特征鲁棒上界，即干净样本$\boldsymbol{x}$在预训练模型$f_{\tilde{\theta}}$的中间层特征分布；$\boldsymbol{F}_{\tilde{\theta}}(\boldsymbol{x}_{\tilde{\theta}}^{\mathrm{adv}})$表示特征鲁棒下界，即对抗样本$\boldsymbol{x}_{\tilde{\theta}}^{\mathrm{adv}}$在预训练模型$f_{\tilde{\theta}}$的中间层特征分布。

式(7)促使目标模型上生成的对抗样本的中间层特征更靠近特征鲁棒上界，远离特征鲁棒下界，从而在特征层提高目标模型防御攻击的能力。

综上所述，总体的对比对抗损失函数定义为

$$L = L_{\mathrm{C}} + \alpha L_{\mathrm{O}} + \beta L_{\mathrm{F}} \tag{8}$$

式中，α和β为调节参数，在3.5节进行了详细分析。

2.4 对比半监督对抗训练

根据2.2节所介绍的对抗样本生成方法和2.3节所介绍的对比对抗损失，建立对抗训练模型。

1)内层最大化，即

$$\boldsymbol{x}_\theta^{\mathrm{adv}} \leftarrow \arg\max_{\boldsymbol{x}_\theta^{\mathrm{adv}}} L_f(\boldsymbol{x}, \boldsymbol{x}_\theta^{\mathrm{adv}}) \tag{9}$$

2)外层最小化，即

$$f_\theta \leftarrow \arg\min_\theta \{L_{\mathrm{C}} + \alpha L_{\mathrm{O}} + \beta L_{\mathrm{F}}\} \tag{10}$$

内层最大化和外层最小化提交训练以提高目标模型的对抗鲁棒性。总体算法流程如算法1所示。其中仅在步骤1)中用到样本标签，对抗训练过程不需要样本标签的参与。

算法1 对比半监督对抗训练方法

输入：标记样本$\tilde{\boldsymbol{x}}$及其标签$\tilde{\boldsymbol{y}}$，无标记样本$\hat{\boldsymbol{x}}$；

输出：鲁棒的目标模型f_θ

1)利用标记样本集合$\{\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}\}$预训练模型$f_{\tilde{\theta}}$。

2)根据预训练模型$f_{\tilde{\theta}}$对无标记样本预测其伪标签$\hat{\boldsymbol{y}}$。

3)同时利用标记样本和无标记样本构建训练样本集合$\{\boldsymbol{x}, \boldsymbol{y}\}$。

4)利用预训练模型$f_{\tilde{\theta}}$的参数初始化目标模型f_θ。

5)Repeat：

(1)提取目标模型f_θ的中间层特征，并构建如式(3)所示的对抗样本生成损失；

(2)最大化式(3)，并按照式(4)所示的优化公式求解对抗样本$\boldsymbol{x}_\theta^{\mathrm{adv}}$；

(3)根据干净样本$\boldsymbol{x}$在目标模型f_θ输出层分布，依据式(5)构建分类损失；

(4)根据干净样本$\boldsymbol{x}$和对抗样本$\boldsymbol{x}_\theta^{\mathrm{adv}}$在预训练模型和目标模型输出层分布，依据式(6)构建输出层对比损失；

(5)根据干净样本$\boldsymbol{x}$和对抗样本$\boldsymbol{x}_\theta^{\mathrm{adv}}$在预训练模型和目标模型中间层特征分布，依据式(7)构建特征层对比损失；

(6)综合步骤3)—5)，依据式(7)构建总体对比对抗损失函数，并最小化损失函数，优化目标模型f_θ；

Until 达到最大迭代次数。

3 实验结果与分析

3.1 数据集与实验设置

3.1.1 数据集

本文采用公开的高光谱图像数据集PaviaU和Indian Pines(http://www.ehu.eus/ccwintco/index.php/HyperspectralRemote)验证提出算法的有效性。

PaviaU数据集是由ROSIS光谱仪对意大利Pavia大学成像得到,该数据集空间分辨率为610×304像素,包含115个光谱波段,其空间分辨率为1.3 m,光谱分辨率为4 nm。实验前去除了12个噪声波段。数据集包含9个不同的地物类别,图2显示了PaviaU高光谱图像的伪彩色图和真值图。

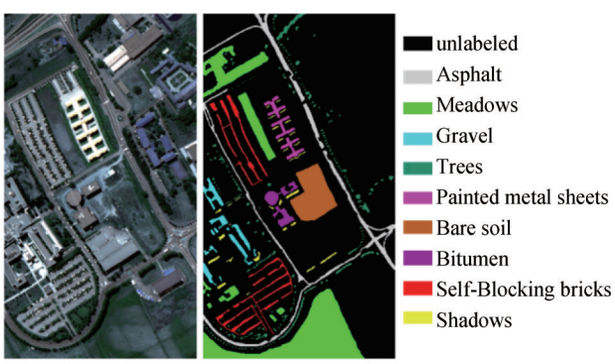

图2 PaviaU数据集

Fig. 2 PaviaU dataset

Indian Pines数据集是由机载可视红外成像光谱仪(airborne visible infrared imaging spectrometer, AVIRIS)在1992年对美国印第安纳州一块印度松树林进行成像得到,该数据集空间分辨率为145×145像素,包含220个光谱波段,其空间分辨率为20 m,光谱成像仪成像波长范围为0.4~2.5 μm。实验前去除了20个噪声波段,共有200个光谱波段和20个不同的地物类别用于实验,图3显示了Indian Pines高光谱图像的伪彩色图和真值图。

数据集划分:PaviaU数据集和Indian Pines数据集中每个类别选择的训练样本和测试样本数量如表1和表2所示。在后续所述的所有实验中,监督对抗训练方法中所有训练样本都作为标记样本;半监督对抗训练方法(包括提出方法)中训练样本的20%作为标记样本,其余训练样本作为无标记样本参与训练。

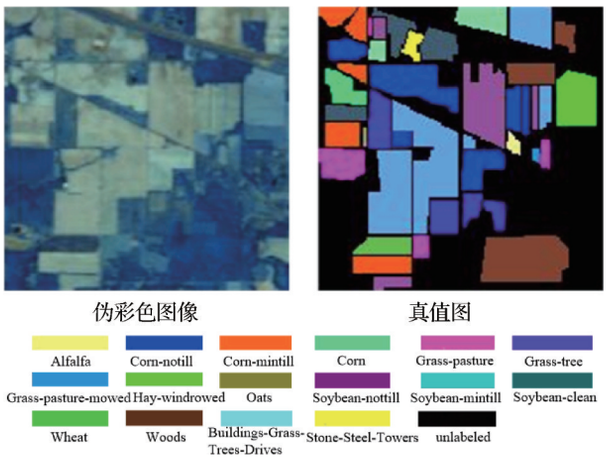

图3 Indian Pines数据集

Fig. 3 Indian Pines dataset

表1 PaviaU数据集训练样本和测试样本数量

Table 1 The number of training examples and test examples for PaviaU dataset

类别	训练样本	测试样本
1	900	5 731
2	900	177 490
3	900	1 199
4	900	2 164
5	900	445
6	900	4 129
7	900	430
8	900	2 782
9	900	47

3.1.2 实验设置

本文采用的对抗训练目标模型包括ResNet18(He等,2016)、VGG11(Simonyan和Zisserman,2015)、VGG19(Simonyan和Zisserman,2015)和Inception_V3(Szegedy等,2016),并采用FGSM(Goodfellow等,2015)、PGD(Madry等,2019)、APGD(Auto-PGD)(Croce和Hein,2020)、AutoAttack(Croce和Hein,2020)、MI-FGSM(Dong等,2018)、DI-FGSM(Xie等,2019)以及2.2节介绍的基于特征的攻击方法(feature-based attack, FBA)生成对抗样本对防御模型进行攻击,以验证防御模型的防御性能,这6种攻击方法的攻击参数设置如表3所示(FGSM只有攻击强度参数,无其他额外要说明的参数设置)。在后面的所有实验中,分别对原始高光谱数据集PaviaU和

表 2　Indian Pines 数据集训练样本和测试样本数量

Table 2　The number of training examples and test examples for Indian Pines dataset

类别	训练样本	测试样本
1	15	31
2	300	1 128
3	300	530
4	90	147
5	90	393
6	300	430
7	15	13
8	90	388
9	15	5
10	300	672
11	300	2 155
12	300	293
13	90	115
14	300	965
15	90	296
16	15	78

表 3　不同对抗攻击方法参数设置（扰动强度：0. 03）

Table 3　Parameter settings of different adversarial attack methods（perturbation budget：0. 03）

攻击方式	参数设置
FBA	迭代次数：20
PGD	步长：0.005，迭代次数：10
DI-FGSM	步长：2/255，迭代次数：20，转换概率：0.9
MI-FGSM	步长：2/255，迭代次数：20，衰减因子：1
APGD	迭代次数：10
AutoAttack	迭代次数：10

Indian Pines 利用相邻波段取均值方法降低光谱维数至 6 个波段进行实验分析，中间层特征均取自目标模型最后一个卷积层的特征，温度 τ 设置为 0. 2，高光谱样本尺寸为 32 × 32 像素。对抗训练中学习率设置为 0. 01，迭代次数设置为 200。超参数 α 和 β 均设置为 1，超参数的选择在 3. 5 节进行了进一步的分析。

3. 2　算法比较

在 PaviaU 和 Indian Pines 高光谱图像数据集上将提出的防御方法与 5 个主流的对抗训练方法进行了对比，包括 AT（adversarial training）（Goodfellow 等，2015）、RST（https://github.com/yaircarmon/semisup-adv）（Carmon 等，2019）、TRADES（https://github. com/yaodongyu/TRADES）（Zhang 等，2019）、MART（https://github. com/YisenWang/MART）（Wang 等，2020）和 SRT（https://github. com/THUYimingLi/Semi-supervised_Robust_Training）（Li 等，2022）。其中，AT、TRADES 为监督对抗训练方法，SRT、RST、MART 以及本文方法为半监督对抗训练方法。本文依据对比文献的公开代码，在高光谱图像数据集上重新运行得到对比实验结果，原因在于：一方面对比方法文献中采用自然图像作为测试数据集，而本文方法目的在于证明所提出防御方法在高维图像数据上的优越性；另一方面，为了保证对比实验的公平性，对比实验中对抗训练方法的对抗样本均由 2. 2 节介绍的 FBA 方法生成，目标模型为 ResNet18，实验结果如表 4 和表 5 所示。

从表 4 和表 5 可以看出，没有防御保护的目标模型极易受到对抗样本的攻击，在两组测试的高光谱图像数据集上，预训练的目标模型的分类精度显著降低。该现象表明，对抗样本的存在对高光谱图像分类网络的鲁棒性构成了严重的威胁。经过对抗训练后的目标模型防御已知和未知攻击的能力明显增强，同时也保证了干净样本的分类精度，但是这些方法的防御性能仍存在一些差别。

在两组高光谱图像数据集上，AT 方法能够有效防御 FBA 对抗样本，但是相比于其他对比方法，防御的泛化能力不足，无法有效抵御未知攻击。除此以外，监督对抗训练方法（TRADES）与半监督对抗训练方法（SRT、RST、MART）取得了相近的防御效果，但是半监督对抗训练方法仅利用了 20% 的训练样本作为标记样本参与对抗训练，而 TRADES 方法利用了所有的训练样本进行对抗训练。所以半监督对抗训练方法能够用更少的标记样本达到和监督对抗训练相近的防御能力。相较于以上对比方法，本文提出的 CSAT 方法在两组高光谱图像数据集上均取得了优于其他对比方法的防御性能，分类精度甚至超过监督对抗训练方法。在面对攻击力更强的 DI-FGSM 对抗样本和 MI-FGSM 对抗样本时，防御精度也提高了 3. 45% ～4. 71%。综合以上分析，本文方法在面对已知攻击和未知攻击时，均取得了优越的防御性能。

表4　在PaviaU数据集上不同对抗训练方法的防御性能对比（光谱波段数 = 6）

Table 4　Defense accuracy evaluation of different adversarial training methods on PaviaU dataset (number of spectral bands = 6)

防御方法	攻击方法							
	无	FBA	FGSM	PGD	MI-FGSM	DI-FGSM	APGD	AutoAttack
无	0.999 3	0.199 6	0.346 8	0.082 2	0.049 9	0.045 2	0.062 3	0.023 4
AT-FGSM（Goodfellow等，2015）	0.998 4	0.911 4	0.962 1	0.840 6	0.660 1	0.715 9	0.623 8	0.533 6
AT-FBA（Goodfellow等，2015）	0.999 1	0.995 8	0.866 3	0.788 9	0.759 3	0.783 1	0.740 6	0.724 2
SRT（Li等，2022）	**0.999 7**	0.996 3	0.898 0	0.777 2	0.682 7	0.730 5	0.642 3	0.636 1
RST（Carmon等，2019）	0.999 7	0.998 4	0.893 3	0.822 1	0.753 3	0.822 0	0.729 5	0.707 0
TRADES（Zhang等，2019）	0.999 7	**0.999 2**	0.925 7	0.841 9	0.792 8	0.842 8	0.755 1	0.741 6
MART（Wang等，2020）	0.998 8	0.996 7	0.889 1	0.806 6	0.753 3	0.794 9	0.725 9	0.705 6
本文	0.998 7	0.995 5	**0.980 8**	**0.957 7**	**0.939 9**	**0.963 7**	**0.937 3**	**0.935 1**

注：加粗字体表示各列最优结果。

表5　在Indian Pines数据集上不同对抗训练方法的防御性能对比（光谱波段数 = 6）

Table 5　Defense accuracy evaluation of different adversarial training methods on Indian Pines dataset (number of spectral bands = 6)

防御方法	攻击方法							
	无	FBA	FGSM	PGD	MI-FGSM	DI-FGSM	APGD	AutoAttack
无	0.997 0	0.341 6	0.255 6	0.163 1	0.033 9	0.141 9	0.123 1	0.115 1
AT-FGSM（Goodfellow等，2015）	0.998 7	0.925 1	**0.976 7**	0.900 5	0.812 9	0.860 8	0.823 9	0.822 1
AT-FBA（Goodfellow等，2015）	0.996 3	**0.996 3**	0.924 9	0.8919	0.857 2	0.879 1	0.851 2	0.849 3
SRT（Li等，2022）	0.997 3	0.994 5	0.939 2	0.912 7	0.886 8	0.906 4	0.880 3	0.873 5
RST（Carmon等，2019）	**0.998 7**	0.995 6	0.937 1	0.925 5	0.910 3	0.921 7	0.907 0	0.899 6
TRADES（Zhang等，2019）	0.997 4	0.994 4	0.954 2	0.932 1	0.911 3	0.930 5	0.910 9	0.902 7
MART（Wang等，2020）	0.997 3	0.995 8	0.942 9	0.924 7	0.914 7	0.932 8	0.910 0	0.905 3
本文	0.993 7	0.988 9	0.973 2	**0.967 8**	**0.961 8**	**0.967 3**	**0.962 9**	**0.960 8**

注：加粗字体表示各列最优结果。

3.3　消融实验

本节针对对比对抗损失（式（8））的每一项进行消融分析实验，以进一步验证本文方法的有效性。消融实验结果如表6和表7所示。对比对抗损失包含分类损失L_C、输出层对比损失L_O以及特征层对比损失L_F。

实验1仅利用分类损失L_C和输出层对比损失L_O对目标网络进行对抗训练；

实验2仅利用分类损失L_C和特征层对比损失L_F对目标网络进行对抗训练；

实验3仅利用输出层对比损失L_O和特征层对比损失L_F对目标网络进行对抗训练；

实验4同时利用分类损失L_C、输出层对比损失L_O和特征层对比损失L_F对目标网络进行对抗训练。

根据消融实验结果可以得出以下结论：

1）比较实验1、实验2和实验4的实验结果，在PaviaU高光谱图像数据集上，实验1的结果优于实验2，也就是说，重建对抗样本的输出层预测结果能够更加有效地提高目标网络的对抗鲁棒性；而在Indian Pines高光谱图像数据集上，实验2的结果优于实验1，对抗样本的中间层特征对目标网络对抗鲁棒性的提高更为重要。但是，在实验4中，联合输

出层对比损失和特征层对比损失优化模型，可以明显提高目标模型抵御未知攻击的能力。

2）比较实验3和实验4的实验结果，由于分类损失的约束，干净样本的分类精度有明显提升，尤其在类别数更多的Indian Pines数据集上，分类精度提升得更为明显。本文方法中对抗样本生成和目标网络重训练过程交替进行，目标网络对干净样本的准确分类可以得到高可靠性的对抗样本，从而提高对抗训练过程的可靠性。所以实验4在干净样本和防御攻击的性能上都有明显提升。

表6　在PaviaU数据集上的消融实验

Table 6　Ablation experiments on PaviaU dataset

实验	分类损失	输出层对比损失	特征层对比损失	无	FBA	FGSM	PGD	MI-FGSM	APGD	AutoAttack
–	–	–	–	0.999 3	0.199 6	0.346 8	0.082 2	0.049 9	0.062 3	0.023 4
1	√	√	–	0.999 1	0.994 2	0.978 4	0.944 1	0.922 3	0.918 2	0.915 2
2	√	–	√	**0.999 4**	0.988 2	0.964 8	0.941 7	0.926 3	0.921 7	0.917 1
3	–	√	√	0.997 8	0.989 0	0.969 5	0.948 6	0.914 6	0.907 1	0.900 3
4	√	√	√	0.998 7	**0.995 5**	**0.980 8**	**0.957 7**	**0.939 9**	**0.937 3**	**0.935 1**

注：加粗字体表示各列最优结果，“√”表示采用，“–”表示未采用。

表7　在Indian Pines数据集上的消融实验

Table 7　Ablation experiments on Indian Pines dataset

实验	分类损失	输出层对比损失	特征层对比损失	无	FBA	FGSM	PGD	MI-FGSM	APGD	AutoAttack
–	–	–	–	0.997 0	0.341 6	0.255 6	0.163 1	0.033 9	0.123 1	0.115 1
1	√	√	–	**0.995 3**	**0.990 9**	0.918 4	0.914 4	0.896 7	0.893 8	0.880 7
2	√	–	√	0.974 2	0.963 9	0.925 4	0.916 2	0.910 7	0.909 5	0.900 3
3	–	√	√	0.971 1	0.967 2	0.934 5	0.927 3	0.918 8	0.918 4	0.911 3
4	√	√	√	0.993 7	0.988 9	**0.973 2**	**0.967 8**	**0.961 8**	**0.962 9**	**0.960 8**

注：加粗字体表示各列最优结果，“√”表示采用，“–”表示未采用。

3.4　不同目标模型上的防御性能分析

以上实验的目标模型采用ResNet18，本节进一步分析了不同目标模型上提出方法的防御性能，包括VGG11，VGG19以及Inception_V3，实验结果如表8和表9所示，其中FBA为已知攻击，其余攻击方法为未知攻击。从表8和表9可以看出，VGG11、VGG19以及Inception_V3同样无法有效抵御高光谱对抗样本的攻击，而本文方法在两个数据集上均能有效提高不同目标模型的防御性能。

3.5　超参数分析

本节分析了PaviaU高光谱图像数据集上输出层对比损失权重α和特征层对比损失权重β对提出方法的影响，实验结果如图4所示。

图4中α和β的取值范围为[0.001, 100]，从图4中可以看出，3个子图的变化趋势基本一致。随着输出层对比损失权重α不断增大，抵御对抗攻击的防御效果相应增强，但是过大或过小的特征层对比损失权重β会导致目标模型抵御对抗攻击能力降低，可能的原因是权重β过小导致特征层对比损失失去作用，而权重β过大造成对抗训练的目标模型出现过拟合，泛化能力下降。因此，输出层对比损失权重α的取值建议为[1, 10]，特征层对比损失权重β的取值建议为[0.1, 1]。

4　结　论

本文提出了一种基于对比半监督学习的对抗训练方法，克服防御泛化能力依赖于对抗样本类型及数量的问题。首先通过最大化干净样本与对抗样本在目标模型上的特征距离，生成高迁移性对抗样本，其次利用少量标记样本预训练目标模型，并根据预训练的目标模型构建对比对抗损失函数，通过最小

表8　在PaviaU数据集上不同目标模型上的防御性能分析

Table 8　Defense performance analysis on different target models on PaviaU dataset

目标模型	防御方法	攻击方法							
		无	FBA	FGSM	PGD	MI-FGSM	DI-FGSM	APGD	AutoAttack
ResNet18	无	0.999 3	0.199 6	0.346 8	0.082 2	0.049 9	0.045 2	0.062 3	0.023 4
	本文	0.998 7	0.995 5	0.980 8	0.957 7	0.939 9	0.963 7	0.937 3	0.935 1
VGG11	无	0.998 6	0.145 7	0.225 9	0.057 1	0.048 5	0.062 1	0.044 3	0.033 1
	本文	0.999 1	0.980 5	0.985 8	0.969 4	0.951 9	0.968 6	0.950 3	0.946 6
VGG19	无	0.998 4	0.318 5	0.453 1	0.206 6	0.159 8	0.129 4	0.133 2	0.053 9
	本文	0.999 4	0.987 4	0.976 9	0.992 1	0.972 6	0.982 2	0.971 2	0.970 5
Inception_V3	无	0.998 4	0.164 1	0.559 7	0.147 9	0.102 3	0.104 0	0.151 9	0.025 1
	本文	0.999 1	0.997 1	0.998 9	0.996 5	0.993 6	0.994 9	0.993 5	0.988 2

表9　在Indian Pines数据集上不同目标模型上的防御性能分析

Table 9　Defense performance analysis on different target models on Indian Pines dataset

目标模型	防御方法	攻击方法							
		无	FBA	FGSM	PGD	MI-FGSM	DI-FGSM	APGD	AutoAttack
ResNet18	无	0.997 0	0.341 6	0.255 6	0.163 1	0.033 9	0.141 9	0.123 1	0.115 1
	本文	0.993 7	0.988 9	0.973 2	0.967 8	0.961 8	0.967 3	0.962 9	0.960 8
VGG11	无	0.996 8	0.049 8	0.220 8	0.026 6	0.006 3	0.018 2	0.016 8	0.001 8
	本文	0.996 3	0.982 3	0.973 5	0.969 1	0.963 3	0.967 1	0.963 6	0.949 2
VGG19	无	0.996 3	0.128 6	0.347 4	0.201 7	0.167 9	0.150 8	0.142 4	0.093 9
	本文	0.998 0	0.992 8	0.988 9	0.987 5	0.986 5	0.987 7	0.985 6	0.977 6
Inception_V3	无	0.992 3	0.162 5	0.624 1	0.153 1	0.084 3	0.113 0	0.145 9	0.008 1
	本文	0.995 1	0.975 5	0.970 4	0.960 1	0.951 3	0.956 1	0.951 2	0.935 9

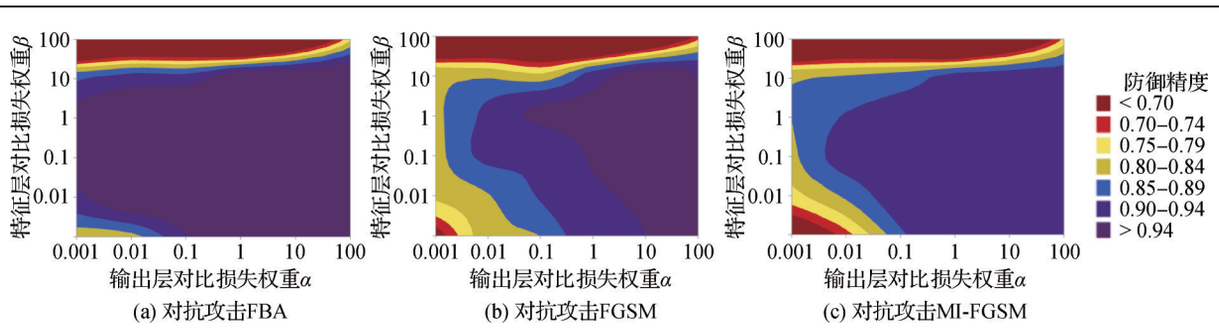

图4　在PaviaU数据集上输出层对比损失权重α和特征层对比损失权重β的参数研究

Fig. 4　Parameter study of contrastive weight α of output layer and contrastive weight β of feature layer on PaviaU dataset

((a) adversarial attack FBA; (b) adversarial attack FGSM; (c) adversarial attack MI-FGSM)

化对抗对比损失函数优化模型。本文充分挖掘了少量标记样本提供的特征信息,利用大量无标记样本进一步提高目标网络的防御性能。提出方法在PaviaU和Indian Pines高光谱图像数据集上进行对比和分析实验,与具有代表性的监督和半监督对抗训练方法相比,提出的对比半监督对抗训练方法取得了更加优越的防御性能,分类精度更是优于监督对抗训练方法,从而验证了提出方法防御高光谱图像对抗样本的有效性。

由于提出方法需要根据标记样本预训练目标模

型，并以此构造对比对抗损失函数，因此对预训练模型的精度具有较高的要求。下一步将针对防御性能依赖于预训练模型精度这一问题进行研究，进一步提高防御模型的泛用性。

参考文献(References)

Carlini N and Wagner D. 2017. Towards evaluating the robustness of neural networks//2017 IEEE Symposium on Security and Privacy (SP). San Jose, USA: IEEE: 39-57 [DOI: 10.1109/SP.2017.49]

Carmon Y, Raghunathan A, Schmidt L, Liang P and Duchi J C. 2019. Unlabeled data improves adversarial robustness//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 11192-11203

Croce F and Hein M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks//Proceedings of the 37th International Conference on Machine Learning. Virtual, Online: JMLR.org: 2206-2216

Dong Y P, Liao F Z, Pang T Y, Su H, Zhu J, Hu X L and Li J G. 2018. Boosting adversarial attacks with momentum//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9185-9193 [DOI: 10.1109/CVPR.2018.00957]

Dong Y P, Pang T Y, Su H and Zhu J. 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4307-4316 [DOI: 10.1109/CVPR.2019.00444]

Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C W, Prakash A, Kohno T and Song D. 2018. Robust physical-world attacks on deep learning visual classification//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1625-1634 [DOI: 10.1109/CVPR.2018.00175]

Goodfellow I J, Shlens J and Szegedy C. 2015. Explaining and harnessing adversarial examples [EB/OL]. [2023-07-16]. http://arxiv.org/pdf/1412.6572.pdf

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hendrycks D, Lee K and Mazeika M. 2019. Using pre-training can improve model robustness and uncertainty//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: ICML: 2712-2721

Huang Q, Katsman I, Gu Z Q, He H, Belongie S and Lim S N. 2019. Enhancing adversarial example transferability with an intermediate level attack//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4732-4741 [DOI: 10.1109/ICCV.2019.00483]

Jin G Q, Shen S W, Zhang D M, Dai F and Zhang Y D. 2019. APE-GAN: adversarial perturbation elimination with GAN//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK: IEEE: 3842-3846 [DOI: 10.1109/ICASSP.2019.8683044]

Kang X D, Duan P H, Xiang X L, Li S T and Benediktsson J A. 2018. Detection and correction of mislabeled training samples for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 56(10): 5673-5686 [DOI: 10.1109/TGRS.2018.2823866]

Kurakin A, Goodfellow I and Bengio S. 2017. Adversarial examples in the physical world [EB/OL]. [2023-07-16]. http://arxiv.org/pdf/1607.02533.pdf

Lamb A, Verma V, Kawaguchi K, Matyasko A, Khosla S, Kannala J and Bengio Y. 2022. Interpolated adversarial training: achieving robust neural networks without sacrificing too much accuracy. Neural Networks, 154: 218-233 [DOI: 10.1016/j.neunet.2022.07.012]

Li X and Li F X. 2017. Adversarial examples detection in deep networks with convolutional filter statistics//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5775-5783 [DOI: 10.1109/ICCV.2017.615]

Li Y M, Wu B Y, Feng Y, Fan Y B, Jiang Y, Li Z F and Xia S T. 2022. Semi-supervised robust training with generalized perturbed neighborhood. Pattern Recognition, 124: #108472 [DOI: 10.1016/j.patcog.2021.108472]

Lin J D, Song C B, He K, Wang L W and Hopcroft J E. 2020. Nesterov accelerated gradient and scale invariance for adversarial attacks [EB/OL]. [2023-07-16]. http://arxiv.org/pdf/1908.06281.pdf

Machado G R, Silva E and Goldschmidt R R. 2021. Adversarial machine learning in image classification: a survey toward the defender's perspective. ACM Computing Surveys, 55(1): #8 [DOI: 10.1145/3485133]

Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A. 2019. Towards deep learning models resistant to adversarial attacks [EB/OL]. [2023-07-16]. http://arxiv.org/pdf/1706.06083.pdf

Miyato T, Maeda S I, Koyama M and Ishii S. 2019. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8): 1979-1993 [DOI: 10.1109/TPAMI.2018.2858821]

Moosavi-Dezfooli S M, Fawzi A and Frossard P. 2016. DeepFool: a simple and accurate method to fool deep neural networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2574-2582 [DOI: 10.1109/CVPR.2016.282]

Naseer M, Khan S H, Rahman S and Porikli F. 2019. Task-generalizable adversarial attack based on perceptual metric [EB/OL]. [2023-07-16]. http://arxiv.org/pdf/1811.09020.pdf

Papernot N, Mcdaniel P, Wu X, Jha S and Swami A. 2016. Distillation as a defense to adversarial perturbations against deep neural networks//2016 IEEE Symposium on Security and Privacy (SP). San

Jose, USA: IEEE: 582-597 [DOI: 10.1109/sp.2016.41]

Ren Y C, Zhu H G, Sui X Y and Liu C. 2023. Crafting transferable adversarial examples via contaminating the salient feature variance. Information Sciences, 644: #119273 [DOI: 10.1016/j.ins.2023.119273]

Shi C, Dang Y N, Fang L, Lyu Z Y and Zhao M H. 2022. Hyperspectral image classification with adversarial attack. IEEE Geoscience and Remote Sensing Letters, 19: #5510305 [DOI: 10.1109/lgrs.2021.3122170]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2023-07-16]. http://arxiv.org/pdf/1409.1556.pdf

Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z. 2016. Rethinking the inception architecture for computer vision//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2818-2826 [DOI: 10.1109/CVPR.2016.308]

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I and Fergus R. 2014. Intriguing properties of neural networks [EB/OL]. [2023-07-16]. http://arxiv.org/pdf/1312.6199.pdf

Tan K, Wang X and Du P J. 2019. Research progress of the remote sensing classification combining deep learning and semi-supervised learning. Journal of Image and Graphics, 24(11): 1823-1841 (谭琨, 王雪, 杜培军. 2019. 结合深度学习和半监督学习的遥感影像分类进展. 中国图象图形学报, 24(11): 1823-1841) [DOI: 10.11834/jig.190348]

Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D and McDaniel P. 2020. Ensemble adversarial training: attacks and defenses [EB/OL]. [2023-07-16]. http://arxiv.org/pdf/1705.07204.pdf

Uesato J, Alayrac J B, Huang P S, Stanforth R, Fawzi A and Kohli P. 2019. Are labels required for improving adversarial robustness? // Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 12214-12223

Wang X S and He K. 2021a. Enhancing the transferability of adversarial attacks through variance tuning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1924-1933 [DOI: 10.1109/CVPR46437.2021.00196]

Wang Y S, Zou D F, Yi J F, Bailey J, Ma X J and Gu Q Q. 2020. Improving adversarial robustness requires revisiting misclassified examples//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: 1-14 [Online]

Wang Z B, Guo H C, Zhang Z F, Liu W X, Qin Z and Ren K. 2021b. Feature importance-aware transferable adversarial attacks//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 7619-7628 [DOI: 10.1109/ICCV48922.2021.00754]

Wu L, Zhu Z X, Tai C and E W N. 2018. Understanding and enhancing the transferability of adversarial examples [EB/OL]. [2023-07-16]. http://arxiv.org/pdf/1802.09707.pdf

Xie C H, Wang J Y, Zhang Z S, Ren Z and Yuille A. 2018. Mitigating adversarial effects through randomization [EB/OL]. [2018-02-28]. http://arxiv.org/pdf/1711.01991.pdf

Xie C H, Zhang Z S, Zhou Y Y, Bai S, Wang J Y, Ren Z and Yuille A L. 2019. Improving transferability of adversarial examples with input Diversity//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2725-2734 [DOI: 10.1109/CVPR.2019.00284]

Yang S S, Yang Y, Zhou L N, Zhan R and Man Y F. 2022. Intermediate-layer transferable adversarial attack with DNN attention. IEEE Access, 10: 95451-95461 [DOI: 10.1109/access.2022.3204696]

Yuan L, Li X M, Pan Z X, Sun J M and Xiao L. 2022. Review of adversarial examples for object detection. Journal of Image and Graphics, 27(10): 2873-2896 (袁珑, 李秀梅, 潘振雄, 孙军梅, 肖蕾. 2022. 面向目标检测的对抗样本综述. 中国图象图形学报, 27(10): 2873-2896) [DOI: 10.11834/jig.210209]

Zhang H Y, Yu Y D, Jiao J T, Xing E P, El Ghaoui L and Jordan M I. 2019. Theoretically principled trade-off between robustness and accuracy//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: ICML: 7472-7482

Zhang J P, Wu W B, Huang J T, Huang Y Z, Wang W X, Su Y X and Lyu M R. 2022a. Improving adversarial transferability via neuron attribution-based attacks//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 14973-14982 [DOI: 10.1109/CVPR52688.2022.01457]

Zhang X R, Chen S T, Zhu P, Tang X, Feng J and Jiao L C. 2022b. Spatial pooling graph convolutional network for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 60: #5521315 [DOI: 10.1109/tgrs.2022.3140353]

Zhou D W, Liu T L, Han B, Wang N N, Peng C and Gao X. 2021. Towards defending against adversarial examples via attack-invariant features. InInternational Conference on Machine Learning. ICML: 12835-12845 [DOI: 10.48550/arXiv.2106.0503]

Zhu H G, Zheng H R, Zhu Y and Sui X Y. 2023. Boosting the transferability of adversarial attacks with adaptive points selecting in temporal neighborhood. Information Sciences, 641: #119081 [DOI: 10.1016/j.ins.2023.119081]

作者简介

石程,女,副教授,主要研究方向为遥感图像处理和对抗机器学习。E-mail: c_shi@xaut.edu.cn

赵明华,通信作者,女,教授,主要研究方向为图像处理和计算机视觉。E-mail: zhaominghua@xaut.edu.cn

刘莹,女,硕士研究生,主要研究方向为高维图像分类中的对抗防御方法。E-mail: ying_liu999@163.com

苗启广,男,教授,主要研究方向为计算机视觉和模式识别。E-mail: qgmiao@xidian.edu.cn

潘治文,男,教授,主要研究方向为计算机视觉、模式识别和对抗机器学习。E-mail: cmpun@umac.mo