

# 用于防御高光谱图像分类中对抗攻击的上下文感知框架

贺旺泉

## 摘要

深度神经网络在高光谱图像处理中发挥着重要作用，但当使用对抗样本（通过对干净样本添加微小扰动生成）进行训练时，模型容易被欺骗。这些扰动对人眼是不可见的，但很容易导致深度学习模型出现错误的分类。最近一项关于高光谱图像分类中对抗样本防御的研究通过利用全局上下文信息提高了深度网络的鲁棒性。然而，现有的方法没有区分不同类别的上下文信息，不可避免地引入了干扰信息。针对这一问题，本文提出了一种鲁棒的上下文感知网络，能够在高光谱图像分类中防御错误的对抗样本。该网络通过聚合空洞卷积学到的特征来生成全局上下文表示，然后通过构建类上下文感知学习场景（包括亲和力损失）来显式建模类内和类间上下文信息，以进一步细化全局上下文，帮助像素获得更可靠的长程依赖关系，提高模型对对抗攻击的整体鲁棒性。在基准数据集上的实验结果表明，该方法比其他先进方法具有更好的鲁棒性和泛化性。

**关键词：**深度学习；高光谱图像；对抗攻击；对抗样本；全局上下文

## 1 引言

随着卫星传感技术的快速发展，地球观测仪器获取的遥感图像从全色到多光谱再到数百个连续波段的高光谱。高光谱图像具有丰富的光谱信息，可以比其他成像方法更准确地表征地物的物理性质。高光谱遥感影像分类是一项关键任务，其目的是为每个像元分配一个唯一的标签，广泛应用于许多遥感应用（包括军事侦察、农业生产、环境监测、矿产勘查等诸多领域<sup>[1-4]</sup>）。

卷积神经网络（Convolutional neural networks, CNNs）也因其强大的拟合能力而被广泛应用于 HSI 分类任务中<sup>[5-7]</sup>。为了降低 3DCNN 带来的复杂性，Roy 等人<sup>[8]</sup>提出了一种 2D 和 3DCNN 混合分类框架。Xie 等人<sup>[9]</sup>提出了一种基于 CNN 的密集残差网络来获取多尺度特征以提高分类精度。这些方法的性能证明了 CNN 在特征提取任务中的有效性。为了进一步提高 HSI 的解译效果，人们探索了其他结合 CNN 的方法。基于 CNN 的空谱双分支网络也成为研究热点。这些网络包括 CNN 和 RNN 或其变体组合<sup>[10-12]</sup>。Hong 等人<sup>[13]</sup>探索了图卷积网络与 CNN 的融合用于高光谱图像分类，并提出了一种小批量 GCN，以减少传统图卷积构建邻接矩阵所带来的高计算成本。这些方法通常具有独立的光谱和空间特征流，并将得到的空谱特征向量进行融合以产生最终结果。此外，Transformer 结合 CNN 在高光谱图像处理中也得到了广泛的探索<sup>[14-16]</sup>。

上述基于深度学习的方法在高光谱图像处理中表现出了优异的性能，但它们通常处理的是干净的图像。事实上，深度学习模型非常容易受到对抗攻击。图 1 给出了对深度学习模型的对抗性攻击的图示说明。如图所示，在干净样本上取得优异性能的模型在对抗样本上的表现可能非常差。对抗样本由原始输入的扰动而产生，深度网络在处理此类样本时，往往对错误类别的标签具有较高的置信度。从人类的视觉角度来看，对抗样本和干净样本之间没有差异。这一现象的存在限制了深度学习模型在一些安全性要求较高的领域的实际适用性，如军事侦查和医疗诊断应用中。

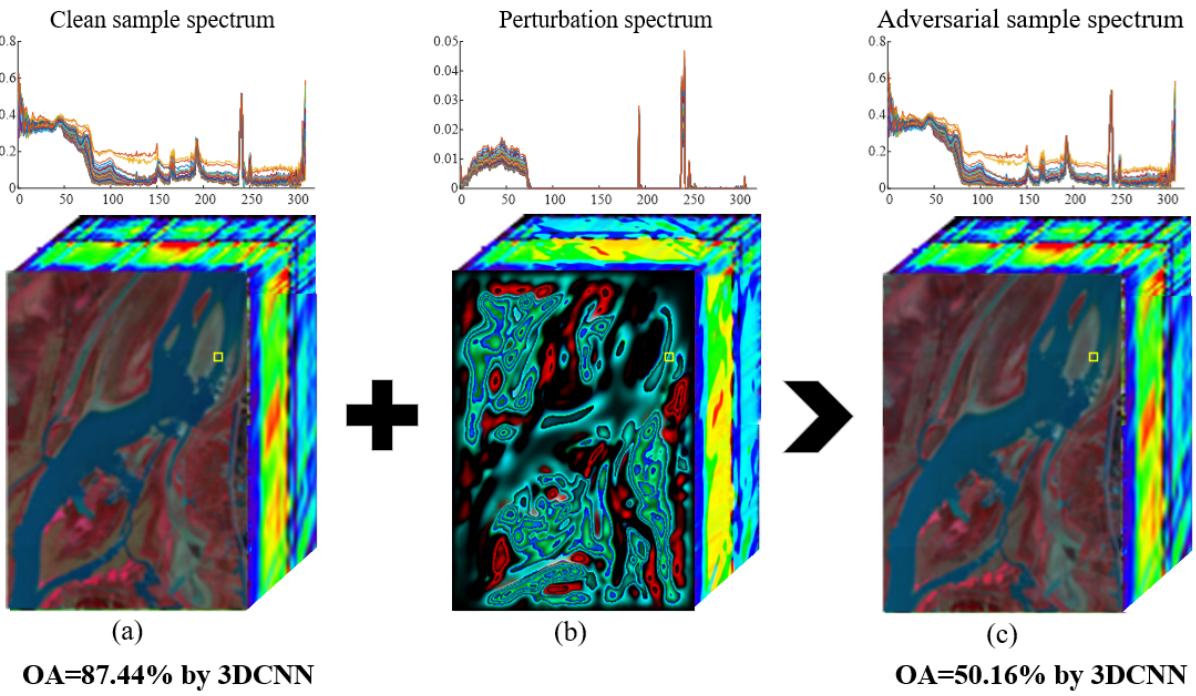


图 1: 高光谱图像中对抗样本攻击示意图

在遥感领域, Czaja 等人<sup>[17]</sup>首先研究了对抗攻击对遥感图像分类的影响, 其结果表明分类模型对对抗样本的鲁棒性普遍较低。针对遥感图像场景分类中的对抗攻击, 在<sup>[18]</sup>中提出了一种基于扰动搜索生成对抗网络的防御框架。合成孔径雷达 (Synthetic Aperture Radar, SAR) 中的一些对抗防御策略也被广泛研究<sup>[19-21]</sup>。遥感领域对对抗样本的研究主要集中在 RGB 和 SAR 图像上。然而, 高光谱处理中的对抗攻击威胁也值得关注。从图 1 可以看出, 对抗攻击在高光谱图像中产生的光谱和空间变化是人眼无法察觉的, 但这些变化会极大地影响深度学习模型的分类性能。在<sup>[22]</sup>中首次明确了对抗样本对高光谱图像的影响, 这些结果表明大多数先进的高光谱分类方法容易被对抗样本欺骗。Park 等人<sup>[23]</sup>提出了一种结合随机光谱采样和光谱形状特征编码的框架来提高高光谱图像分类的鲁棒性, 充分利用高光谱图像中存在的丰富光谱信息。在<sup>[24]</sup>中提出一种改进的 DeepFool 算法, 通过在原始训练集中添加生成的边界对抗样本来处理高光谱图像中的对抗攻击。

对抗训练可以有效提高网络的鲁棒性, 但也不可避免地增加了训练负担。在对抗训练中, 训练好的模型被未知的对抗样本反复攻击<sup>[25]</sup>。设计鲁棒的网络结构来提高模型应对对抗样本是一种有效的防御策略。Xu 等人<sup>[22]</sup>提出了一种自注意力上下文网络来解决高光谱图像分类中的对抗攻击问题。该网络通过在特征图中建立像素之间的关系来捕获全局上下文信息, 并将所分配像素的损失与其他像素共享以建立全局关系, 提高模型的鲁棒性。然而, 这种方式产生的上下文信息不可避免地引入了与被分配像素的不同类别之间的长距离依赖关系, 类间长程依赖关系容易降低基于对抗样本的模型的预测精度。

针对这些问题, 本文提出了一种鲁棒的类上下文感知网络, 专门用于处理高光谱图像分类中的对抗性样本。该网络通过构建包含类内和类间关系的可学习亲和矩阵作为类内和类间上下文先验信息来区分类内和类间上下文信息。整个网络由三个主要阶段组成。首先, 使用一个骨干网络 (基于空洞卷积) 来生成特征图。其次, 设计聚合模块对空谱信息进行聚合; 在第三阶段, 根据聚合的特征生成相应的先验矩阵, 并产生一个类上下文先验 (具有亲和力损失) 来监督先验矩阵中学习到的类内和类间关系。利用学习到的类上下文先验信息, 可以有效增强类内相似性和类间差异性, 使捕获到的全局上下

文信息更有利于模型抵御对抗样本。

本文其余部分的结构如下。第二节回顾了相关工作。第3节详细描述了提出的类上下文感知网络。第四节给出了复现细节。第五节对实验结果进行了分析。最后总结了整个报告，并对未来研究方向提供了展望。

## 2 相关工作

在本小节中，我们简要回顾了几种可用于攻击深度学习算法的代表性白盒攻击方法的技术细节。

### 2.1 对抗攻击白盒算法

1) *Fast gradient sign method (FGSM)*: 为了实现有效的对抗训练，提高神经网络对对抗样本的鲁棒性，在<sup>[26]</sup>中提出了一种名为 FGSM 的攻击方法。该模型的核心思想是在损失函数中使梯度上升，从而使模型做出错误的决策。

给定一张图  $I$  和一个目标标签  $\hat{y}$ , 一个由 FSGM 生成的对抗样本  $I'$  可以表示如下:

$$I' = \text{clip}(I - \varepsilon \cdot \text{sign}(\nabla_I J(\theta, I, \hat{y}))), \quad (1)$$

其中  $\varepsilon$  是一个限制扰动范数的值,  $\nabla_I J(\theta, I, \hat{y})$  表示输入样本  $I$  的目标函数梯度,  $\text{clip}(\cdot)$  表示剪切函数,  $\text{sign}(\cdot)$  是符号函数。通过修改  $\varepsilon$  的值, 我们可以生成具有不同扰动强度的对抗样本。

2) *Projected gradient descent (PGD)*: 作为增强版的 FSGM, PGD<sup>[27]</sup> 通过以更小的步长迭代优化器多次提高攻击的有效性, 并将从每次迭代中学习到的对抗样本投影到规定的范围内。具体来说, 给定一个图像  $I$  和其对应的标签  $\hat{y}$ , 每次迭代生成的对抗样本可表示为:

$$I'_{t+1} = \text{clip}(I'_t - \varepsilon \cdot \text{sign}(\nabla_I J(\theta, I_t, \hat{y}))), \quad (2)$$

其中  $I'_t$  表示迭代  $t$  时生成的对抗样本, 当  $t=0$  时,  $I'_t$  是干净的样本  $I$ .

3) *Carlini and Wagner (C&W)*: C&W 是一种基于优化的攻击方法<sup>[28]</sup>, 其目标是使扰动尽可能小(同时确保模型被欺骗)。用该方法优化扰动的过程可表示为:

$$\min \|I - I'\|_2^2 + c \cdot f(I', \hat{y}) \quad \text{s.t. } I' \in [0, 1]^m, \quad (3)$$

当输入为  $(I, y)$  时,  $f(I', \hat{y})$  表示对应的损失函数,  $c$  是一个与攻击性能有关的超参数。

## 3 本文方法

所提出的模型的体系结构如图 4 所示。它由三个主要阶段组成。在第一阶段, 我们使用三个具有不同空洞率的空洞卷积层和两个池化层进行上下文特征提取。在第二阶段, 我们使用特征聚合模块对获得的特征的上下文进行建模, 该模块由并行的全局和局部上下文校准器组成。在第三阶段, 我们引入亲和学习模块, 学习捕获的上下文特征的类内和类间上下文关系, 为最终的分类获得可靠的上下文关系。下面我们详细描述所提出的类上下文感知网络。

### 3.1 特征聚合

为了获得有效的上下文特征, 我们设计了一个特征聚合模块, 根据全局和局部上下文对第一阶段学习到的特征进行聚合。特征聚合模块包括两个并行上下文校准器, 如图所示 2。

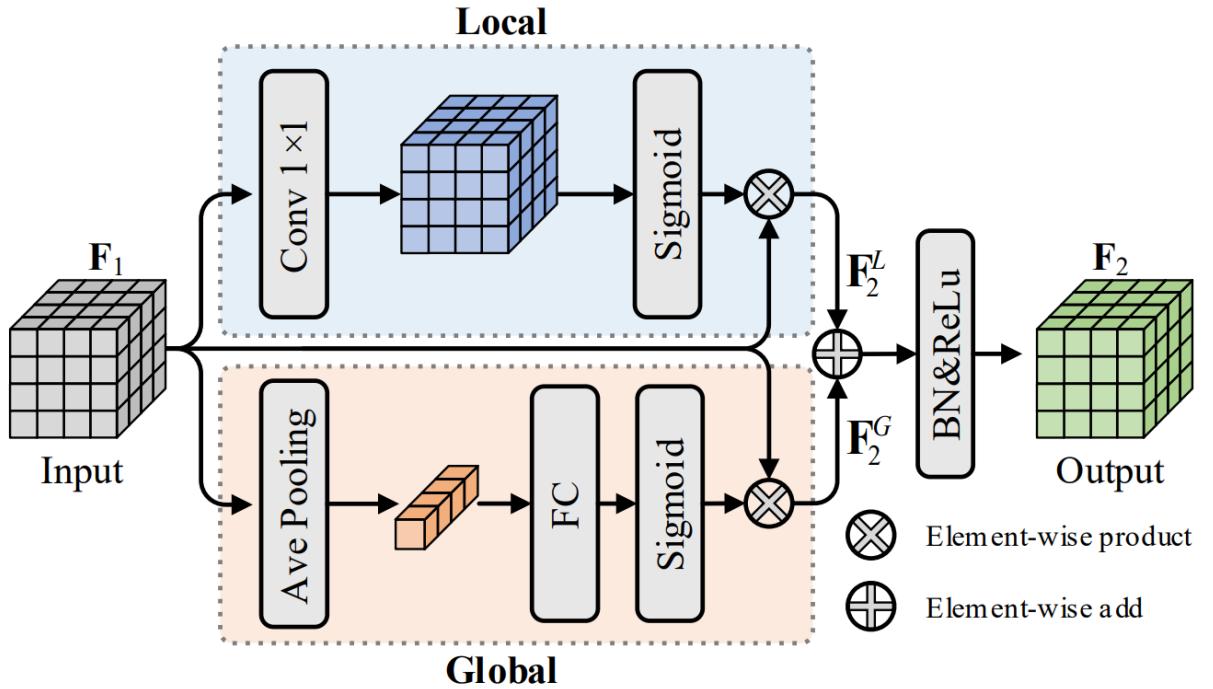


图 2: 特征聚合模块示意图

1) 局部聚合: 局部聚合分支关注像素的局部邻域。对于输入特征  $\mathbf{F}_1 \in \mathbb{R}^{h_0 \times w_0 \times c_0}$ , 我们首先应用一个大小为  $c_0 \times 1 \times 1$  的卷积核进行局部运算, 然后对得到的特征应用 sigmoid 函数估计权重, 最后对原始输入特征映射  $\mathbf{F}_1$  进行逐元素积运算, 该过程可以表示为:

$$\mathbf{F}_2^L = \sigma(\text{Conv}_{1 \times 1}(\mathbf{F}_1)) \otimes \mathbf{F}_1, \quad (4)$$

其中  $\mathbf{F}_2^L \in \mathbb{R}^{h_0 \times w_0 \times c_0}$  表示局部聚合分支的输出,  $\sigma$  为 sigmoid 函数,  $\otimes$  表示按元素相乘。

2) 全局聚合: 全局聚合分支从全局上下文中更新特性的重要性。为了实现该聚合, 首先在输入特征  $\mathbf{F}_1$  上进行全局平均池化, 然后将处理后的一维向量反馈到全连接层用于提高特征间的交互。最后, 利用 sigmoid 函数获取通道权重。在获得权重向量之后, 我们使用元素复制操作使结果向量的形状大小和  $\mathbf{F}_1$  的形状大小相等, 并将展开的向量应用于原始输入, 如下所示:

$$\mathbf{F}_2^G = \text{Expand}(\sigma(\text{FC}(\text{GAP}(\mathbf{F}_1)))) \otimes \mathbf{F}_1, \quad (5)$$

其中  $\mathbf{F}_2^G \in \mathbb{R}^{h_0 \times w_0 \times c_0}$  表示全局聚合分支的输出。FC and GAP 分别表示全连接层和全局平均池化操作。在获得两个并行分支的输出后, 最终的特征输出  $\mathbf{F}_2 \in \mathbb{R}^{h_0 \times w_0 \times c_0}$  由以下输出融合得到:

$$\mathbf{F}_2 = \delta(\text{BN}(\mathbf{F}_2^L \oplus \mathbf{F}_2^G)), \quad (6)$$

其中  $\delta$  为 ReLU 函数, BN 表示批处理归一化层,  $\oplus$  表示按元素相加。

### 3.2 类上下文先验学习

为了更好地捕获类内和类间上下文, 我们引入类上下文亲和学习来提高全局上下文捕获的可靠性。具体来说, 我们在第二阶段为特征映射  $\mathbf{F}_2$  生成亲和预测和相应的亲和标签。亲和性预测包括特征图  $\mathbf{F}_2 \in \mathbb{R}^{h_0 \times w_0 \times c_0}$  中的所有像素, 亲和标签可以通过亲和损失函数的监督, 引导亲和预测中的每个像素建立类内和类间的上下文关系。

为了生成亲和度预测图, 我们使用  $c_1 \times 1 \times 1$  卷积层和 BN 层来改变特征图  $\mathbf{F}_2$  的维数, 得到特征图  $\mathbf{F}_3 \in \mathbb{R}^{h_0 \times w_0 \times c_1}$ , 将其转化成大小为  $c_1 \times c_1$  的特征图, 然后利用 sigmoid 函数生成亲和度预测图  $\hat{\mathbf{A}} \in \mathbb{R}^{c_1 \times c_1}$ , 其中  $c_1 = h_0 \times w_0$ 。

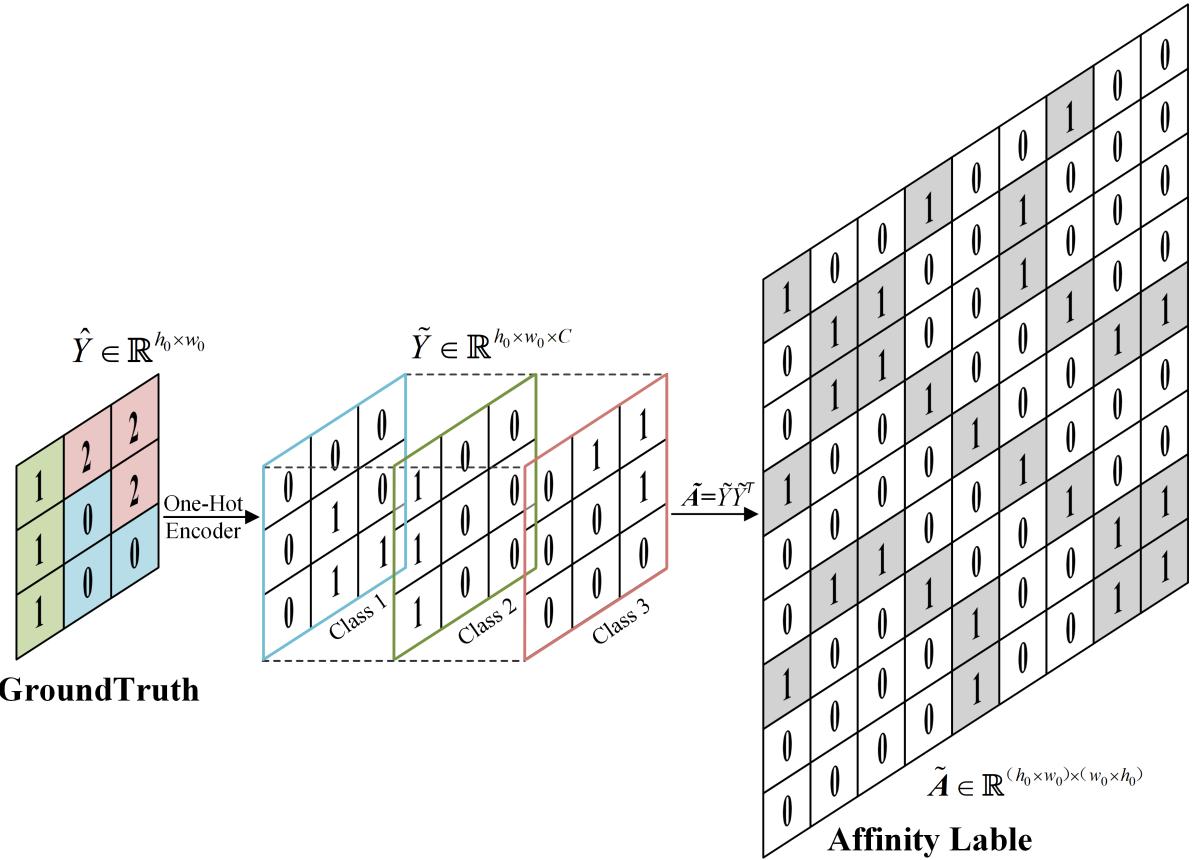


图 3: 亲和标签生成过程示意图

从原始高光谱图像对应的 **GroundTruth**, 可知一个像素是否与其他像素属于同一类别。因此, 我们可以根据基本事实生成具有类内和类间上下文先验的亲和标签映射。如图 3 所示, 我们将 **GroundTruth** 的大小下采样至  $\mathbf{F}_1$  来得到一个更新后的 **GroundTruth**  $\hat{Y} \in \mathbb{R}^{h_0 \times w_0}$ , 然后用 one-hot 操作对每个标签  $\hat{Y}$  进行编码, 得到变换矩阵  $\tilde{Y} \in \mathbb{R}^{h_0 \times w_0 \times C}$ , 其中  $C$  表示类别数。在此之后, 我们将编码后的矩阵转化为  $(h_0 \times w_0) \times C$ , 并将其乘以其转置矩阵, 得到亲和标签映射  $\tilde{\mathbf{A}}$ 。该过程可以表示为:

$$\tilde{\mathbf{A}} = \tilde{Y}\tilde{Y}^\top \in \mathbb{R}^{c_1 \times c_1}. \quad (7)$$

亲和标签  $\tilde{\mathbf{A}}$  中的类内和类间上下文先验被表示为二进制映射。值为 1 表示对应的行像素和列像素属于同一个类, 而值为 0 表示它们属于不同的类。为了监督亲和度预测图  $\hat{\mathbf{A}}$  并从亲和标签映射  $\tilde{\mathbf{A}}$  中学习类内和类间上下文信息, 可以考虑引入二元交叉熵损失函数来解决如下问题:

$$\mathcal{L}_b = -\frac{1}{(c_1)^2} \sum_{i=1}^{(c_1)^2} (\tilde{a}_i \log \hat{a}_i + (1 - \tilde{a}_i) \log (1 - \hat{a}_i)), \quad (8)$$

其中  $\{\hat{a}_i \in \hat{\mathbf{A}}, \tilde{a}_i \in \tilde{\mathbf{A}}, i \in [1, (c_1)^2]\}$ 。亲和度预测映射  $\mathbf{F}_2$  中的每一行像素对应于特征映射中的一个像素。一元损失忽略了像素之间的语义相关性, 为了进一步提高像素之间的语义相关性, 将像素中的亲和度预测映射  $\hat{\mathbf{A}}$  分为类内和类间像素。二元交叉熵损失的全局项  $\mathcal{L}_g$  被引入为类内上下文和类间上下文之间的关系建模, 其定义如下:

$$\mathcal{T}_j^p = \log \frac{\sum_{i=1}^{c_1} \tilde{a}_{ij} \hat{a}_{ij}}{\sum_{i=1}^{c_1} \hat{a}_{ij}} \quad (9)$$

$$\mathcal{T}_j^r = \log \frac{\sum_{i=1}^{c_1} \tilde{a}_{ij} \hat{a}_{ij}}{\sum_{i=1}^{c_1} \tilde{a}_{ij}}, \quad (10)$$

$$\mathcal{T}_j^s = \log \frac{\sum_{i=1}^{c_1} (1 - \tilde{a}_{ij}) (1 - \hat{a}_{ij})}{\sum_{i=1}^{c_1} (1 - \tilde{a}_{ij})} \quad (11)$$

$$\mathcal{L}_g = -\frac{1}{c_1} \sum_{j=1}^{c_1} (\mathcal{T}_j^p + \mathcal{T}_j^r + \mathcal{T}_j^s) \quad (12)$$

其中  $\mathcal{T}_j^p$ ,  $\mathcal{T}_j^r$  和  $\mathcal{T}_j^s$  分别表示精确率、召回率和特异性。在亲和损失中, 权重  $\lambda_b$  and  $\lambda_g$  被引入来平衡一元损失  $\mathcal{L}_b$  和全局项  $\mathcal{L}_g$ , 该过程可表示为:

$$\mathcal{L}_{aff} = \lambda_b \mathcal{L}_b + \lambda_g \mathcal{L}_g, \quad (13)$$

其中  $\mathcal{L}_{aff}$  表示亲和损失。在该报告中  $\lambda_b$  and  $\lambda_g$  被设置为 1.

### 3.3 类上下文感知网络

在亲和学习中, 可以使用亲和损失来监督亲和预测图, 学习亲和标记图中的类内和类间上下文先验信息, 进一步细化全局上下文特征。应用亲和预测映射  $\hat{A}$  可以实现显式的类内上下文学习得到特征图  $\mathbf{F}_2$ , 该过程可以表示为:

$$\mathcal{Z}_{tra} = \hat{A} \mathbf{F}'_2 \in \mathbb{R}^{c_1 \times c_0}, \quad (14)$$

其中  $\mathcal{Z}_{tra}$  表示类内上下文特征,  $\mathbf{F}'_2$  表示通过转化特征图  $\mathbf{F}_2$  到  $c_1 \times c_0$  ( $c_1 = h_0 \times w_0$ ) 所生成的. 同理, 通过构造相反的上下文亲和预测图, 类间上下文显式学习可以表示为:

$$\mathcal{Z}_{ter} = (\mathbf{1} - \hat{A}) \mathbf{F}'_2 \in \mathbb{R}^{c_1 \times c_0}, \quad (15)$$

其中  $\mathcal{Z}_{ter}$  表示类间上下文特征,  $\mathbf{1}$  表示值为 1 的矩阵, 其大小与  $\hat{A}$  一致。通过这种类内和类间显式学习, 可以选择性地强调特征图中像素的类内和类间上下文。

在分类阶段, 我们将阶段 1 中生成的特征图与阶段 2 中经过类内和类间上下文先验学习后的上下文特征合并, 其过程可以表示为:

$$\mathbf{F} = \text{Concat}(\mathbf{F}_2, \mathcal{Z}'_{tra}, \mathcal{Z}'_{ter}) \in \mathbb{R}^{h_0 \times w_0 \times 3c_0}, \quad (16)$$

其中  $\mathcal{Z}'_{tra}$  和  $\mathcal{Z}'_{ter}$  分别表示通过转化  $\mathcal{Z}_{tra}$  和  $\mathcal{Z}_{ter}$  到  $h_0 \times w_0 \times c_0$  得到的特征图,  $\mathbf{F}$  表示阶段 3 融合后的输出特征图, 最后的结果图由卷积层进行 softmax 和上采样得到。假设预测结果为  $\hat{Y}$ , 标签为  $\mathbf{Y}$ , 总损失函数  $\mathcal{L}$  可定义如下

$$\mathcal{L}_{cls} = -\frac{1}{h_0 w_0} \sum_{i=1}^{h_0} \sum_{j=1}^{w_0} \sum_{k=1}^C \mathbf{Y}_{(i,j,k)} \log \left( \hat{Y}_{(i,j,k)} \right), \quad (17)$$

$$\mathcal{L} = \lambda_{aff} \mathcal{L}_{aff} + \lambda_{cls} \mathcal{L}_{cls}, \quad (18)$$

其中  $\mathcal{L}_{cls}$  表示分割损失,  $\lambda_{aff}$  和  $\lambda_{cls}$  表示平衡  $\mathcal{L}_{aff}$  和  $\mathcal{L}_{cls}$  的权重. 在该报告中, 它们被设置为 1。

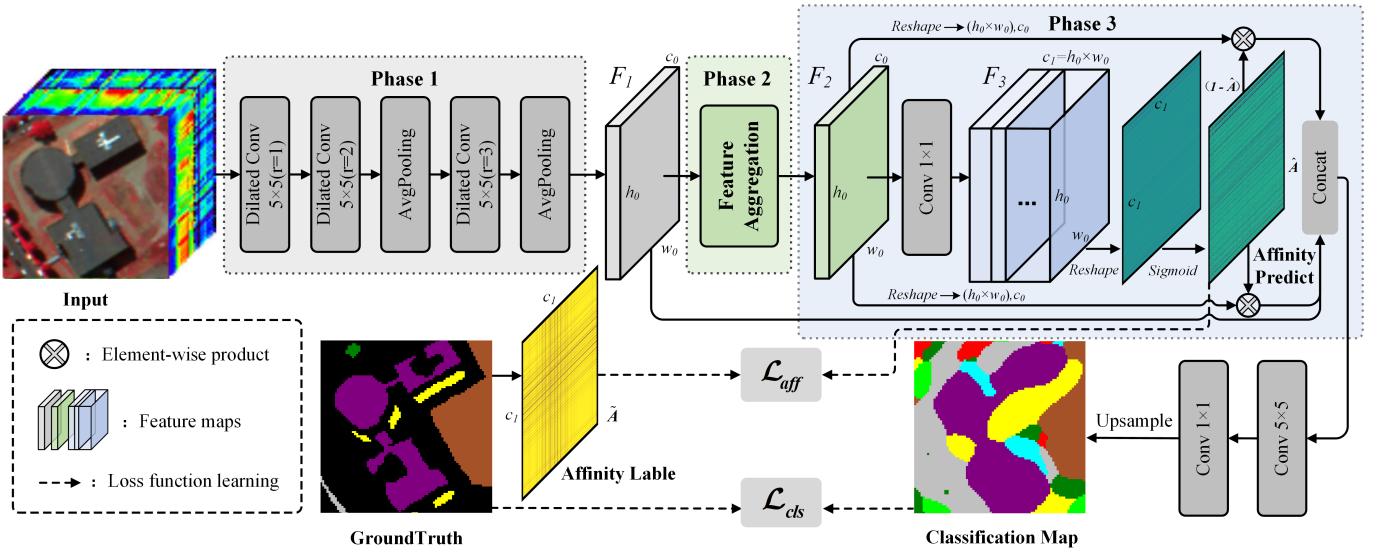


图 4: 提出的类上下文感知网络架构

## 4 复现细节

### 4.1 与已有开源代码对比

有限的卷积核感受野大小限制了像素捕捉长程信息，这不利于提高特征的判别性，降低了模型对对抗样本的防御能力<sup>[29]</sup>。在开源代码<sup>[22]</sup>中提出的自注意力上下文模型通过建立像素与远程位置之间的关系来捕获像素的全局上下文信息，使错误预测标签的损失与这些关系共享，从而提高模型对对抗样本的鲁棒性。然而，这种方式建立的上下文关系没有明确规范化，不可避免地引入了与给定像素类别不同的上下文关系，影响了对抗攻击下模型预测的正确率。为了获得更清晰的全局上下文关系，提高模型对对抗攻击的抵抗能力，我提出一种鲁棒的类上下文感知网络，将类之间的上下文关系建模为先验知识，通过区分类内和类间上下文关系进一步细化全局上下文，从而提高像素之间长程关系的可靠性，产生更强的鲁棒性。在复现过程中，我主要复现了所提类上下文感知模型架构，对于网络中数据输入与评价部分借鉴了以下代码库：1) <https://github.com/YonghaoXu/SACNet>. 2) <https://github.com/ycszen/ContextPrior>.

### 4.2 实验环境搭建

本节我们详细介绍实验细节。实验部分的组织结构如下，首先对用于分析的数据集进行了描述。第二，详细说明了实验的细节设置，包括网络中的参数设置和评价指标。

#### 4.2.1 数据集

使用三个公开的数据集来验证所提出方法的有效性。包括 Indian Pines<sup>[30]</sup> 和 Pavia University<sup>1</sup> 数据集，它们覆盖了农村、城市和农田场景。对应的假彩色图和 GroundTruth 如图 5 所示。实验中使用的训练样本和测试样本数量如图 6 所示。

<sup>1</sup>[http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral\\_Remote\\_Sensing\\_Scenes](http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes)

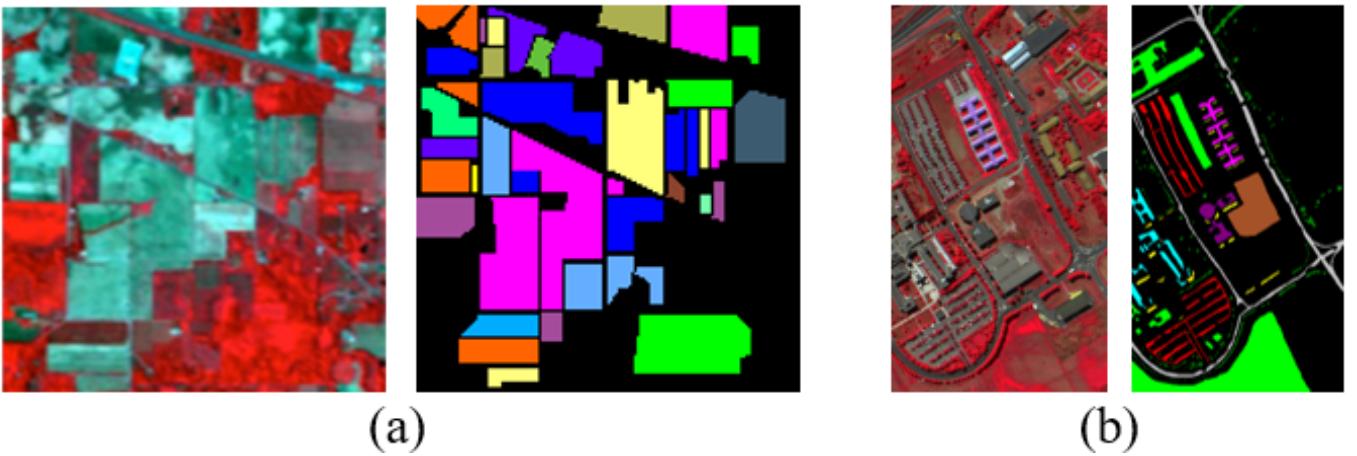


图 5: 假彩色图与对应的 Groundtruth: (a) Indian Pines. (b) Pavia University.

Indian Pines					Pavia University				
No	Name	Training	Test	Toal	No	Name	Training	Test	Toal
1	Alfalfa	20	26	46	1	Asphalt	100	6531	6631
2	Corn-N	80	1348	1428	2	Meadows	100	18549	18649
3	Corn-M	80	750	830	3	Gravel	100	1999	2099
4	Corn	80	157	237	4	Tress	100	2964	3064
5	Grass-M	80	403	483	5	Metal sheets	100	1245	1345
6	Grass-T	80	650	730	6	Bare soil	100	4929	5029
7	Grass-P	15	13	28	7	Bitumen	100	1230	1330
8	Hay-W	80	398	478	8	Bricks	100	3582	3682
9	Oats	10	10	20	9	Shadows	100	847	947
10	Soybeans-N	80	892	972					
11	Soybeans-M	80	2375	2455					
12	Soybeans-C	80	513	593					
13	Wheat	80	125	205					
14	Woods	80	1185	1265					
15	Buildings	50	336	386					
16	Stone	50	43	93					
Total		1025	9224	10249	Total		900	41876	42776

图 6: 不同类别使用的训练集和测试集: (a) Indian Pines. (b) Pavia University.

1) *Indian Pines*: 该图像是在美国印第安纳州西北部的松林地区拍摄的。其场景空间大小为  $145 \times 145$  并包括 200 个光谱波段用于分析。该图像中有 10249 个标记像素，共属于 16 个土地覆盖类别。

2) *Pavia University*: 该图像拍摄于意大利帕维亚大学。其场景空间大小为  $610 \times 340$  包括 103 个光谱波段可供分析。在这幅图像中，总共标记了 42776 个像素，共属于 9 个土地覆盖类别。

#### 4.2.2 实验细节与设置

1) 实验设置: 在接下来的实验中，我们使用 FSGM 攻击来模拟对抗样本的生成，以验证所提出的方法在对抗攻击下的鲁棒性其中  $\varepsilon$  设置为 0.06。所有实验都在一台配有 Intel(R) Core(TM) i9-10900KF

3.70GHz CPU、Nvidia GTX 2080 Ti GPU 和 32GB RAM 的计算机上进行。

2) 参数优化: 所提的类上下文感知网络是基于 PyTorch 框架实现的。使用 Adam 优化器优化训练过程, 学习率设置为  $5e-4$ , 衰减设置为  $5e-5$ , 训练周期设置为 500。

3) 评价指标: 在实验中, 使用总体精度 (OA)、平均精度 (AA) 和 kappa 系数来评价不同分类方法的性能。为了减少实验误差, 所有实验重复 10 次, 并取均值和方差作为最终结果。

### 4.3 创新点

论文的主要贡献可以总结如下:

1) 引入类上下文感知模块, 明确区分类内和类间关系。该模型通过嵌入可学习的亲和度损失来捕捉类内和类间上下文之间的关系, 以改善特征的类内相似性和类间差异性。

2) 针对高光谱图像分类任务中的对抗攻击, 提出了一种鲁棒的类上下文感知网络。该方法通过区分类内和类间上下文关系来进一步细化全局上下文信息, 从而提高模型的鲁棒性。

3) 在公开的高光谱图像数据集中实验结果表明, 所提方法在分类性能上优于现有的对抗防御方法。

## 5 实验结果分析

本节我们对实验结果进行分析, 主要包括三个部分: 第一部分是消融实验, 分析不同模块对所提方法的贡献; 第二部分是采用干净测试样本和对抗测试样本, 将所提方法与几种先进方法进行对比实验; 第三部分探讨了不同攻击扰动的影响。

### 5.1 消融实验

在本小节中, 我们评估了所提方法不同模块对精度的贡献, 包括特征聚合 (FA) 和亲和学习 (AL)。在两个数据集上的实验结果如图 7 所示, 其中 Backbone 表示所提出框架的第一阶段, 这是一个由空洞卷积组成的 FCN。

Method	Feature aggregation (FA)	Affinity learning (AL)	Indian Pines		Pavia University	
			OA	Kappa	OA	Kappa
Backbone	-	-	87.28	85.53	78.70	72.06
+FA	✓	-	94.44(↑7.16)	93.63(↑8.10)	81.74(↑3.04)	76.31(↑4.25)
+AL	-	✓	92.36(↑5.08)	91.26(↑5.73)	80.71(↑2.10)	75.55(↑3.49)
+FA&AL	✓	✓	<b>95.63(↑8.35)</b>	<b>93.75(↑8.22)</b>	<b>89.44(↑10.74)</b>	<b>86.03(↑13.97)</b>

图 7: 所提出的类上下文感知架构中不同模块的性能贡献, 包括特征聚合模块和亲和学习模块。括号中的值表示与主干之间的精度差值, 粗体中的数字表示最佳值。所有结果都以% 表示。

从表中可以看出, 单独应用 FA 和 AL 在 OA 和 Kappa 上都取得了比 Backbone 更高的精度, 表明空谱信息的聚合以及对类内和类间上下文关系的捕获, 可以有效提高模型在对抗攻击下的性能。当单独使用这两个模块时, FA 比 AL 具有更高的精度。以印度松树为例, 与 Backbone 相比, FA 对 OA 和 Kappa 分别有 2.08% 和 2.37% 的提高。这一趋势在帕维亚大学图像中也很明显, 这表明在亲和学习之前, 需要特征聚合来获得有价值的上下文信息。此外, 同时应用 FA 和 AL 得到的实验结果进一步验证了我们的判断。与 Backbone 相比, 特征聚合后的上下文亲和学习可以大大提高模型在对抗样本下的分类性能。实验结果表明, 在特征聚合后, 通过学习类上下文信息的先验知识来捕获类内和类间的

上下文关系，可以有效地抑制不可靠的上下文信息的影响，从而提高模型在对抗样本中的分类性能。

## 5.2 对比实验

在本小节中，我们将所提出的方法与其他几种先进的方法进行比较，包括 3DCNN<sup>[31]</sup>，空谱残差网络 (SSRN)<sup>[32]</sup>，空谱全卷机网络 (SSFCN)<sup>[33]</sup>，和自监督上下文网络 (SACNet)<sup>[22]</sup>。在这些方法中，具体的网络结构和参数设置可以在相应的参考文献中找到。测试样本包括干净样本和对抗样本。

1) *Indian Pines*: Indian Pines 的可视化结果和定量实验结果见图 8 和图 9. 大多数在干净样本上表现良好的分类方法在对抗样本上的准确率大大降低。在干净数据集中表现最好的 SSRN 在对抗样本中 OA 降低了 61.75%，AA 降低了 73.14%。一些类别的准确率接近于 0，如 Alfalfa, Hay-W 和 Moth C，这表明大多数先进的基于深度学习的 HSI 分类方法非常容易受到对抗性攻击。SACNet 和提出的类上下文感知网络在对抗样本上表现最好，这一结果表明，通过关注像素的全局上下文信息，可以有效地抵抗对抗样本攻击。与 SACNet 相比，该方法在未攻击和被攻击状态下的 OA 分别提高了 5.26% 和 6.7%，表明该方法通过进一步区分类内和类间上下文信息来细化全局上下文，能够提高模型对对抗样本的泛化能力和鲁棒性。从图 8 中所示的分类图，可以看出，在 Hay-W 类和 Oats 类中，所提方法在干净样本和对抗样本上的性能都与 GroundTruth 一致，进一步体现了所提方法在防御对抗攻击方面的优势。

2) *Pavia University*: 与 Indian Pines 数据集不同，该数据集的地物多为城市类型，分布相对密集，增加了分类难度。不同方法在 Pavia University 的实验结果显示在图 11. 3DCNN、SSRN 和 SSFCN 在干净样本和对抗样本上的准确率存在显著差异，说明较好的分类方法容易受到对抗样本的欺骗。SACNet 和本文提出的类上下文感知网络在对抗样本上获得了令人满意的精度，表明通过捕获像素的长范围依赖关系来关注高光谱图形的全局上下文信息有利于缓解对抗样本的欺骗。如图 10，无论是干净样本还是对抗样本，与其他方法相比，提出的类上下文感知网络错误分类的类别要少得多，因此获得了更接近真实分布的性能。这一发现表明，该方法通过关注类内上下文和类间上下文，可以细化全局上下文信息，提高高光谱图形在对抗攻击下防御能力。

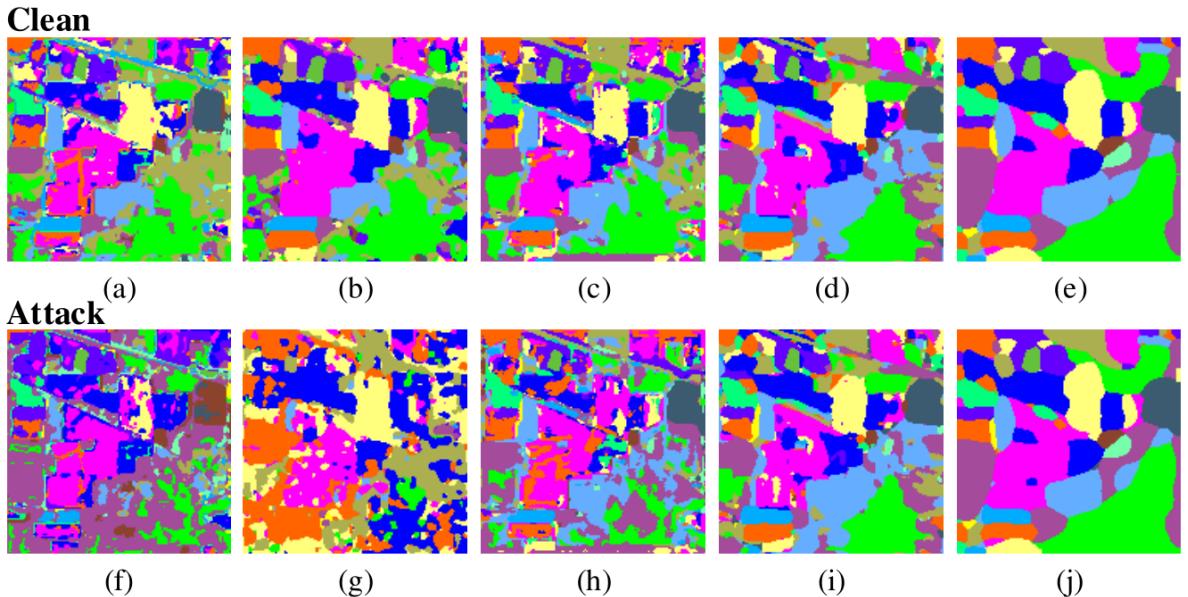


图 8: 干净样本 (第一列) 和对抗样本 (第二列) 在 Indian Pines 获得的分类图。

Class	Test on clean samples					Test on adversarial samples				
	3DCNN	SSRN	SSFCN	SACNet	RCCA	3DCNN	SSRN	SSFCN	SACNet	RCCA
Alfalfa	88.08	99.62	90.00	96.15	94.23	98.46	3.08	92.31	95.38	87.69
Corn-N	68.06	94.31	82.79	89.29	96.60	60.05	68.98	60.82	89.22	97.43
Corn-M	69.20	97.00	85.44	82.12	96.33	27.56	83.77	73.44	79.97	95.33
Corn	91.08	99.62	97.20	98.85	99.68	71.40	2.80	94.27	98.34	98.09
Grass-M	94.42	98.61	95.78	97.07	98.91	48.88	10.27	89.08	96.30	98.56
Grass-T	98.25	99.63	94.65	98.14	99.52	51.63	18.95	88.18	96.54	97.72
Grass-P	100.00	100.00	98.46	100.00	100.00	94.62	3.08	98.46	98.46	98.46
Hay-W	97.86	100.00	98.19	97.06	100.00	32.16	1.81	97.59	97.06	100.00
Oats	87.00	100.00	84.00	100.00	100.00	77.00	15.00	86.00	100.00	100.00
Soybeans-N	78.81	95.09	85.49	90.67	96.82	30.46	68.49	32.67	86.39	96.08
Soybeans-M	65.55	93.88	74.15	86.52	92.40	78.76	15.42	38.27	84.35	91.01
Soybeans-C	81.62	97.90	86.16	86.86	92.26	64.37	1.13	68.23	86.08	92.44
Wheat	99.36	99.84	99.04	98.56	99.20	71.92	1.92	98.40	87.92	99.20
Woods	94.87	99.16	95.90	97.32	99.05	54.89	20.68	67.70	93.55	98.60
Buildings	85.03	98.60	88.15	97.47	99.55	35.06	68.36	69.88	95.33	99.58
Stone	98.60	100.00	98.60	98.84	100.00	97.44	19.30	96.28	96.74	99.07
OA	79.01 (±0.99)	<b>96.50</b> (±0.45)	85.76 (±0.91)	91.00 (±0.84)	96.26 (±0.43)	56.45 (±4.94)	34.75 (±9.27)	61.11 (±4.81)	88.93 (±2.25)	<b>95.63</b> (±3.53)
AA	87.36 (±0.81)	<b>98.33</b> (±0.27)	90.88 (±1.04)	94.68 (±0.66)	97.79 (±0.26)	62.17 (±7.06)	25.19 (±5.43)	78.22 (±2.21)	92.60 (±1.89)	<b>96.83</b> (±1.05)
kappa	76.14 (±1.11)	<b>95.97</b> (±0.52)	83.75 (±1.03)	89.66 (±0.96)	95.70 (±0.37)	49.63 (±5.71)	26.96 (±9.52)	56.39 (±4.87)	87.29 (±2.57)	<b>94.97</b> (±3.97)

图 9: 在 Indian pines 上不同方法对干净样本和对抗样本得到的分类精度。括号内的数字表示对应的标准差, 加粗的数字表示最佳值。

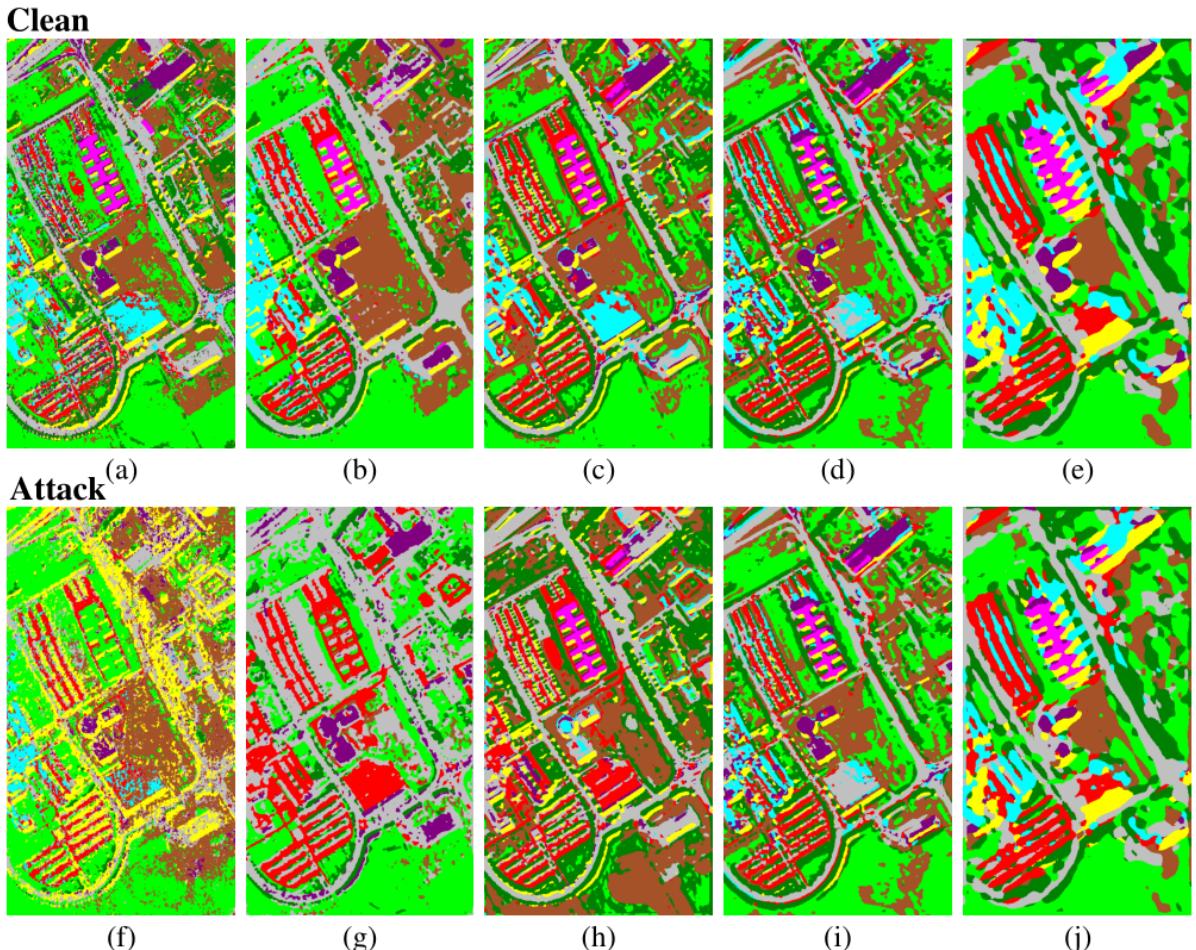


图 10: 干净样本 (第一列) 和对抗样本 (第二列) 在 Pavia University 获得的分类图。

Class	Test on clean samples					Test on adversarial samples				
	3DCNN	SSRN	SSFCN	SACNet	RCCA	3DCNN	SSRN	SSFCN	SACNet	RCCA
Asphalt	86.28	98.41	89.51	83.85	84.03	3.40	47.60	95.51	84.03	85.12
Meadows	88.51	98.76	95.19	87.98	92.12	71.56	98.70	13.66	86.81	92.60
Gravel	72.85	93.65	92.97	87.40	97.45	17.31	19.03	4.13	85.20	85.44
Tress	89.02	98.69	97.56	92.30	74.97	80.18	36.92	98.47	92.05	73.68
Metal sheets	99.28	100.00	99.86	98.60	98.55	40.05	6.66	99.71	98.78	95.34
Bare soil	76.81	98.88	91.39	87.36	97.57	15.27	16.96	49.09	84.90	95.01
Bitumen	84.31	99.11	95.56	94.13	96.99	42.52	40.34	15.98	79.43	64.47
Bricks	74.81	96.40	91.43	91.32	94.39	29.39	33.12	73.67	91.73	94.64
Shadows	98.54	99.95	99.83	98.11	92.44	0.69	0.02	99.76	97.52	91.26
OA	85.30 (±3.52)	<b>98.34</b> (±0.36)	93.84 (±0.29)	88.53 (±3.17)	91.08 (±1.12)	45.49 (±5.22)	60.88 (±9.61)	45.65 (2.56)	87.22 (±3.10)	<b>89.44</b> (±1.63)
AA	85.60 (±5.35)	<b>98.20</b> (±0.27)	94.81 (±0.32)	91.23 (±1.43)	92.06 (±0.70)	33.37 (±7.53)	33.26 (±12.19)	61.11 (±3.82)	<b>88.94</b> (±3.98)	86.40 (±3.23)
Kappa	80.61 (±4.73)	<b>97.79</b> (±0.47)	91.83 (±0.39)	85.01 (±3.94)	88.25 (±1.43)	31.62 (±6.32)	38.35 (±19.9)	37.63 (±2.15)	83.30 (±3.91)	<b>86.03</b> (±2.12)

图 11: 在 Pavia University 上不同方法对干净样本和对抗样本得到的分类精度。括号内的数字表示对应的标准差, 加粗的数字表示最佳值。

### 5.3 攻击强度分析

在本节中, 我们将详细评估具有不同强度的攻击扰动对分类精度的影响, 通过设置  $\varepsilon$  范围在 [0.01-0.1], 步长为 0.01。实验在两个 HSI 数据集上比较了五种最先进的分类方法。

如图 12 所示, 随着攻击扰动的增加, 不同方法的分类准确率均呈下降趋势, 说明扰动越大, 分类模型对对抗样本的防御能力越弱。模型的分类性能能通过修改扰动的强度来控制, 这进一步反映了深度学习模型在 HSI 分类中的脆弱性。我们还发现, 在具有更平滑地物分布的数据集如 Indian Pines 中, 分类模型在对抗样本下表现较差。以最先进的分类方法 SSRN 为例, 当攻击扰动  $\varepsilon$  为 0.01 时, Indian Pines 的 OA 值下降到 50% 以下, 而 Pavia University 的准确率可以达到 75% 左右, 这表明地物分布更复杂的数据集有利于减轻对抗性攻击造成的危害。所提的类上下文感知网络和 SACNet 的精度并没有随着攻击扰动程度的增加而发生显著变化, 这表明通过关注全局上下文可以有效抵抗高光谱图像分类中的对抗性攻击威胁。当攻击扰动  $\varepsilon$  达到 0.1 时, 所提的类上下文感知网络仍然在两个数据集上获得了最好的性能, 进一步证明了所提方法对对抗样本的鲁棒性。

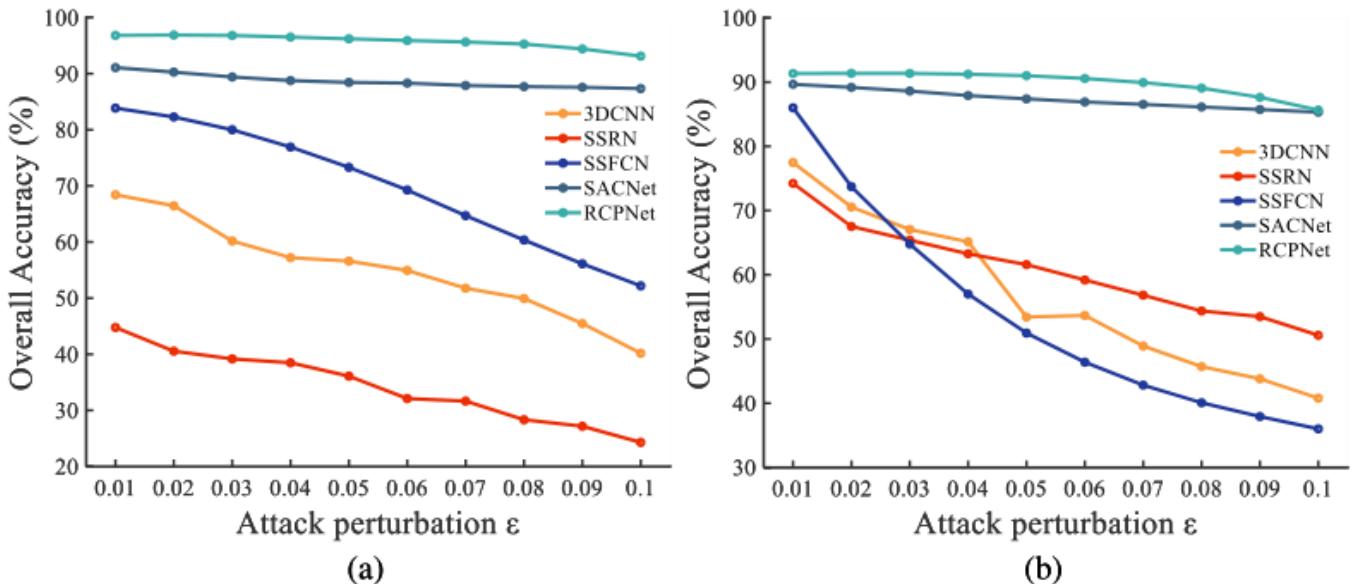


图 12: 不同攻击扰对分类性能的影响: (a) Indian pines, (b) Pavia University.

## 6 总结与展望

针对基于深度学习的高光谱图像分类模型易受对抗样本攻击的问题，本报告提出一种鲁棒的类上下文感知框架，通过亲和度损失监督区分类内上下文和类间上下文信息，以获取更精细的全局上下文。这些上下文关系有助于在整个图像中像素之间建立更可靠的全局依赖关系，从而提高模型对对抗样本的鲁棒性。在两个公开数据集上的实验表明，所提方法在对对抗样本的准确性和鲁棒性之间取得了很好的平衡。

在未来的研究中，鲁棒分类模型在对抗攻击中的准确率和效率仍有提升的空间。现有的应对对抗攻击的鲁棒模型普遍在干净样本上的表现不佳，因此设计一个具有高鲁棒性和高泛化性的统一高光谱图像分类框架至关重要。

## 参考文献

- [1] SHIMONI M, HAELTERMAN R, PERNEEL C. Hyperpectral Imaging for Military and Security Applications: Combining Myriad Processing and Sensing Techniques[J]. IEEE Geosci. Remote Sens. Mag., 2019, 7(2): 101-117. DOI: 10.1109/MGRS.2019.2902525.
- [2] GUAN X, SHEN H, LI X, et al. Climate Control on Net Primary Productivity in the Complicated Mountainous Area: A Case Study of Yunnan, China[J]. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., 2018, 11(12): 4637-4648. DOI: 10.1109/JSTARS.2018.2863957.
- [3] LIU K, SU H, LI X. Estimating High-Resolution Urban Surface Temperature Using a Hyperspectral Thermal Mixing (HTM) Approach[J]. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., 2016, 9(2): 804-815. DOI: 10.1109/JSTARS.2015.2459375.
- [4] MANOLAKIS D, PIEPER M, TRUSLOW E, et al. Longwave Infrared Hyperspectral Imaging: Principles, Progress, and Challenges[J]. IEEE Geosci. Remote Sens. Mag., 2019, 7(2): 72-100. DOI: 10.1109/MGRS.2018.2889610.
- [5] DONG W, ZHOU C, WU F, et al. Model-Guided Deep Hyperspectral Image Super-Resolution[J]. IEEE Trans. Image Process., 2021, 30: 5754-5768. DOI: 10.1109/TIP.2021.3078058.
- [6] LU X, ZHANG J, YANG D, et al. Cascaded Convolutional Neural Network-Based Hyperspectral Image Resolution Enhancement via an Auxiliary Panchromatic Image[J]. IEEE Trans. Image Process., 2021, 30: 6815-6828. DOI: 10.1109/TIP.2021.3098246.
- [7] LI S, SONG W, FANG L, et al. Deep Learning for Hyperspectral Image Classification: An Overview [J]. IEEE Trans. Geosci. Remote Sens., 2019, 57(9): 6690-6709. DOI: 10.1109/TGRS.2019.2907932.
- [8] ROY S K, KRISHNA G, DUBEY S R, et al. HybridSN: Exploring 3-D-2-D CNN Feature Hierarchy for Hyperspectral Image Classification[J]. IEEE Geosci. Remote Sensing Lett., 2020, 17(2): 277-281. DOI: 10.1109/LGRS.2019.2918719.
- [9] XIE J, HE N, FANG L, et al. Multiscale Densely-Connected Fusion Networks for Hyperspectral Images

- Classification[J]. IEEE Trans. Circuits Syst. Video Technol., 2021, 31(1): 246-259. DOI: 10.1109/TCSVT.2020.2975566.
- [10] TU B, HE W, HE W, et al. Hyperspectral Classification via Global-Local Hierarchical Weighting Fusion Network[J]. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., 2021, 15: 184-200. DOI: 10.1109/JSTARS.2021.3133009.
- [11] YANG J, WU C, DU B, et al. Enhanced Multiscale Feature Fusion Network for HSI Classification[J]. IEEE Trans. Geosci. Remote Sens., 2021, 59(12): 10328-10347. DOI: 10.1109/TGRS.2020.3046757.
- [12] WANG D, DU B, ZHANG L, et al. Adaptive Spectral-Spatial Multiscale Contextual Feature Extraction for Hyperspectral Image Classification[J]. IEEE Trans. Geosci. Remote Sens., 2021, 59(3): 2461-2477. DOI: 10.1109/TGRS.2020.2999957.
- [13] HONG D, GAO L, YAO J, et al. Graph Convolutional Networks for Hyperspectral Image Classification [J]. IEEE Trans. Geosci. Remote Sens., 2021, 59(7): 5966-5978. DOI: 10.1109/TGRS.2020.3015157.
- [14] HE X, CHEN Y. Optimized Input for CNN-Based Hyperspectral Image Classification Using Spatial Transformer Network[J]. IEEE Geosci. Remote Sensing Lett., 2019, 16(12): 1884-1888. DOI: 10.1109/LGRS.2019.2911322.
- [15] YANG X, CAO W, LU Y, et al. Hyperspectral Image Transformer Classification Networks[J]. IEEE Trans. Geosci. Remote Sens., 2022, 60: 1-15. DOI: 10.1109/TGRS.2022.3171551.
- [16] SONG R, FENG Y, CHENG W, et al. BS2T: Bottleneck Spatial-Spectral Transformer for Hyperspectral Image Classification[J]. IEEE Trans. Geosci. Remote Sens., 2022: 1-1. DOI: 10.1109/TGRS.2022.3185640.
- [17] CZAJA W, FENDLEY N, PEKALA M, et al. Adversarial examples in remote sensing[C] // Proc. of the 26th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst. 2018: 408-411.
- [18] CHENG G, SUN X, LI K, et al. Perturbation-Seeking Generative Adversarial Networks: A Defense Framework for Remote Sensing Image Scene Classification[J]. IEEE Trans. Geosci. Remote Sens., 2021, 60: 1-11. DOI: 10.1109/TGRS.2021.3081421.
- [19] LI H, HUANG H, CHEN L, et al. Adversarial Examples for CNN-Based SAR Image Classification: An Experience Study[J]. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., 2021, 14: 1333-1347. DOI: 10.1109/JSTARS.2020.3038683.
- [20] DU C, HUO C, ZHANG L, et al. Fast C&W: A Fast Adversarial Attack Algorithm to Fool SAR Target Recognition With Deep Convolutional Neural Networks[J]. IEEE Geosci. Remote Sensing Lett., 2022, 19: 1-5. DOI: 10.1109/LGRS.2021.3058011.
- [21] DU C, ZHANG L. Adversarial Attack for SAR Target Recognition Based on UNet-Generative Adversarial Network[J]. Remote Sens., 2021, 13(21): 4358.

- [22] XU Y, DU B, ZHANG L. Self-Attention Context Network: Addressing the Threat of Adversarial Attacks for Hyperspectral Image Classification[J]. *IEEE Trans. Image Process.*, 2021, 30: 8671-8685. DOI: 10.1109/TIP.2021.3118977.
- [23] PARK S, LEE H J, RO Y M. Adversarially Robust Hyperspectral Image Classification via Random Spectral Sampling and Spectral Shape Encoding[J]. *IEEE Access*, 2021, 9: 66791-66804. DOI: 10.1109/ACCESS.2021.3076225.
- [24] SHI C, DANG Y, FANG L, et al. Hyperspectral Image Classification With Adversarial Attack[J]. *IEEE Geosci. Remote Sensing Lett.*, 2021, 19: 1-5. DOI: 10.1109/LGRS.2021.3122170.
- [25] AKHTAR N, MIAN A, KARDAN N, et al. Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey[J]. *IEEE Access*, 2021, 9: 155161-155196. DOI: 10.1109/ACCESS.2021.3127960.
- [26] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.
- [27] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv:1706.06083, 2017.
- [28] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symp. Secur. Privacy (SP). 2017: 39-57.
- [29] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.
- [30] BAUMGARDNER M F, BIEHL L L, LANDGREBE D A. 220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3[J]. *Purdue University Research Repository*, 2015, 10(7): 991.
- [31] LI Y, ZHANG H, SHEN Q. Spectral-Spatial Classification of Hyperspectral Imagery with 3D Convolutional Neural Network[J]. *Remote Sens.*, 2017, 9(1): 67.
- [32] ZHONG Z, LI J, LUO Z, et al. Spectral-Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework[J]. *IEEE Trans. Geosci. Remote Sens.*, 2018, 56(2): 847-858. DOI: 10.1109/TGRS.2017.2755542.
- [33] XU Y, DU B, ZHANG L. Beyond the Patchwise Classification: Spectral-Spatial Fully Convolutional Networks for Hyperspectral Image Classification[J]. *IEEE Trans. Big Data.*, 2020, 6(3): 492-506. DOI: 10.1109/TBDA.2019.2923243.