DCN-T: Dual Context Network With Transformer for Hyperspectral Image Classification

摘要

由于复杂成像条件引起的空间变异性，高光谱图像 (HSI) 分类具有挑战性。先前的方法存在表征能力有限的问题，因为它们是在有限的注释数据上训练专门设计的搜索网络。我们提出了一个三光谱图像生成通道，将 HSI 转换为高质量的三光谱图像，，从而能够使用现成的 ImageNet 预训练骨干网络进行特征提取。由于观察到 HSI 中存在许多具有不同语义和几何属性的同质区域，可以用来提取有用的上下文，我们提出了一个端到端分割网络，命名为 DCN- T。它采用 transformer 对基于相似性聚类发现的同质区域内部和区域之间的区域适应和全局聚集空间语境进行有效编码。为了充分利用 HSI 的丰富光谱，我们采用了一种集成方法，通过投票方案将三光谱图像的所有分割结果集成到最终预测中。在三个公共基准上进行的大量实验表明，我们提出的方法优于最先进的 HSi 分类方法。代码将在 https://github. com/DotWang/DCN-T 上发布。

关键词：高光谱图像分类；Transformer；上下文网络；

1 引言

与其他形式的数据相比，高光谱图像具有独特的优势，因为它们可以准确的描述物体的物理特征。通过利用先进的传感器，HSI 接收密集而狭窄的波长范围内的电磁信号，具有丰富的数百个波段的光谱信息。因此，HSI 在各种识别任务中得到了广泛的应用，如精准农业 [26] 和环境监测。这些任务通常通过实现逐像素分类方法来实现。

HSI 面临着一个被称为“空间变异性”的重大挑战，即同一物体在不同地区表现出不同的特征。这归因于诸如大气干扰 [8]、光等因素，以及在成像过程中发生的其他此类影响。由此产生的混合类别问题可能导致类间相似性，即不同类别呈现出相似的光谱特征，从而增加了分类的难度。

为了解决上述问题，有必要提取判别特征。随着卷积神经网络 (cnn) 成为基于深度学习的 HSI 分类模型中应用最广泛的网络，研究人员设计了各种网络结构，包括总体框架构建 [14] [11] 和有效模块利用。一些方法也使用了自动网络架构搜索算法! [2]。此外，循环神经网络 [13] 和图卷积神经网络 [18] 也进行了探索。通常，这些网络是在 HSI 上从零开始训练的。

大量的研究 [3] 已经证明了使用预训练权值来增强深度网络性能和收敛性的有效性。有许多现成的网络在大规模数据集上进行预训练，例如，ImageNet [4] 包含来自现实世界的数百万张图像。它们在提取判别特征方面已经获得了出色的能力。然而，由于自然图像与 HSI 之

间通道数的差异，这些网络不能直接适用于 HSI 分类。一种幼稚的解决方案是将所有光谱通道转换为三光谱图像或仅选择三个通道，然而与原始丰富的光谱相比，这将导致显着的信息损失。此外，通过随机采样和以集合方式评估来生成大量的三光谱图像在计算上是不可行的。具体来说，对于 L 通道的 HSI 可以产生的三光谱图像的最大数量为 P= L(L - 1)(L - 2)，其中 L 通常大于 100。

本文认为，通过选择特定的通道进行降维，同时保留尽可能多的有价值的信息，可以生成合适的三光谱图像，这种方法可以使用预训练的网络。为此，我们提出了一种新的三光谱图像生成管道，将 HSI 转换为一系列高质量的三光谱图像。具体来说，本文将所有通道分成几组，每组负责生成的三光谱图像中的一个通道。每个三光谱图像都可以由三个随机选择的任意组组合而成。分组操作旨在平衡精度和效率，使我们能够探索适合降维的比例。随机选择保证了特定类别的偏好能够得到满足，提高了生成图像的多样性。此外，不同组的组合考虑了波长顺序，以保持原始信息。最后，这个专门设计的管道生成大量三光谱图像，形成一个三光谱数据集，类似于预训练期间使用的 RGB 图像。

现有文献表明 [20]，从 ImageNet 预训练中得到的通用表示在遥感任务中是有益的。因此，一旦原始 HSI 被转换为三光谱图像，我们就可以使用现成的 ImageNet 预训练主干进行 HSI 分类。我们的重点主要是提取空间上下文，这在以前的研究 [7] 中被证明对 HSI 分类是有用的。传统的基于空间特征的 HSI 分类方法主要依赖于利用目标像素周围斑块中的邻近信息。然而，所获得的信息通常被限制在一个有限的范围内，这与窗口大小密切相关。因此，出现了图像级的方法，将整个图像直接输入到网络中，同时预测所有像素位置的类别 [19]，类似于计算机视觉中的分割任务。尽管如此，卷积是局部操作，在捕获远程上下文方面效率低下。然后，在一些文献 [24] 中中利用非局部自注意 (SA) 机制来解决这个问题。然而，它会带来很高的计算开销。因此，找到一种有效的方法来捕获有用的上下文以改进 HSI 分类是一个需要解决的关键问题。

参考位置的上下文可以用位置集中的像素来表示。对于自然图像分割，一些方法 [25] 使用同一类别的像素来识别位置集，但由于可用的训练样本缺乏足够的注释，这对于 HSI 来说是不可行的。然而，在 HSI 中，参考像素周围存在许多同质区域。具有相似的像素值和专门的语义和几何属性。语义属性意味着内部像素倾向于属于同一类别，而几何属性表明这些区域的边界与物体轮廓相匹配。因此，同质区域可以提供可用于增强分类性能的有价值的空间上下文信息。由于同一类别的同质区域可能没有连接，因此在捕获区域内上下文后，有必要对跨不同区域的上下文信息进行聚合。为了充分利用这些同质区域内的特征，我们利用了一个包含多头 SA (MHSA) 机制的先进 transformer。在技术上，我们将 HSI 转换成三光谱图像，并使用 ImageNet 预训练的主干来获得相应的特征，然后对相邻像素表示进行基于相似性的聚类以获得同质区域。然后，本文开发了一个基于 transformer 的双上下文模块 (DCM) 来提取两种类型的上下文，即 regional adaptation context (RAC) 和 global aggregation context(GAC)，它们分别对应于区域内和区域间的上下文。我们称整个网络为带 transformer 的双上下文网络 (DCN-T)。

如前所述，每个三光谱图像是通过从原始 HSI 中选择特定通道产生的，因此只包含光谱信息的一个子集。由于原始 HSI 由多个通道组成，因此可以将不同三光谱图像的分割结果进行组合。为了实现这一点，本文提出了一种投票方案，使分割图的融合能够提高分类精度。综上所述，本文的主要贡献有三个方面。

1) 设计了一种新的从 HSI 生成高质量三光谱图像的通道，该通道允许使用现成的 ImageNet 预训练主干提取判别特征并解决空间变异性问题。

2) 开发了一个端到端分割网络，命名为 DCN-T。具体来说，引入了一种用于轮廓自适应同质区域生成的聚类方案，并采用 transformer 来捕获分别位于这些区域或区域之间的区域自适应和全局聚集空间上下文。

3) 提出了一种新的投票方案，通过整合所有三光谱图像的分割结果，有效地利用了 HSI 的丰富光谱，从而提高了分类精度。在三个公共 HSI 基准上进行的大量定性和定量实验表明，所提出的方法优于最先进的方法。

2 相关工作

2.1 HSI 分类的降维

除了空间变异性之外，HSI 还有另一个经典问题，由于波段丰富，被称为维度诅咒或休斯现象。这个问题通常是由有限样本和高维特征之间的不平衡造成的，容易导致过拟合。这个问题暗示了 HSI 中的通道往往是多余的。解决这个问题最直观的思路是进行合理的降维。

传统上，降维有两种范式。第一种是特征选择，目的是通过一些人工设计的标准来选择最具辨别力的波段。然而，不同对象的光谱冗余通常发生在不同的信道上，增加了波段选择的难度。此外，在这些方法中，除了选择的波段之外，其他信道不可避免地会被丢弃，尽管它们可能仍然有价值。另一种是特征提取，将原有的高维特征直接投影到低维空间，如主成分分析。尽管如此，光谱信息不可避免地会丢失。

考虑到上述挑战，大多数基于深度学习的像素级分类方法在提取光谱特征 [10] 时通常使用所有波段，例如，以参考像素为中心，将大小为 $s \times s \times L$ 的 patch 送入网络 [1]。其中 s 为 patch 的高度和宽度，L 为 HSI 波段的个数。也有一些降低通道维数的方法。但是，他们仍然是通过向网络提供大小为 $s \times s \times p$ 的补丁来进行补丁级分类，其中 p 为降维后的通道数。在最近的图像级分类或基于分割的方法中，将整幅图像馈送到网络 [23]，其中输入数据包含所有通道，导致巨大的计算成本。与上述方法相比，本文提出将原始 HSI 转换为一系列三光谱图像，在降维和保持光谱信息完整性之间实现了更好的权衡。

2.2 基于 CNN 的 HSI 分类方法

由于卷积具有出色的局部感知能力和计算效率，在 HSI 分类领域出现了许多基于 CNN 的方法。最初，将像素向量或补丁馈送到网络中。[10]。例如，[10] 构建了一个五层 1D CNN 来提取光谱特征。[1] 分别采用 2D 和 3D 卷积核来获取空间或光谱-空间特征。参考文献 [28] 首先利用 CNN 沿信道方向捕获光谱特征，然后对空间信息进行聚合。而 [12] 则使用 CNN 从设计的像素对中提取判别表示。除了上述像素级或补丁级分类方法外，最近还提出了许多图像级 (即无补丁) 方法 [19] [23]。在这些网络中，全卷积网络 (fully convolutional networks, FCN) 作为 CNN 的一个特殊家族，已经得到了广泛的应用，其输出与输入的大小相同。[23] 在 FCN 的基础上引入展开卷积来扩大接收域，[27] 采用 encoder-decoder 结构来逐步恢复细节。然而，由于卷积是局部操作，CNN 不擅长提取远程上下文信息以获得全局感知，而全局感知对于 HSI 分类很重要。

2.3 基于注意力的 HSI 分类方法

注意机制旨在模仿人类视觉，对场景的不同区域施加不同的重要性，经常用于 HSI 分类以增强特征。这种机制可以与其他组件灵活结合。一些方法直接采用 squeeze-and-excitation 或卷积块注意模块 [22] 等经典注意模块，对不同通道或空间位置进行加权。一些方法定义了一组可训练参数，通过加权加法将来自不同分支的特征进行合并。与上述显式特征增强方式不同，SA 机制用于隐式增强特征，因为它可以有效地捕获不同位置之间的关系。例如，[15] 直接使用非局部 SA 模块 [21] 获取空间上下文，而在图像级网络中使用了一种更轻的 SA 模块，称为纵横关注]。此外，利用图注意力对 HSI 的不规则空间结构进行处理。在 SA 机制的基础上，提出了一种更强大的变异，MHSA。MHSA 同时在多个子空间中进行不同的 SA，从而可以捕获更丰富的上下文，从而获得更有效的特征表示。MHSA 是 transformer 的核心部件，下面将介绍。

2.4 基于 transformer 的 HSI 分类

transformer 首先被提出用于机器翻译，随后在 NLP 领域取得了巨大成功 [5]。虽然 SA 机制已经应用于 CV 领域 [21]，但直到 2020 年底，transformer 才成功应用于图像理解 [6]。transformer 在 HSI 分类界引起了越来越多的关注。其中 BERT [5] 直接被采用在平坦的输入 patch 上将每个像素类比为 NLP 中的 “词向量”。然后一些方法 [9] 采用 transformer 提取光谱上下文，其中特征嵌入是通过取输入 patch 的一个或多个通道来生成的。不同之处在于一些方法在 CNN 提取的特征上使用了 transformer，而另一些方法 [9] 采用了多个 transformer 编码器堆叠的级联方案。一些文献 [16] 使用 transformer 通过二维或三维 CNN 提取的扁平特征图获得空间上下文。与这些补丁级分类方法不同，本文我们提出了一种新的 DCN-T，它是一种图像级分割网络。最近的一项工作在直接分割整幅图像的特征后，采用多个 transformer 编码器分别提取空间和光谱上下文。与之不同的是，本文不仅使用 transformer 编码器捕获同质区域内和区域之间的上下文文本以获得判别性的空间特征，而且还使用 transformer 解码器通过聚合这些区域的信息来进一步细化特征。一些方法分别通过可训练的线性投影层将每个单波段输入转换为三个通道，以利用 ImageNet 预训练网络。然而，生成的三个通道是高度相关的，也就是说，结果可以看作是三个单波段图像的简单叠加，影响了提取特征的性能。在本文中，通过精心设计的管道将 HSI 转换为一系列高质量的三光谱图像，以匹配 ImageNet 预训练网络的输入大小。然后，在 backbone 提取三光谱图像特征的基础上，利用 transformer 进一步提取同质区域内和同质区域之间的区域和全局上下文，大大提高了特征的表示能力和 HSI 分类性能。

3 本文方法

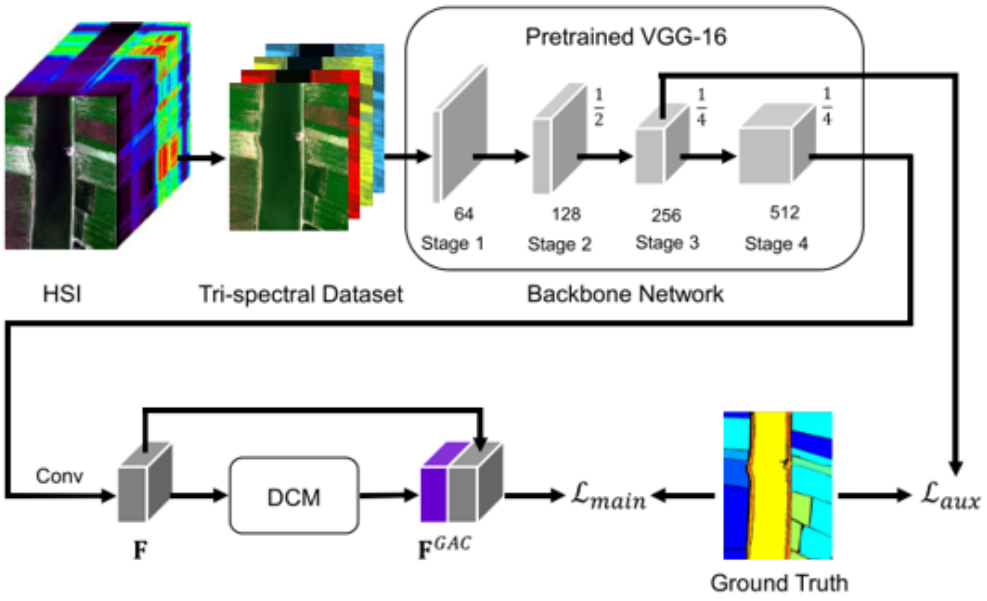

图 1. 模型结构总图

图 1 为 DCN-T 的模型结构总图，主要由三部分组成，分别为：(1) 三光谱图像生成，(2) 双上下文模块，(3) 投票预测，生成的三光谱图像集被送到改进过的 VGG-16 骨干网中，该网络采用 ImageNet 预训练参数，并额外增加了一个卷积层，为后续的双上下文模块产生高表征特征 F，双上下文模块的输出 F^{GAC} 与 F 相连用于最终的预测，在主干网的阶段 3 特征上定义了一个辅助损耗。

3.1 三光谱图像生成

三光谱图像生成包括三个步骤: 分组聚合、随机波段选择、图像拉伸。

1) 分组和聚合: 我们首先将每个 HSI 平均分割成几个子图像。例如, 对于一个 HSI$X_{HSI} = \{x_1, ...x_L\} \in R^{H\times W\times L}$, 其中 H、W、L 分别为高度、宽度和通道数, 子图像的每一组都有 $L \div G$ 通道,G 为组数，由于相邻的通道是高度相关的 [13]，为了便于后续的聚合，我们按照光谱方向顺序拆分通道，而不是随机分组。

然后，我们通过聚合每个组中的通道来进行降维，以解决休斯问题，我们计算这些通道的平均值，每个子图像现在变成了一个两维表，它将被用作一个通道来构建一个三光谱图像。

2）随机波段选择：$\widetilde{X}_{HSI}$ 中任意三个子图像都可以生成三光谱图像。由于不同的类别可能更喜欢不同的波段，即不同的波段能更好地反映，所以本文随机选择三幅子图像作为相应的通道。此外，受 RGB 图像中通道顺序的启发，本文进一步考虑每个波段的波长，以确定生成图像中通道的相对顺序。具体来说，我们采用最大波长对应的子图像作为第一通道，中间的作为第二通道，最小的作为最后通道。这种简单的策略既保持了原有的光谱信息，又进一步排除了不必要的组合，大大减少了生成的三光谱图像的数量。这样，得到的三光谱图像集合可以表示为:

$$X_{TRI}^{0} = RBS(\widetilde{X}_{HSI}) \tag{1}$$

这里，RBS() 表示随机波段选择过程，随机性体现在三光谱图像中的三个通道在 $\widetilde{X}_{HSI}$ 中不必相邻。M 为生成的三光谱图像的个数。

3）图像拉伸：初始三光谱图像是模糊的，本文采用线性 2% 拉伸，这是一种流行且有效的拉伸方法，已被广泛使用，例如在 ENVI 软件中，用于预处理遥感图像以生成高对比度图像。本文之所以采用这种拉伸，是因为虽然三光谱图像是像自然图像一样，但它仍然描绘了遥感场景。线性 2% 拉伸将所有通道的累积直方图中 2% 和 98% 位置之间的值进行缩放到 $[0, 255]$ 而小于 2% 或大于 98% 的值将直接分配给 0 和 255。

3.2 双上下文模块

双上下文模块首先介绍用于上下文捕获的 transform 结构，然后介绍细节，包括同质区域的生成以及 RAC 和 GAC 的提取。

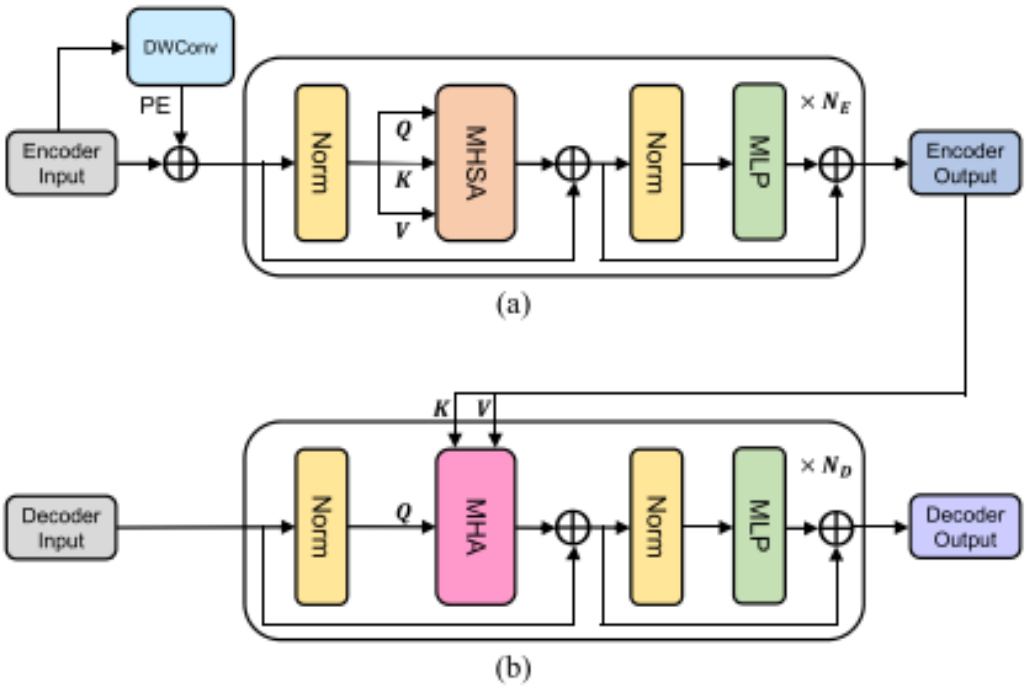

图 2. transformer 结构图

1)transformer：图 2 展示了所用的 transformer 的结构，它包含一个 encoder 和一个 decoder，其层数分别由 N_e 和 N_p 控制，与原始 transformer [17] 相比，此方法删除了 decoder 部分中的 MHSA。(a)encoder: 假设 encoder 的输入为 X_{EI}，考虑到生成的不规则均匀区域得到的 token 长度不等，本方法将不采用预定义的位置编码与嵌入特征的大小进行拟合。使用 3 × 3 深度卷积 (DWConv)，这是一种群卷积，其群数等于通道数来产生位置编码。(b)decoder: 根据图 2 所示，decoder 的流程为:

$$TD\left(X_{DI}, X_{EO}\right) = Res\left(MLP\left(Norm\left(.\right)\right), Res\left(MHA\left(Norm\left(.\right), X_{EO}\right), X_{DI}\right)\right) \tag{2}$$

TD 是 transformer decoder 的缩写,X_{DI} 是另一个需要的输入，$X_{EO} = TE(X_{EI})$ 是 encoder 的输出。

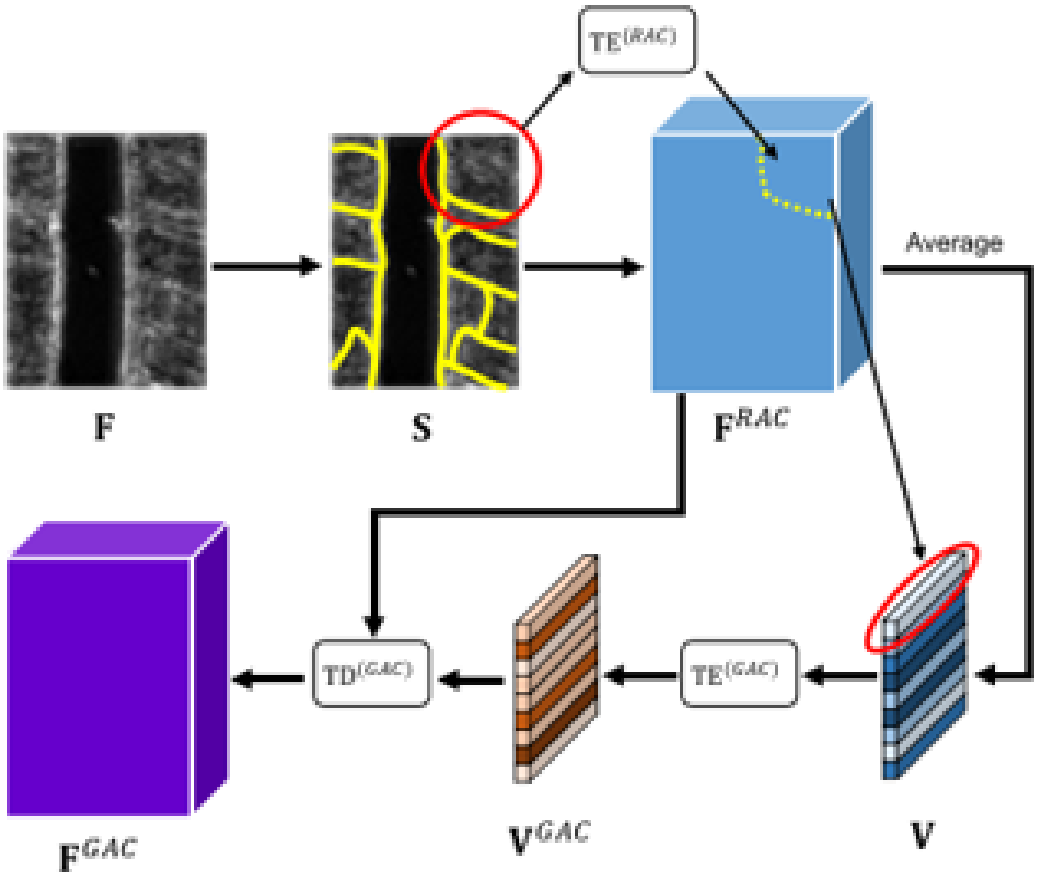

图 3. DCM 细节

2）同质区域生成：假设输入特征为 $F = \{F_1, ..., F_N\} \in R^{C \times H_1 \times W_1}$, 获得同质区域的一个直接思路是对相似的像素表示进行分组，这些像素表示通常在空间上相邻。具体来说，首先将 F 分为 Z 个规则网格，每个网格的大小为 $s \times s$, 通过平均每一个网格取初始化聚类中心，然后通过迭代获得同质区域，具体的说，测量每个像素和聚类中心直接的相似性，构建亲和力矩阵，不断的将像素分配给新的聚类，对于每个像素，只需要考虑周围的聚类，避免不必要的计算。

3) 区域自适应上下文：在生成的均匀区域 S 上捕获 RAC，“自适应”是指这些同质区域是自适应获得的，在网络训练过程中动态调整，以连续拟合特征图对象的轮廓。由于 RAC 是每个同质区域 S_i 内部像素之间的关系，因此这些关系可以通过注意力机制直接建模。我们只使用 reansformer encoder 对 RAC 进行编码。

4) 全局聚合上下文：在获得 RAC 后，本方法进一步构建不同区域之间的连接，使每个像素都能感知到全局信息。生成的上下文信息被称为 GAC，其中的“聚合”是指将这些区域的特征聚合在一起。在聚合之前，需要对每个区域的向量进行平面化和嵌入，然后使用 transformer 对这些嵌入进行处理。但是，由于这些同质区域的形状是多种多样的，所以这些区域的像素数是不一样的。用同一组参数对每个区域进行嵌入是很困难的。因此，我们使用一个描述符

来表示每个区域。每个区域的描述符是通过平均内部像素表示来获得的，即：

$$v_i = \frac{1}{n_i} \sum_{j \in S_i} f_j^{RAC} \tag{3}$$

n_i 是 S_i 的像素数，f_j^{RAC} 是 $F_{S_i}^{RAC}$ 的像素表示，$j \in S_i$，然后使用一个 transformer encoder 来捕捉这些描述符之间的关系，在采用一个 transformer decoder 将向量恢复为 2D 特征图。

3.3 投票预测

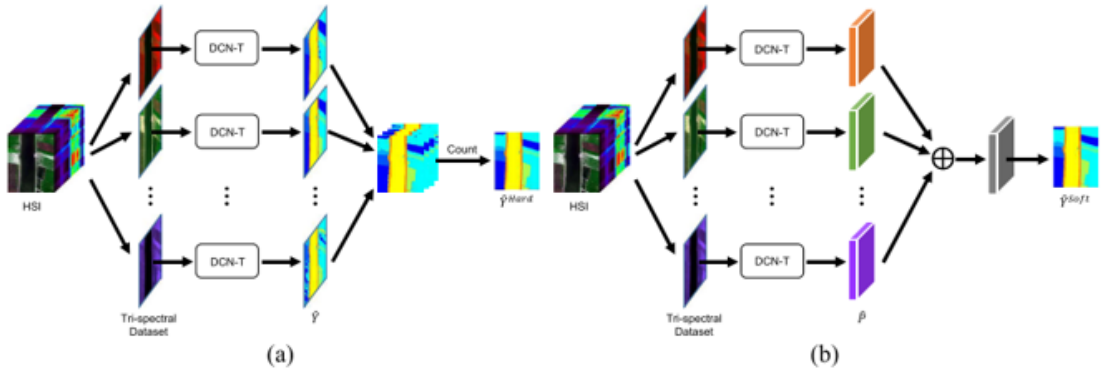

图 4. 投票机制细节

从每一张三光谱图像中可以得到一张分类图。需要综合利用不同的结果来提升准确率。在本文中采用投票机制，其中分别探索了两种方案，即“硬投票”和“软投票”，见图 4。

a) 硬投票：“硬投票”是遵循“少数服从多数”原则的最典型的投票方式假设我们已经从生成的包含 M 幅图像的三光谱图像集中获得了结果图 $\left\{\widehat{Y}^{(1)}, ..., \widehat{Y}^{(M)}\right\}$，然后，第 i 行和第 j 列像素的投票结果基于每个类别的计数：

$$\widehat{Y}_{ij}^{\text{Hard}} = \underset{c \in \{1, \cdots, C_n\}}{\arg\max} \sum_{k=1}^{M} \mathcal{I}\left(\widehat{Y}_{ij}^{(k)} = c\right), \tag{4}$$

其中 I() 是一个二元指标，判断第 k 张图像中对应像素的类别是否为 c。

b) 软投票：“软投票”不使用现成的分类图，而是考虑“硬投票”中被忽略的概率，并通过加法将它们组合起来：

$$\widehat{Y}_{ij}^{Soft} = \underset{c \in \{1, \cdots, C_n\}}{\arg\max} \sum_{k=1}^{M} \widehat{P}_{ij}^{(k)}, \tag{5}$$

$\widehat{P}_{ij}^{k}$ 是一个大小为 $1 \times C_n$ 的向量。

4 复现细节

4.1 与已有开源代码对比

与原开源代码对比，本实验在实验设置上进行了修改，在对于 epoch 的数量以及 batch size 上针对实际实验环境以及数据集均进行相应调整，并对 backbone 进行了修改，在两个预

训练 backbone 上分别进行了实验，并且对于任意数据集分别进行了不同训练样本量的实验对比。

4.2 实验环境搭建

本实验采用单个 16G NVIDIA Tesla P100 GPU 作为主要算力，并使用 python 虚拟环境作为环境支持，在 ResNet 和 VGG16 两个 backbone 以及 Whu-Hi 数据集中的三个子数据集龙口、汉川、洪湖上进行相关实验，backbone 均采用 ImageNet 预训练参数，并且针对每一个数据集都进行样本训练分别为 25、50、100、300 的不同样本训练量实验，以准确率作为模型效果评估的唯一指标，初始学习率为 0.001，采用随机梯度下降法来更新网络参数。各类超参数均采取最优配置。

4.3 创新点

1) 设计了一种新的从 HSI 生成高质量三光谱图像的通道，该通道允许使用现成的 ImageNet 预训练主干提取判别特征并解决空间变异性问题。

2) 开发了一个端到端分割网络，命名为 DCN-T。具体来说，引入了一种用于轮廓自适应同质区域生成的聚类方案，并采用 transformer 来捕获分别位于这些区域或区域之间的区域自适应和全局聚集空间上下文。

3) 提出了一种新的投票方案，通过整合所有三光谱图像的分割结果，有效地利用了 HSI 的丰富光谱，从而提高了分类精度。在三个公共 HSI 基准上进行的大量定性和定量实验表明，所提出的方法优于最先进的方法。

5 实验结果分析

表 1. ResNet 50 为 backbone 实验结果

Samples	25	50	100	300	original(100)
LongKou	91.96	95.45	98.58	99.45	98.91
HongHu	81.03	89.86	95.45	98.55	95.85
HanChuan	85.97	95.14	97.56	99.37	96.38

表 2. VGG16 为 backbone 实验结果

Samples	25	50	100	300	original(100)
LongKou	93.79	96.28	99.09	99.56	98.91
HongHu	66.18	69.32	72.93	74.43	95.85
HanChuan	92.52	96.56	97.76	99.04	96.38

从实验结果可以看出论文复现效果较好，可以打到你预期效果，并且训练样本数量越大其准确率越高，并且提升幅度较大，但是对于不同的 backbone 对于互不相同的数据集并不是总能够表现出良好的效果，，在这三个数据集中，LongKou 数据集表现最佳，并且 backbone 为 VGG16 时其在 HongHu 数据集表现效果极差。

6 总结与展望

本文提出了一种新的 HSI 分类方法，命名为 DCN-T。为了解决丰富光谱带来的高维问题，设计了一个三光谱图像生成管道，将 HSI 转换成一系列具有多种精细光谱信息的高质量三光谱图像。采用 ImageNet 预训练骨干网，提取具有高度代表性的特征，解决 HSI 的空间变异性问题。在此基础上，确定同质区域，然后引入一种基于 transformer 的上下文模块提取区域内以及区域间的上下文，提高特征表示和预测精度。最后，使用硬投票或软投票方案对这些三光谱图像的预测结果进行组合，以进一步提高性能。

从实验结果上来看，方法的实现较为完整，可以达到预期的效果，但是此方法对于在不同 backbone 下对于不同的数据集并不总是能表现出很好的效果，例如，在 ResNet 50 下的 HongHu 数据集上表现的效果远比在 VGG16 上表现得效果要好，说明其模型得鲁棒性欠缺，并不能对于任意 backbone 以及任意数据集都能有不错得效果，下一步可以针对模型的鲁棒性开展相应研究。

参考文献

[1] Yushi Chen, Hanlu Jiang, Chunyang Li, Xiuping Jia, and Pedram Ghamisi. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 54(10):6232–6251, 2016.

[2] Yushi Chen, Kaiqiang Zhu, Lin Zhu, Xin He, Pedram Ghamisi, and Jón Atli Benediktsson. Automatic design of convolutional neural network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):7048–7066, 2019.

[3] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

[7] Jie Feng, Xiande Wu, Ronghua Shang, Chenhong Sui, Jie Li, Licheng Jiao, and Xiangrong Zhang. Attention multibranch convolutional neural network for hyperspectral image classification based on adaptive region search. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):5054–5070, 2021.

[8] Juan Mario Haut, Mercedes E. Paoletti, Javier Plaza, Antonio Plaza, and Jun Li. Visual attention-driven hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10):8065–8080, 2019.

[9] Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. Spectralformer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

[10] Wei Hu, Yangyu Huang, Wei Li, Fan Zhang, and Hengchao Li. Deep convolutional neural networks for hyperspectral image classification. *J. Sensors*, 2015:258619:1–258619:12, 2015.

[11] Hyungtae Lee and Heesung Kwon. Going deeper with contextual cnn for hyperspectral image classification. *IEEE Transactions on Image Processing*, 26(10):4843–4855, 2017.

[12] Wei Li, Guodong Wu, Fan Zhang, and Qian Du. Hyperspectral image classification using deep pixel-pair features. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):844–853, 2017.

[13] Lichao Mou, Pedram Ghamisi, and Xiao Xiang Zhu. Deep recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3639–3655, 2017.

[14] Anirban Santara, Kaustubh Mani, Pranoot Hatwar, Ankit Singh, Ankur Garg, Kirti Padia, and Pabitra Mitra. Bass net: Band-adaptive spectral-spatial feature learning neural network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):5293–5301, 2017.

[15] Hao Sun, Xiangtao Zheng, Xiaoqiang Lu, and Siyuan Wu. Spectral–spatial attention network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5):3232–3245, 2020.

[16] Le Sun, Guangrui Zhao, Yuhui Zheng, and Zebin Wu. Spectral–spatial feature tokenization transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[17] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[18] Sheng Wan, Chen Gong, Ping Zhong, Bo Du, Lefei Zhang, and Jian Yang. Multiscale dynamic graph convolutional network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5):3162–3177, 2020.

[19] Di Wang, Bo Du, and Liangpei Zhang. Fully contextual network for hyperspectral scene parsing. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

[20] Di Wang, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–20, 2023.

[21] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[22] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In-So Kweon. Cbam: Convolutional block attention module. *ArXiv*, abs/1807.06521, 2018.

[23] Yonghao Xu, Bo Du, and Liangpei Zhang. Beyond the patchwise classification: Spectral-spatial fully convolutional networks for hyperspectral image classification. *IEEE Transactions on Big Data*, 6(3):492–506, 2020.

[24] Yonghao Xu, Bo Du, and Liangpei Zhang. Self-attention context network: Addressing the threat of adversarial attacks for hyperspectral image classification. *IEEE Transactions on Image Processing*, 30:8671–8685, 2021.

[25] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*, 2020.

[26] Xia Zhang, Yanli Sun, Kun Shang, Lifu Zhang, and Shudong Wang. Crop classification based on feature band set construction and object-oriented approach using hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(9):4117–4128, 2016.

[27] Zhuo Zheng, Yanfei Zhong, Ailong Ma, and Liangpei Zhang. Fpga: Fast patch-free global learning framework for fully end-to-end hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(8):5612–5626, 2020.

[28] Zilong Zhong, Jonathan Li, Zhiming Luo, and Michael Chapman. Spectral–spatial residual network for hyperspectral image classification: A 3-d deep learning framework. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):847–858, 2018.