arXiv:2304.00495v1 [cs.CV] 2 Apr 2023

基于互联融合的多模态高光谱图像分类

Lu Huo、Jiahao Xia、Leijie Zhang、Haimin Zhang、
Min Xu 工程与信息技术学院，悉尼科技大学

摘要——现有的多模态融合方法，如拼接、求和和基于编码器-解码器的融合，近期已被用于结合高光谱图像（HSI）与激光雷达（LiDAR）的模态特征。然而，这些方法仅从有限的角度考虑高光谱图像-激光雷达信号之间的关系。具体而言，它们忽略了高光谱图像与激光雷达各模态间的上下文信息，以及激光雷达的模态内特征。本文为特征融合提供了新的视角，以全面探索高光谱图像与激光雷达模态之间的关系。本文提出了一种互连融合（IF）框架。首先，提取高光谱图像输入的中心图像块，并将其复制至与高光谱图像输入相同的尺寸。随后，通过计算复制后的中心图像块、高光谱图像输入及对应激光雷达输入之间的自注意力和交叉注意力，生成融合矩阵中的九种不同视角。这样可以充分利用模态内和模态间特征，并以模态内和模态间的方式考虑上下文信息。融合矩阵中的这九个相互关联的元素能够相互补充并消除偏差，从而准确生成用于分类的多模态表征。本文在三个广泛使用的数据集上进行了大量实验：特伦托、MUUFL和Houston。与现有方法相比，IF框架在这些数据集上取得了最先进的结果。

索引词——高光谱图像分类、激光雷达、多模态、多模态、多传感器信息融合、特征融合、遥感、Transformer。

I. 引言

高光谱图像（高光谱图像）中的每个像素都包含海量光谱信息 [1], [2], ，可据此提取精细特征，以识别地球表面的物质 [3]–[4], [5]。然而，由于缺乏高度信息，高光谱图像几乎无法用于识别存在遮挡的物体，例如云层。由于高度不同，利用高光谱图像数据区分相似类别中的同一物体（例如屋顶上的草与地面上的草）是不可能的[6], [7]。为获取高度信息，合成孔径雷达（SAR） [8], [9] 或激光雷达（LiDAR） [10] 与高光谱图像信号协同工作，用于检测存在遮挡或完全隐藏在障碍物后的物体。因此，不同模态的融合在遥感领域日益受到研究关注。

大多数多模态融合方法会对一种模态的特征表示进行编码，然后融合多个模态的特征以进行分类 [11], [12], [13]。传统特征融合方法，如拼接、求和，

本文由 IEEE 出版技术组制作。该团队位于新泽西州皮斯卡塔韦。

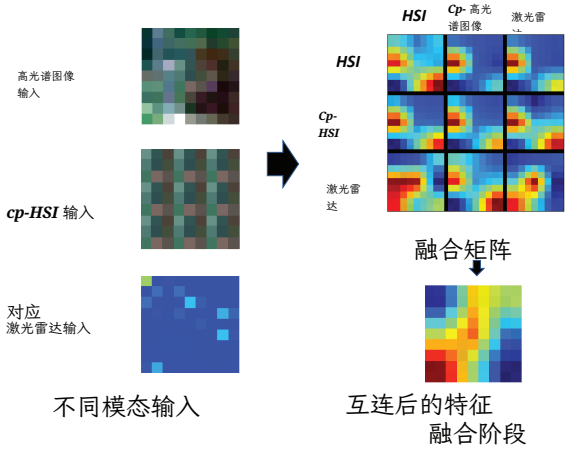


图 1. 草地样本由三种输入组成：高光谱图像、高光谱图像的中心图像块（cp-HSI）以及对应的激光雷达信号。高光谱图像信号的假彩色图像中，绿色部分表示草地，棕色部分表示土壤区域。这三种输入可针对高光谱图像-激光雷达信号生成九个互联视角。此外，这九个相互关联的视图可以相互补充、协同工作，生成融合特征，从而消除各视角带来的偏差，并聚焦于草地区域。

以及基于编码器-解码器的融合，从有限的视角考虑特征融合，无法全面考量模态间关系 [11], [12], [13], [14], [15]。特征拼接后，来自相应模态内特征的融合神经元会被激活，而来自另一模态内特征的神经元则会受到抑制 [6]。大多数高光谱图像-激光雷达融合方法仅考虑单一模态（如高光谱图像信号）内的空间相关性，却忽略了模态间图像块的上下文信息，从而限制了最终性能。此外，其中一些融合方法关注高光谱图像内的模态内关系，却忽视了激光雷达中的模态内关系，导致信息缺失 [16]。此外，现有特征融合方法对未见过的训练样本缺乏泛化能力，且容易过拟合 [17]。

如图 1 所示，本文提出了一种互连融合（IF）框架，以丰富模态内和模态间信息，并获得令人满意的数据表示。我们将高光谱图像输入划分为 3×3 个图像块，提取高光谱图像输入的中心图像块，并将其复制至与高光谱图像输入相同的尺寸，作为额外输入送入框架。随后，引入融合矩阵，通过应用自注意力和交叉注意力操作，捕获并表示三种多模态之间的互连相关性

输入，包括高光谱图像、复制得到的高光谱图像中心图像块，以及相应的激光雷达数据。之后，融合矩阵中的九个相互关联的元素可以协同工作，通过特征融合层准确生成融合特征表示。

总体而言，我们的主要贡献总结如下：
1）将高光谱图像的中心图像块作为额外输入引入我们的框架，从而提取中心图像块与邻近块之间的上下文信息，以丰富模态内和模态间的信息。

2）不同于拼接、求和以及基于编码器-解码器的融合等传统融合方法，我们首次将融合矩阵中的九个互联元素作为高光谱图像-激光雷达信号的特征图引入；这些元素能够相互补充、协同工作，从而消除各元素带来的偏差。3）与其他方法相比，所提出的方法在三个广泛使用的数据集（特伦托、MUUFL和Houston）上优于现有方法。此外，我们还扩展了实验，研究图像块大小变化的影响以及早期、中期和后期等融合策略的影响。我们发现，当输入图像块包含更多像素时，IF框架在不同融合策略下均能实现稳定的性能，这说明IF是一种稳健的框架，能够在不同融合策略下取得相近的性能。

II. 相关工作

A. 单模态信号表示方法

高光谱图像领域常用的单模态信号表示方法包括基于卷积神经网络（CNN）的方法 [18], [11], [19], [12], [20], [13]。带像素对特征的深度CNN（CNN-PPF） [18] 和一维CNN [21] 结合 ℓ_2 正则化，采用像素对CNN结构。然而，针对高光谱图像区域的逐像素信息仅在像素级探索光谱特征，而未考虑任何空间关系。相邻像素的光谱数据具有相似属性，因此可以从中挖掘更多判别能力，以处理高光谱图像数据 [22], [23], [24], [25], [26], [27], [28], [19], [29]。然而，如果CNN模型同时提取光谱和空间信息，就会过多关注空间方向知识，而非光谱信息，这意味着每个像素的大量高光谱信息未得到充分利用 [30]。此外，基于CNN的模型还存在诸多局限，例如忽略长程依赖和位置信息 [31]。

视觉Transformer（ViT） [32] 的性能优于卷积神经网络结构。此外，基于 Transformer 的模型参数更少，过拟合风险更低，同时能够提供比基于CNN的模型更稳健且泛化能力更强的特征 [33], [34]。SpectralFormer [30] 能够从高光谱图像的相邻波段中学习分组光谱嵌入。此外，一些基于 Transformer 的方法采用光谱-空间Transformer架构，以同时获取空间和光谱注意力 [35], [36]。遗憾的是，大多数基于 Transformer 的模型参数量庞大，计算成本高昂 [37]。此外，这些模型的性能严重依赖大规模训练数据 [32]。因此，训练基于 Transformer 的模型仍面临诸多挑战。在我们的框架中，我们采用ViT作为骨干网络，以提取通用的密集视觉特征 [31] ，并结合中心图像块与邻近块之间的邻域关系来提取上下文相关特征。为缓解ViT对数据需求量大的局限，我们的框架深度设为三层。

B. 多模态融合方法

多模态高光谱图像领域的特征融合算法通常采用传统融合方法，包括求和、拼接、编码器-解码器和交叉融合 [11], [12], [13], [14], [15]。双分支卷积神经网络分别使用两个子网络从高光谱图像和激光雷达数据中提取特征，然后通过拼接组合这两种不同的特征 [11]。此外，双分支卷积神经网络的权重共享（Co-CNN）也已用于高光谱图像和激光雷达的特征提取。之后，使用求和或最大化策略在特征层融合这两类特征 [20]。编码器-解码器架构（EndNet）也已应用于该领域。该方法使用自编码器整合并重建不同模态的信息[12]。另一种基于自编码器的模型是交错感知卷积神经网络（IP-CNN），它采用双向自编码器，对高光谱图像和激光雷达数据源进行信息融合、特征提取和重建 [13]。此外，共享与特定特征学习（S2FL）模型将多模态高光谱图像数据划分为两大类：不同模态共享的空间，以及各模态特有的空间 [38]。此外，MDL-RS框架采用基于紧致性的交叉融合，以获取不同模态之间更多样化的关系[6]。然而，卷积神经网络结构提取的特征明显偏重纹理，且位置信息严重不足 [31]。因此，卷积神经网络结构在捕获输入各元素之间的全局交互方面仍然存在局限。

近年来，ViT 已应用于高光谱图像多模态领域，因为与卷积神经网络相比，ViT 能够取得有竞争力的性能。多模态融合 Transformer（MFT）网络通过来自另一种模态的分类标记来融合不同模态的特征 [39]。然而，这些融合方法忽略了多模态信号的上下文信息以及模态自身的信息，因此性能提升有限；而我们的框架则提取中心图像块与邻近块之间的上下文信息，以捕获模态内和模态间的信息。

III. 互联融合框架

K

所提出的IF框架包含三个阶段：模态内特征编码、互联融合和分类（见图2和算法1）。第一阶段采用视觉 Transformer 编码器，分别提取复制的中心图像块、高光谱图像输入和激光雷达输入的三种模态内特征

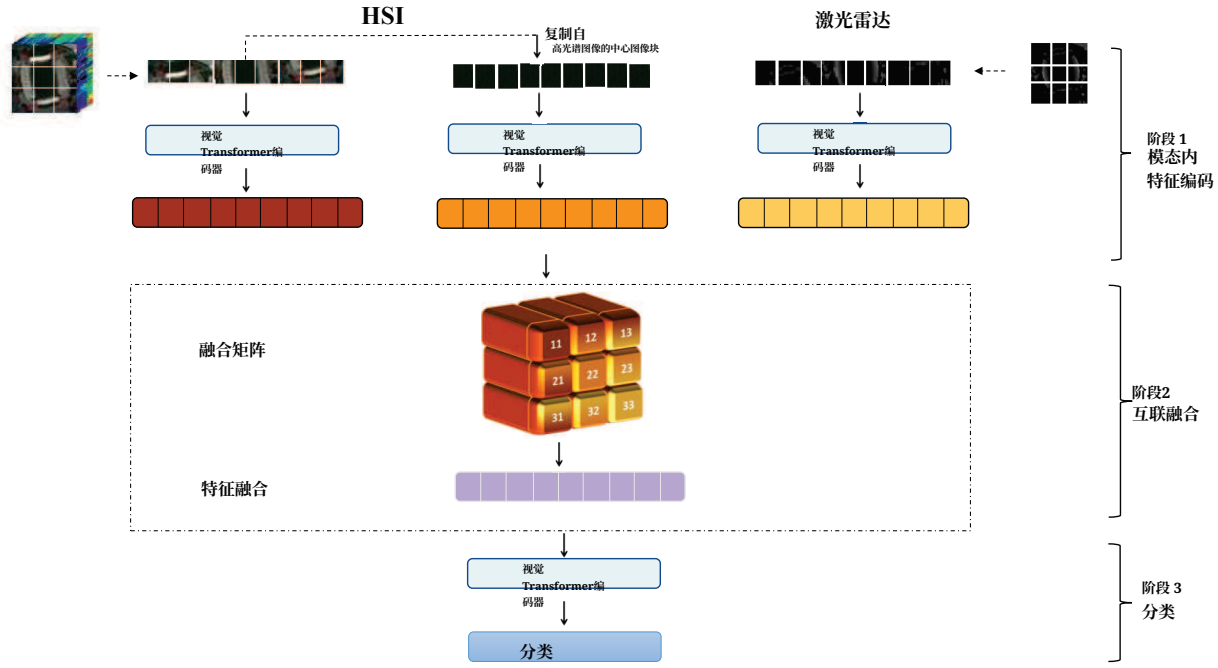


图 2. 所提出的 IF 框架包含三个阶段：模态内特征编码、互联融合和分类。在模态内特征编码阶段，将高光谱图像的中心图像块作为额外输入，以探索中心图像块与邻近块之间的上下文信息，适用于模态内和模态间。对于互联融合阶段，融合矩阵中的九个互联视角可以表示包含上下文信息的模态内和模态间关系。这九个相互关联的视图可以相互补充，并通过特征融合层消除彼此的偏差。在最后阶段，可以计算融合特征，并获得分类结果。

。随后，第二阶段引入融合矩阵，通过计算三种模态内特征的自注意力和交叉注意力，汇集模态内和模态间信息的九个视角。之后，融合矩阵中的九个相互关联元素可以相互补充，并通过特征融合层形成融合特征。此外，最后阶段采用视觉Transformer编码器获取更加抽象的融合特征和分类结果。

A. 模态内特征编码

对于模态内特征编码，引入高光谱图像的中心图像块作为额外输入，以进一步探索中心图像块与邻近块之间的上下文信息。输入包含三个数据源：高光谱图像 X_{h1}、对应的激光雷达 X_l，以及高光谱图像数据的中心图像块 X_{h2}，它们分别输入三个独立编码器：用于高光谱图像的 H、H_p，以及用于激光雷达的 L。高光谱图像和激光雷达的输入在块嵌入后具有相同的图像块数量 p 和嵌入维度 D。X_{h2} 的数据来自高光谱图像的中心图像块 X_{h1}，并重复 p 次（图像块数量与高光谱图像和激光雷达样本相同）。随后，高光谱图像和激光雷达样本经过 Transformer 层，该层包含多头自注意力（MSA）、前馈网络（FFN）和层归一化（LN）模块。多头自注意力（MSA）可通过拼接不同注意力头的自注意力（SA）结果来计算然后通过权重矩阵 W 进行投影（见公式 1、2）：

$$MSA(z) = [SA_1(z) ..., SA_h(z)]W \qquad W \in \mathbb{R}^{hD_q \times D} \quad (1)$$

其中，h 是多头自注意力的头数，$SA(z)$可计算为：

$$SA(z) = softmax(\frac{qk^{\intercal}}{\sqrt{D_q}})v \qquad [q,k,v] = zU_{qkv} \qquad U_{qkv} \in \mathbb{R}^{D \times 3D_q} \quad (2)$$

其中，q、k 和 v 分别表示查询、键和值。多头自注意力的结果 Z_{h1}^l、Z_{h2}^l 和 Z_l^l 可计算为：

$$\hat{Z}_i^l = MSA(LN(Z_i^{l-1})) + Z_i^{l-1} \qquad i \in \{h1, h2, l\} \quad (3)$$

然后，前馈神经网络模块可以更新多头自注意力的输出。

$$Z_i^l = MLP(LN(\hat{Z}_i^l)) + \hat{Z}_i^l \qquad i \in \{h1, h2, l\} \quad (4)$$

因此，编码器 H、H_p 和 L 分别生成了三种不同的结果 Z_{h1}^l、Z_{h2}^l 和 Z_l^l。此外，经过模态内数据处理后，这三个输出具有相同数量的图像块和相同维度的嵌入。

B. 互联融合

IF框架不仅关注不同模态的特征融合，还致力于利用全面的

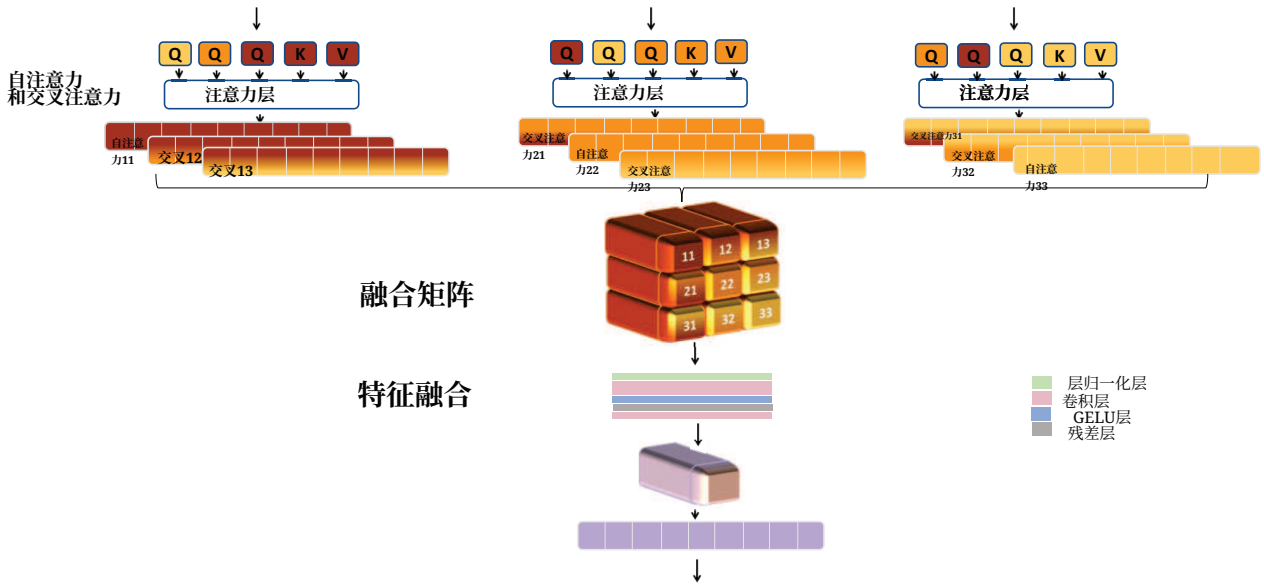


图3. 互联融合阶段利用融合矩阵，结合中心图像块与邻近块的邻域相关性，汇集模态内和模态间信息的九个视角。随后，融合矩阵中的九个相互关联的视图可以互为补充，并通过特征融合层形成融合特征。

算法1 实现核心的伪代码

IF框架的运行1: 初始化：设置 $t = 0$，输入数据 X_{h1}，X_{h2}来自高光谱图像数据源，X_l 来自激光雷达数据源。具体而言，将高光谱图像输入 X_{h2} 的中心图像块复制至与高光谱图像输入 X_{h1}相同的大小。2: 对于 $t = 0$、1、2、$\cdots$，执行3: 模态内特征编码：分别通过视觉Transformer 编码器（参见公式 (4)）从输入数据 X_{h1}，X_{h2}，X_l 中提取特征表征 Z_{h1}^l，Z_{h2}^l 和 Z_l^l；4: 互联融合：计算九个互联视角：$Z_{ij}^c = f_{views}(Z_{h1}^l, Z_{h2}^l, Z_l^l)$ $i,j \in \{1,2,3\}$；定义融合矩阵：$Z^m = \begin{bmatrix} Z_{11}^c & Z_{12}^c & Z_{13}^c \\ Z_{21}^c & Z_{22}^c & Z_{23}^c \\ Z_{31}^c & Z_{32}^c & Z_{33}^c \end{bmatrix}$ ˆ，计算综合且经过校正的特征 Z^f，然后更新至 Z^f（参见公式 (9) 和公式 (10)）；5: 分类：通过视觉Transformer编码器处理组合特征 Z^f（参见公式 (11) 和 (12)），并获取分类结果；6: 结束循环

不同模态之间的交互。对于互连融合阶段，提出融合矩阵中的九个互联元素，以表示模态内和模态间的信息，以及不同模态之间的上下文信息。融合矩阵中的九个互联元素可以相互补充，并通过特征融合层共同构成融合特征。

首先，在前一阶段计算三个模态内特征 Z_{h1}^l、Z_{h2}^l 和 Z_l^l。然后，通过对这三个模态内特征进行自注意力和交叉注意力，生成九个互联视角 Z_{ij}^c：

$$Z_{ij}^c = f_{views}(Z_{h1}^l, Z_{h2}^l, Z_l^l) \qquad i,j \in \{1,2,3\} \tag{5}$$

ˆ 当 $i = j$ 时，可通过由相同输入 i生成的k、q和 v 计算自注意力值 Z_{ij}^c。相反，ˆ 当 $i \neq j$ 时，可通过来自相同输入 i 的 k和 v 计算交叉注意力值 Z_{ij}^c，而 q 则来自另一个数据源j:

$$\hat{Z_{ij}^c} = MSA(q_j, k_i, v_i) + Z_i^l \qquad i,j \in \{1,2,3\} \tag{6}$$

ˆ Z_{ij}^c 的值可以通过残差连接和前馈神经网络模块进行更新:

$$Z_{ij}^c = MLP(LN(\hat{Z_{ij}^c})) + \hat{Z_{ij}^c} \qquad i,j \in \{1,2,3\} \tag{7}$$

然后，可以利用这九个视图定义 3×3 融合矩阵。

$$Z^m = \begin{bmatrix} Z_{11}^c & Z_{12}^c & Z_{13}^c \\ Z_{21}^c & Z_{22}^c & Z_{23}^c \\ Z_{31}^c & Z_{32}^c & Z_{33}^c \end{bmatrix} \tag{8}$$

在该融合矩阵中，模态内特征包括Z_{11}^c 高光谱图像的 Z_{33}^c 相应激光雷达的 Z_{22}^c 高光谱图像中心图像块的，以及从两个不同视角得到的中心图像块与邻近块之间的高光谱图像关系 Z_{12}^c、Z_{21}^c。此外，Z_{13}^c、Z_{31}^c、Z_{23}^c、Z_{32}^c 展示了模态间相关性。更具体地说，Z_{13}^c 和 Z_{31}^c表示位于对应位置的高光谱图像与激光雷达图像块之间的关系。相比之下，Z_{23}^c 和 Z_{32}^c 提供了高光谱图像中心图像块与邻近激光雷达图像块之间的相关性。此外，Z_{12}^c、Z_{21}^c、Z_{23}^c、Z_{32}^c 展示了

模态内和模态间的上下文信息。因此，需要注意的是，融合矩阵中的九个互联元素包含了模态内和模态间的相关性，并结合了上下文信息。

由于多模态遥感信号的信息有限，融合矩阵的每个元素都存在偏差。IF框架中设计的特征融合层能够有效生成融合特征，并从不同视角消除偏差，因为融合矩阵中的这九个相互关联的视图可以相互补充（详见第4节 消融研究）。随后， 3×3融合矩阵 Z^m 可以压缩为一个融合特征 Z^f:

$$\hat{Z^f} = f_{comp}(Z^m) \tag{9}$$

ˆ 随后，该特征 Z^f 通过前馈神经网络模块和残差连接进行更新:

$$Z^f = \hat{Z^f} + FFN(\hat{Z^f}) \tag{10}$$

C. 分类

在分类阶段，将融合结果 Z^f 视为单一特征，并由视觉Transformer编码器处理:

$$\hat{Z^u} = MSA(LN(Z^f)) + Z^f \tag{11}$$

ˆ此外， Z^u 由多头自注意力模块生成，随后由前馈神经网络模块处理:

$$Z^u = MLP(LN(\hat{Z^u})) + \hat{Z^u} \tag{12}$$

最后，可以计算结果 Z^u 。最终分类结果可以根据 cls 标记计算得出。

IV. 实验

本节采用三个知名的高光谱图像多模态数据集评估 IF 框架的性能。此外，我们比较了不同融合策略和图像块尺寸下的性能，并针对 IF 框架进行了不同的消融实验。

A. 数据集

休斯敦2013数据集采集于休斯敦大学校园及其附近的城市区域，采集时间为 2013 年 IEEE GRSS 数据融合竞赛期间。该数据集包含高光谱图像和激光雷达衍生的数字表面模型（DSM）信号。数据集大小为 349 × 1905，空间分辨率为 2.5 米。高光谱图像信号包含 144 个波段，覆盖 0.38 至 1.05μ 米。此外，该数据集包含 15 个类别，包括健康草地、胁迫草地、人工草地、树木、土壤、水体、住宅区、商业区、道路、高速公路、铁路、停车场1、停车场2、网球场和跑道。各类别的训练和测试设置见表I。

表I 休斯敦数据集各类别的标准训练集和测试集

类别		样本数量	
No.	Name	训练	测试
1	健康草	198	1251
2	受胁迫草	190	1254
3	人造草地	192	697
4	树木	188	1244
5	Soil	186	1242
6	水体	182	325
7	住宅区	196	1268
8	商业区	191	1244
9	Road	193	1252
10	高速公路	191	1227
11	铁路	181	1235
12	停车场1	192	1233
13	停车场2	184	469
14	网球场	181	428
15	跑道	187	660
总计		2832	15029

Trento数据集采集于意大利特伦托的农村地区。采集图像尺寸为600×166像素，空间分辨率为1米。此外，该数据集包含数字表面模型和高光谱图像信号，共有六个不同类别：建筑物、林地、苹果树、道路、葡萄园和地面。高光谱图像包含63个波段，范围从0.40到0.98 μ米。各类别的训练集和测试集设置如表 II 所示。

表 II 特伦托数据集中每个类别的标准训练集和测试集。

类别		样本数量	
No.	Name	训练	测试
1	苹果树	129	4034
2	建筑物	125	2903
3	地面	105	479
4	林地	154	9123
5	葡萄园	184	10501
6	道路	122	3174
总计		819	30214

MUUFL 数据集 [40] 于 2010 年在南密西西比大学校园内采集。该数据集的大小为 325×220 像素，高光谱图像信号包含 72 个波段。该数据集还包含两种不同高度的激光雷达信号。在我们的实验中，我们沿像素位置拼接这两种激光雷达信号。此外，该数据集包含 11 个类别：树木、大部分为草地、混合地表、泥土和沙地、道路、水体、建筑物阴影、建筑物、人行道、黄色路缘和布料面板。各类别的训练集和测试集设置见表 III。

B. 实现细节

我们在配备 Intel(R) Xeon(R) Gold 6132 CPU 和 Nvidia Quadro P5000 GPU 的 Linux 系统上运行模型。我们在 IF 框架中实现了三种融合策略设置——早期、中期和后期。若总层数为 N，阶段 1 的层数为 M，则阶段 3 的深度为 N-M-1 （见图 4 ）。为缓解 ViT 对数据量的需求限制，我们将框架的总深度设为三。对于早期融合设置，阶段 1 的深度为零，阶段 3 的层数为两层。相反，对于晚期融合设置，阶段 1 的深度为两层，阶段 3 的层数为零。对于中期融合设置，阶段 1 的深度为一层，阶段 3 的层数也为一层。因此，可以通过调整阶段 1 和阶段 3 的深度来定制 IF 框架的融合策略设置

表 III MUUFL 数据集各类别的标准训练集和测试集

类别		样本数量	
No.	Name	训练	测试
1	树木	150	23246
2	大部分为草地	150	4270
3	混合地表	150	6882
4	泥土和沙地	150	1826
5	Road	150	6687
6	水体	150	466
7	建筑物阴影	150	2233
8	建筑物	150	6240
9	人行道	150	1385
10	黄色路缘	150	183
11	布料面板	150	269
总计		1650	53687

表 IV 特征融合层的详细信息

层 No.	层类型	详细设置
1	层归一化	形状 [3,3]
2	卷积层	卷积核大小：(3,3)；步幅：(1,1)；填充：1
3	卷积层	卷积核大小：(1,1)；步幅：(1,1)；填充：0
4	激活函数	GELU()
5	残差层	–
6	卷积层	卷积核大小：(3,3)；步幅：(1,1)；填充：0

此外，我们借鉴 ConvNet [41]，以有效提取融合特征。特征融合层的详细设置见表 IV。因此，经过特征融合层后，我们可以得到融合特征 Z^f。



图 4. 若总层数为 N，且阶段 1 的层数为 M，则阶段 3 的深度为 N-M-1。

C. 结果比较

我们在表 V、VI 和 VII 中将我们的方法与以往方法进行性能比较：CNN-PPF [18], 双分支卷积神经网络 [11], 上下文 CNN [19], EndNet[12], MDL-RS[6], Co-CNN [20], S2FL[38], SPECTRALFORMER [30], MFT [39] IP-CNN [13]。值得注意的是，在这些表格中，红色值表示当前评估中的最佳结果，蓝色值表示次佳结果。

休斯敦数据集上的评估 如表 V 所示，IF取得了最佳性能。IF的总体精度（OA）、平均精度（AA）和Kappa系数分别比排名第二的方法（IP-CNN）高出约5%、3%和6%。此外，对于大多数类别，IF框架都能取得有竞争力的结果，而IP-CNN的性能排名第二，这意味着IF能够全面捕捉不同模态之间的关系，且不会像IP-CNN那样面临数据泄漏的风险。IF的实现使性能显著提升了14%，这表明与单一高光谱图像模态相比，我们提出的方法能够有效挖掘更多信息。与双分支卷积神经网络、上下文 CNN、Co-CNN、EndNet、MDL-RS和S2FL相比，IF的结果提升了8%，因为针对高光谱图像-激光雷达信号收集的九种不同视角包含了丰富的信息，有助于揭示模态内和模态间关系，包括上下文信息。这九种不同视角有助于消除由多种视角带来的潜在偏差。与SPECTRALFORMER和MFT相比，IF的性能显著提升。尽管IF、SPECTRALFORMER和MFT都采用Transformer作为骨干网络，但SPECTRALFORMER和MFT未能充分挖掘不同模态的特征，而IF能够从高光谱图像多模态数据源中获取全面的信息。

MUUFL 数据集上的评估 如表 VI 所示，与排名第二的 IP-CNN 相比，IF 的总体精度高出 4% 以上。同样，与 IP-CNN 相比，IF 的平均精度和 Kappa 系数分别约高出 2% 和 5%，这表明 IF框架能够从不同视角融合不同模态，并结合上下文信息，具有显著的能力。与 CNN-PPF 相比，IF 的性能提高了 7%，这意味着 IF 能够有效融合不同模态的信息，同时不影响单一高光谱图像模态的性能。与双分支卷积神经网络、上下文 CNN、Co-CNN、EndNet、MDL-RS 和 S2FL 相比，IF框架取得了有竞争力的结果，因为它考虑了从不同视角出发的不同模态间交互，而非采用视角有限的传统融合方式，或仅进行不足的多模态交互。与 SPECTRALFORMER 和 MFT 相比，IF 的性能同样具有竞争力，因为 IF 包含丰富的不同模态间信息。

在 Trento 数据集上的评估 如表VII所示，IF框架的结果略低于IP-CNN，因为IP-CNN需要两阶段训练，即在同一数据集上进行无监督训练和监督训练，这可能导致信息泄漏。此外，IF可能会在融合矩阵上出现过拟合风险，因为该矩阵需要容纳不同模态之间的大量交互。与其他方法相比，IF框架取得了具有竞争力的结果，表明它在探索不同模态间更全面的关系以及从九个视角消除偏差方面具有优势。

因此，IF框架能够在特伦托、MUUFL和休斯敦数据集上取得最佳性能，同时不会像IP-CNN一样面临数据泄漏风险。此外，与其他融合方法相比，IF取得了具有竞争力的结果，因为IF通过融合矩阵中的九个互联视角，从不同模态中提取全面特征。这九个相互关联的元素可以相互补充，从而消除潜在偏差。相比之下，其他融合方法仅从有限视角采用传统融合方法，这可能导致不同

表 V 我们在休斯敦数据集上将 IF 与 CNN-PPF、双分支卷积神经网络、上下文 CNN、EndNet、MDL-RS、S2FL、Co-CNN、SPECTRALFORMER、MFT 和 IP-CNN 进行了比较。

No.	类别	性能											
		CNN-PPF[18]	双分支 卷积神经网络[11]	双分支 CNN（融合）	上下文 卷积神经网络[19]	EndNet[12]	MDL-遥感[6]	S2FL[38]	Co- 卷积神经网络 [20]	光谱变换器[30]	MFT[39]	IP- 卷积神经网络[13]	IF
1	健康草	83.57	83.10	83.10	84.89	81.58	83.10	81.67	98.51	82.65	82.34	85.77	96.58
2	受胁迫草	98.21	84.10	81.20	87.40	83.65	81.58	82.99	97.83	83.33	88.78	87.34	99.44
3	人造草地	98.42	100.00	100.00	99.86	100.00	100.00	100.00	70.60	75.78	98.15	100.00	91.49
4	树木	97.73	93.09	92.90	93.49	93.09	99.72	92.05	99.06	91.10	94.35	94.26	95.64
5	Soil	96.50	100.00	99.81	99.72	99.91	99.81	99.43	100.00	98.30	99.12	98.42	99.72
6	水体	97.20	99.30	100.00	98.77	95.10	95.10	99.30	41.11	89.04	99.30	99.91	92.31
7	住宅区	85.82	92.82	92.54	82.81	82.65	90.02	76.31	83.14	81.72	88.56	94.59	98.23
8	商业区	56.51	82.34	94.87	78.78	81.29	87.94	68.09	98.39	67.81	86.89	91.81	99.43
9	Road	71.20	84.70	83.85	82.51	88.29	81.59	69.22	94.81	74.47	87.91	89.35	95.18
10	高速公路	57.12	65.44	69.89	59.41	89.00	86.68	52.12	92.98	56.76	64.70	72.43	98.65
11	铁路	80.55	88.24	86.15	83.24	83.78	89.37	90.42	90.88	59.93	98.64	96.57	99.62
12	停车场1	62.82	89.53	92.60	92.13	90.39	85.69	87.42	91.02	70.00	94.24	95.60	99.23
13	停车场2	63.86	92.28	79.30	94.88	82.46	83.16	70.18	97.09	66.20	90.29	94.37	94.39
14	网球场	100.00	96.76	100.00	99.77	100.00	100.00	100.00	100.00	92.04	99.73	99.86	99.60
15	跑道	98.10	99.79	100.00	98.79	98.10	98.73	98.31	97.85	77.45	99.58	99.99	91.33
总体精度 (%)		83.33	87.98	88.91	86.90	88.52	89.60	81.95	90.43	76.87	89.80	92.06	97.50
AA (%)		83.21	90.11	90.42	89.11	89.95	90.83	84.50	90.22	77.77	91.51	93.35	96.72
Kappa系数 [%]		81.88	86.98	87.96	85.89	87.59	88.75	80.40	89.68	75.03	88.93	91.42	97.29

表 VI 我们将 IF 与 CNN-PPF、双分支卷积神经网络、上下文 CNN、EndNet、MDL-RS、S2FL、Co-CNN、SPECTRALFORMER、MFT 和 IP-CNN 在 MUUFL 数据集上的表现进行了比较。

No.	类别	性能											
		CNN -PPF[18]	双分支 卷积神经网络[11]	双分支 卷积神经网络（融合）	上下文 卷积神经网络[19]	EndNet [12]	MDL-遥感[6]	S2FL [38]	Co- 卷积神经网络[20]	光谱变换器[30]	MFT[39]	IP- 卷积神经网络[13]	IF
1	树木	89.07	92.35	94.28	91.29	89.00	88.57	72.40	98.90	97.30	97.90	94.40	98.87
2	大部分为草地	85.71	59.30	82.84	63.09	84.73	84.00	75.97	78.60	69.35	92.11	92.26	96.12
3	混合地表	80.15	94.47	84.58	81.84	73.39	73.79	54.72	90.66	78.48	91.80	87.96	95.94
4	泥土和沙地	93.10	93.74	97.32	93.92	91.95	93.48	82.20	90.60	82.63	91.59	97.15	99.16
5	Road	88.98	92.76	92.96	89.44	87.72	88.54	71.26	96.90	87.91	95.60	94.38	97.87
6	水体	98.93	98.42	99.05	92.92	99.14	98.42	94.42	75.98	58.77	88.19	99.79	98.42
7	建筑物阴影	89.07	95.68	93.28	84.73	91.58	90.46	77.34	73.54	85.87	90.27	96.30	97.17
8	建筑物	92.15	94.01	95.42	81.22	91.30	92.95	86.19	96.66	95.60	97.26	96.13	98.87
9	人行道	75.45	86.64	84.13	81.30	83.75	81.37	59.21	64.93	53.52	61.35	94.01	93.36
10	黄色路缘	100.00	100.00	100.00	98.91	100.00	100.00	98.91	19.47	08.43	17.43	100.00	96.97
11	布料面板	100.00	96.64	97.48	99.63	99.63	99.26	98.88	62.76	35.29	72.79	99.63	99.16
总体精度 (%)		90.97	90.35	91.95	86.07	87.02	87.54	72.49	90.93	88.25	94.34	93.86	97.96
平均精度 (%)		90.24	91.27	92.58	87.12	90.20	90.14	79.23	77.18	68.47	81.48	95.64	97.45
Kappa系数 (%)		84.46	87.27	89.35	81.89	83.24	83.31	65.81	88.22	84.40	92.51	91.99	97.27

模态的信息不足，并且包含来自有限视角的判别信息。因此，IF能够为高光谱图像多模态数据融合带来全新的视角。

D. 融合策略与图块尺寸实验

此外，我们还探讨了图像块尺寸和融合策略的变化对最终性能的影响。我们比较了输入数据源的四种不同尺寸设置（包括 3×3、6×6、9×9 和 12×12）以及三种不同的融合策略设置，即早期、中期和后期。在阶段 1 中，我们在所有情况下都将输入的图像块数量设为 9，这意味着 3×3 的小图像块尺寸处于像素级，而 6×6、9×9 和 12×12 处于图块级。如表VIII、IX和X所示，在三个数据集中，3×3 在不同融合策略中性能最差，因为像素级输入提供的邻域信息不足，无法探究中心图像块与邻近图像块之间的关系，进而分析模态内与模态间的关系。随着图像块尺寸增大，性能呈现明显的上升趋势，这表明图像块尺寸对最终性能有积极影响。值得注意的是，当输入处于图块级时，无论采用何种融合策略，IF框架都能取得相近的性能，这意味着IF框架在不同融合策略下均能保持稳定的性能。

E. 消融研究

为进一步研究各个所提出模块的影响，我们在三个广泛使用的数据集上进行了消融研究。为探究使用多模态的有效性，我们实现了一个仅基于高光谱图像的模型。借助其他模态提供的充分信息，我们发现，在 Houston、特伦托和 MUUFL 数据集上的总体精度分别显著提升了 7.19%、3.86 % 和 3.01%。这表明，不同模态中的丰富信息对高光谱图像分类至关重要。如图 5 所示，融合矩阵中的元素

表VII 我们在Trento数据集上将 IF 与 CNN-PPF、双分支卷积神经网络、上下文 CNN、EndNet、MDL-RS、S2FL、Co-CNN、SPECTRALFORMER、MFT 和 IP-CNN 进行了比较。具体而言，IP-CNN 需要包括无监督训练和监督训练在内的两阶段训练，这将导致信息泄漏。

No.	类别	性能											
		CNN -PPF[18]	双分支 卷积神经网络[11]	双分支 CNN（融合）	上下文 卷积神经网络[19]	EndNet [12]	MDL- 遥感[6]	S2FL [38]	Co- 卷积神经网络[20]	光谱变换 器[30]	MFT[39]	IP- 卷积神经网络[13]	IF
1	苹果树	90.11	98.07	98.51	99.26	88.19	88.58	79.26	99.87	96.76	98.23	99.00	99.15
2	建筑物	83.34	95.21	92.49	86.81	98.49	95.86	92.44	83.84	97.25	99.34	99.40	98.92
3	地面	71.13	93.32	100.00	97.91	95.19	93.58	92.78	87.09	58.47	89.84	99.10	84.49
4	林地	99.04	99.93	97.32	97.31	99.30	99.22	94.94	99.98	99.24	99.82	99.92	100.00
5	葡萄园	99.37	98.78	100.00	99.82	91.96	83.82	72.83	99.61	93.52	99.93	99.66	99.94
6	道路	89.73	89.98	92.56	84.63	90.14	76.51	86.40	98.75	73.39	88.72	90.21	90.14
总体精度 (%)		94.76	97.92	97.46	96.11	94.17	90.65	83.95	97.69	93.51	98.32	98.58	98.54
平均精度 (%)		88.97	96.19	96.80	94.29	93.88	89.60	86.44	94.86	86.44	95.98	97.88	95.44
Kappa系数 (%)		93.04	96.81	96.61	94.81	92.22	86.28	78.99	96.91	91.36	97.75	98.17	98.05

表VIII 休斯敦数据集性能随不同融合阶段和输入图像尺寸变化的变化

	早期				中期				Late			
指标	3 × 3	6 × 6	9 × 9	12 × 12	3 × 3	6 × 6	9 × 9	12 × 12	3 × 3	6 × 6	9 × 9	12 × 12
OA	85.67	94.36	96.37	97.50	84.82	94.02	95.95	96.38	82.74	94.78	93.69	97.03
AA	86.05	93.89	95.42	96.72	84.87	93.25	95.22	95.83	83.11	94.46	93.15	96.52
Kappa系数	84.46	93.88	96.06	97.29	83.52	93.51	95.60	96.07	81.27	94.33	93.15	96.78

表IX

TRENTO数据集的性　　能随不同融合阶段和输入图像尺寸的变化而变化

	早期				中期				Late			
指标	3 × 3	6 × 6	9 × 9	12 × 12	3 × 3	6 × 6	9 × 9	12 × 12	3 × 3	6 × 6	9 × 9	12 × 12
OA	97.28	98.24	98.54	98.39	97.05	97.51	98.50	98.33	97.20	97.81	98.37	98.32
AA	93.50	96.18	95.44	96.41	93.18	92.91	95.97	96.00	93.00	93.82	96.04	95.45
Kappa系数	96.35	97.65	98.05	97.85	96.04	96.66	97.99	97.76	96.24	97.07	97.81	97.75

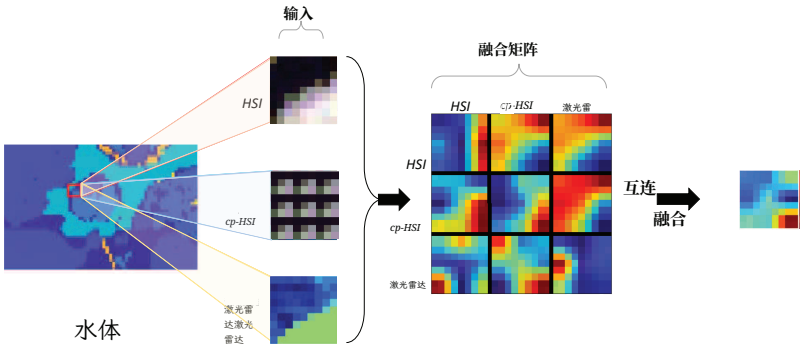
表 X MUFFL 数据集的性能随不同融合阶段和输入图像尺寸的变化情况

	早期				中期				Late			
指标	3 × 3	6 × 6	9 × 9	12 × 12	3 × 3	6 × 6	9 × 9	12 × 12	3 × 3	6 × 6	9 × 9	12 × 12
OA	95.49	97.08	97.48	97.84	95.89	97.29	97.73	97.96	95.74	97.09	97.48	97.78
AA	95.56	96.55	97.14	97.54	95.42	96.79	97.51	97.45	95.05	97.04	97.14	97.51
Kappa系数	94.00	96.11	96.63	97.11	94.50	96.38	96.97	97.27	94.31	96.11	96.63	97.04

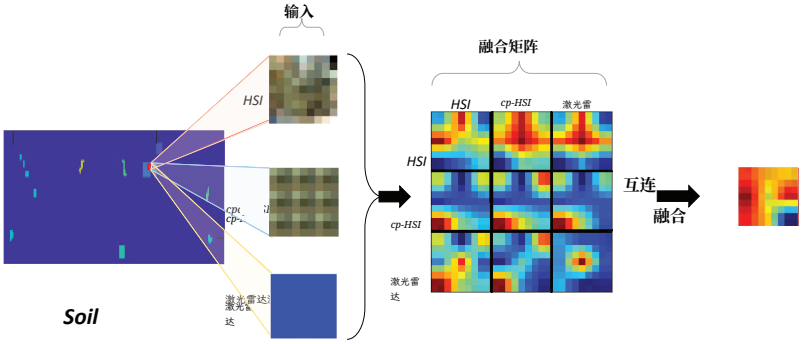
表 XI

此处对不同消融设置下的分类性能进行评估。不使用上下文信息意味着移除阶段 1 中的中心块模态内特征编码模块。不使用融合矩阵表示通过拼接融合矩阵中的九个元素。此外，高光谱图像模态表示实现一个仅基于高光谱图像的模型。

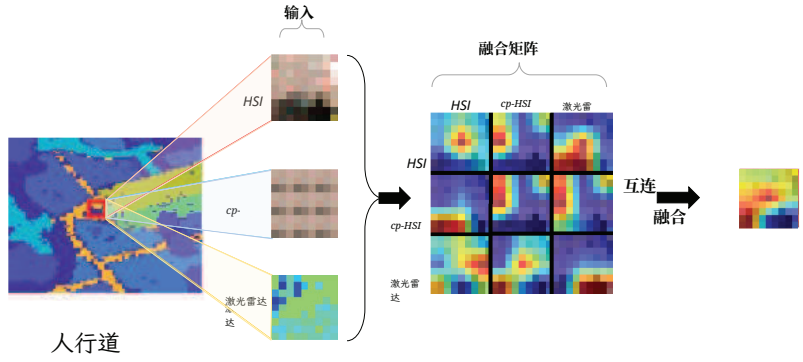
	指标	不含上下文信息	无融合矩阵	高光谱图像模态	IF框架
Houston	OA	96.41	91.35	90.31	**97.50**
	AA	95.76	95.31	90.10	**96.72**
	Kappa系数	96.10	95.56	88.96	**97.29**
特伦托	OA	97.73	97.51	95.56	**98.54**
	AA	93.92	95.26	90.97	**95.44**
	Kappa系数	96.96	96.67	94.05	**98.05**
MUUFL	OA	96.47	96.00	95.10	**97.96**
	AA	96.89	95.82	92.41	**97.45**
	Kappa系数	96.25	95.31	93.44	**97.27**



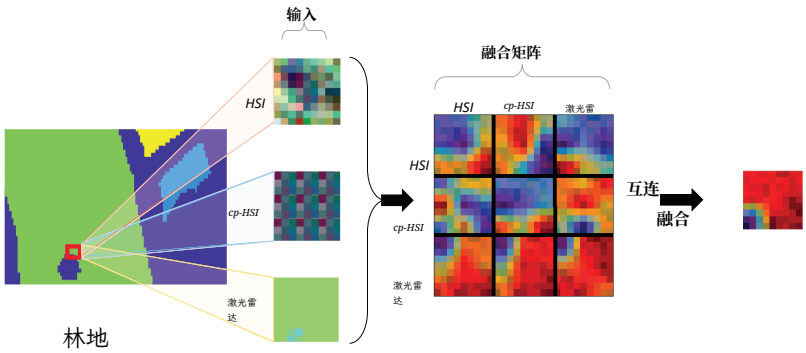
(a) 水体



(b) 土壤



(c) 人行道



(d) 木材

图5. 已对三个输入、融合矩阵以及互联融合阶段后的融合特征进行可视化。第一个输入显示了高光谱图像输入的伪彩色可视化。第二个输入是高光谱图像的中心图像块。第三个输入表示对应的激光雷达可视化结果（蓝色表示没有高度信息）。随后，计算并可视化了融合矩阵中的九个视角。经过互联融合阶段后，九个互联元素能够相互补充，并消除各个视角中的偏差。

融合矩阵左上角对应高光谱图像特征。我们观察到，仅利用一种模态的信息，难以将注意力集中在输入的正确部分。

然而，直接融合不同模态的信息并不能保证高光谱图像多模态分类性能显著提升。为研究所提出的融合矩阵的影响，我们通过拼接融合矩阵中的九个元素来替换融合矩阵和特征融合层。与单模态相比，采用拼接层的多模态模型在 Houston、特伦托和 MUUFL 数据集上的总体精度分别仅提高了 1.04%、2.05% 和 1.10%，远不及所提出的融合矩阵带来的提升。主要原因是，尽管多模态包含更多信息，但也会给模型引入更多噪声。然而，所提出的融合矩阵能够通过自适应调整不同模态的权重，消除这种负面影响。如图 5 所示，融合矩阵中每个元素的注意力主要集中在错误的空间位置上。然而，经过互联融合阶段后，整体注意力集中在恰当的空间位置，与人类感知一致。

最后，我们移除中心块内模态特征编码模块，以实现不考虑邻域关系的模型。可以观察到，在 Houston、特伦托和 MUUFL 数据集上，不考虑邻域关系的模型总体精度分别下降了 1.10%、0.79% 和 1.49%。这一结果表明，在高光谱图像多模态分类中，也应纳入邻域块的上下文信息。如图 5 所示，融合矩阵的四个顶点表示不含上下文信息的融合矩阵。这四个顶点的特征可能会聚焦于错误的部分。

V. 结论

我们提出了一种互联融合框架，用于考量不同模态间丰富的相关性。通过提取中心图像块与邻近块之间的上下文信息，可以丰富模态内与模态间的信息。高光谱图像与激光雷达信号包含模态内与模态间的信息。此外，融合矩阵通过九种不同视角生成模态内与模态间的相关性。融合矩阵中的九个相互关联元素可以相互补充、协同作用，从而消除各个视角带来的偏差。因此，我们的方法能够深入理解高光谱图像-激光雷达信号的特征融合。

致谢

作者谨此感谢休斯敦大学和 GRSS-DFC 提供休斯敦大学数据集。

参考文献

[1] B. Du、M. Zhang、L. Zhang、R. Hu 和 D. Tao，“Pltd：面向高光谱图像的基于图块的低秩张量分解”，IEEE 多媒体汇刊，第 19 卷，序号 1，第 67–79 页，2017年。

[2] C. Shi 和 C.-M. Pun，“利用堆叠自编码器循环神经网络进行多尺度超像素高光谱图像分类”，IEEE 多媒体汇刊，第 22 卷，501，2019 2，第 487 期，第 – 页。[3] P. Bajcsy 和 P. Groves，“高光谱波段选择方法”，《摄影测量工程与遥感》，第 70 卷，第 7 期，第 793–802 页，2004年。[4] D. Landgrebe，“高光谱图像数据分析”，IEEE 信号处理杂志，第 19 卷，第 1 期，第 17–28 页，2002年。[5] W. Sun 和 Q. Du，“高光谱波段选择：综述”，IEEE 地球科学与遥感杂志，第 7 卷，139，2019 2，第 118 期，第 – 页。[6] D. Hong、L. Gao、N. Yokoya、J. Yao、J. Chanussot、Q. Du 和 B. Zhang，“更多样化意味着更好：多模态深度学习与遥感影像分类的结合”，IEEE 地球科学与遥感汇刊，第 59 卷，第 5 期，第 4340–4354 页，2020年。[7] J. Li、D. Hong、L. Gao、J. Yao、K. Zheng、B. Zhang 和 J. Chanussot，“多模态遥感数据融合中的深度学习：全面综述”，arXiv 预印本 arXiv:2205.01380，2022年。[8] Y.-L. Chang、C.-C. Han、H. Ren、C.-T. Chen、K.-S. Chen 和 K.-C. Fan，“高光谱与合成孔径雷达图像的数据融合”，《光学工程》，第 43 卷，第 8 期，第 1787–1797 页，2004年。[9] B. Huang、Y. Li、X. Han、Y. Cui、W. Li 和 R. Li，“利用稀疏表示结合合成孔径雷达影像去除光学卫星影像中的云”，IEEE 地球科学与遥感快报，第 12 卷，第 5 期，第 1046–1050 页，2015年。[10] C. Debes、A. Merentitis、R. Heremans、J. Hahn、N. Frangiadakis、T. van Kasteren、W. Liao、R. Bellens、A. Pi˘zurica、S. Gautama、W. Philips、S. Prasad、Q. Du 和 F. Pacifici，“高光谱与激光雷达数据融合：2013年 GRSS 数据融合竞赛的成果”，IEEE《应用地球观测与遥感精选主题期刊》，第 7 卷，第 6 期，第 2405–2418 页，2014年。[11] X. Xu、W. Li、Q. Ran、Q. Du、L. Gao 和 B. Zhang，“基于卷积神经网络的多源遥感数据分类”，IEEE 地球科学与遥感汇刊，第 56 卷，第 2 期，第 937–949 页，2017年。[12] D. Hong、L. Gao、R. Hang、B. Zhang 和 J. Chanussot，“用于高光谱与激光雷达数据分类的深度编码器-解码器网络”，IEEE 地球科学与遥感快报，2020年。[13] M. Zhang、W. Li、R. Tao、H. Li 和 Q. Du，“使用 IP-CNN 对高光谱与激光雷达数据进行分类的信息融合”，IEEE 地球科学与遥感汇刊，第 60 卷，12，2021 1，第 – 页。[14] X. Wu 、 D. Hong 和 J. Chanussot，“ 用于多模态遥感数据分类的卷积神经网络”，IEEE 地球科学与遥感汇刊，第 60 卷，第 1–10 页，2021年。[15] X. Wang、Y. Feng、R. Song、Z. Mu 和 C. Song，“用于高光谱与激光雷达数据融合分类的多重注意力层次密集融合网络”，《信息融合》，第 82 卷，第 1–18 页，2022年。[16] P. Xu、X. Zhu 和 D. A. Clifton，“使用 Transformer 的多模态学习：综述”，arXiv 预印本 arXiv:2206.06488，2022年。[17] W. Wang、D. Tran 和 M. Feiszli，“是什么让多模态分类网络的训练如此困难？”，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，2020年，第 12 695–12 705 页。[18] W. Li、G. Wu、F. Zhang 和 Q. Du，“利用深度像素对特征进行高光谱图像分类”，IEEE 地球科学与遥感汇刊，第 55 卷，第 2 期，第 844–853 页，2016年。[19] H. Lee 和 H. Kwon，“利用上下文卷积神经网络深入开展高光谱图像分类”，IEEE《图像处理汇刊》，第 26 卷，第 10 期，第 4843–4855 页，2017年。[20] R. Hang、Z. Li、P. Ghamisi、D. Hong、G. Xia 和 Q. Liu，“使用耦合卷积神经网络对高光谱与激光雷达数据进行分类”，IEEE 地球科学与遥感汇刊，第 58 卷，第 7 期，第 4939–4950 页，2020 年。[21] Y. Chen 、 H. Jiang 、 C. Li 、 X. Jia 和 P. Ghamisi，“基于卷积神经网络的高光谱图像深度特征提取与分类”，IEEE 地球科学与遥感汇刊，第 54 卷，第 10 期，第 6232–6251 页，2016年。[22] L. He、J. Li、C. Liu 和 S. Li，“光谱-空间高光谱图像分类的最新进展：综述与新指南”，IEEE 地球科学与遥感汇刊，第 56 卷，第 3 期，第 1579–1597 页，2017年。[23] P. Ghamisi、E. Maggiori、S. Li、R. Souza、Y. Tarablaka、G. Moser、A. De Giorgi、L. Fang、Y. Chen、M. Chi 等，“光谱-空间高光谱图像分类的新前沿：基于数学形态学、马尔可夫随机场、分割、

、稀疏表示和深度学习的最新进展”，IEEE 地球科学与遥感杂志，第 6 卷，第 3 期，第 10–43 页，2018年。[24] Z. Zhong、J. Li、Z. Luo 和 M. Chapman，“用于高光谱图像分类的光谱-空间残差网络：三维深度学习框架”，IEEE 地球科学与遥感汇刊，第 56 卷，第 2 期，第 847–858 页，2017年。[25] S. Jia、J. Liao、M. Xu、Y. Li、J. Zhu、W. Sun、X. Jia 和 Q. Li，“用于高光谱图像分类的三维 Gabor 卷积神经网络”，IEEE 地球科学与遥感汇刊，第 60 卷，第 1–16 页，2022年。[26] S. K. Roy、D. Hong、P. Kar、X. Wu、X. Liu 和 D. Zhao，“用于噪声标签高光谱图像分类的轻量级异构核卷积”，IEEE 地球科学与遥感快报，第 19 卷，第 1–5 页，2022年。[27] A. Plaza、J. A. Benediktsson、J. W. Boardman、J. Brazile、L. Bruzzone、G. Camps-Valls、J. Chanussot、M. Fauvel、P. Gamba、A. Gualtieri 等，“高光谱图像处理技术的最新进展”，《遥感环境》，第 113 卷，第 S110–S122 页，2009年。[28] M. J. Khan、H. S. Khan、A. Yousaf、K. Khurshid 和 A. Abbas，“高光谱图像分析的现代趋势：综述”，IEEE Access，第 6 卷，第 14 118–14 129 页，2018年。[29] R. Ji、Y. Gao、R. Hong、Q. Liu、D. Tao 和 X. Li，“光谱-空间约束高光谱图像分类”，IEEE 地球科学与遥感汇刊，第 52 卷，第 3 期，第 1811–1824 页，2013年。[30] D. Hong、Z. Han、J. Yao、L. Gao、B. Zhang、A. Plaza 和 J. Chanussot，“SPECTRALFORMER：重新思考基于 Transformer 的高光谱图像分类”，IEEE 地球科学与遥感汇刊，第 60 卷，第 1–15 页，2021年。[31] S. Amir、Y. Gandelsman、S. Bagon 和 T. Dekel，“作为密集视觉描述符的深度 ViT 特征”，arXiv 预印本 arXiv:2112.05814，第 2 卷，第 3 期，第 4 页，2021年。[32] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly 等，“一幅图像胜过 16×16 个词：用于大规模图像识别的 Transformer”，arXiv 预印本 arXiv:2010.11929，2020年。[33] M. M. Naseer、K. Ranasinghe、S. H. Khan、M. Hayat、F. Shahbaz Khan 和 M.-H. Yang，“视觉 Transformer 的有趣特性”，《神经信息处理系统进展》，第 34 卷，第 23 296–23 308 页，2021年。[34] H.-Y. Zhou、C. Lu、S. Yang 和 Y. Yu，“卷积神经网络与 Transformer：谁的视觉表征更具迁移性？”，载于 IEEE/CVF 国际计算机视觉会议论文集，2021年，第 2230–2238 页。[35] Z. Zhong、Y. Li、L. Ma、J. Li 和 W.-S. Zheng，“用于高光谱图像分类的光谱-空间 Transformer 网络：因式分解架构搜索框架”，IEEE 地球科学与遥感汇刊，第 60 卷，第 1–15 页，2021年。[36] L. Sun、G. Zhao、Y. Zheng 和 Z. Wu，“用于高光谱图像分类的光谱-空间特征词元化 Transformer”，IEEE 地球科学与遥感汇刊，第 60 卷，14，2022 1，第 – 页。[37] K. Han 、 Y. Wang 、 H. Chen 、 X. Chen 、 J. Guo 、 Z. Liu 、 Y. Tang、A. Xiao、C. Xu、Y. Xu 等，“视觉 Transformer 综述”，IEEE 模式分析与机器智能汇刊，2022年。[38] D. Hong、J. Hu、J. Yao、J. Chanussot 和 X. X. Zhu，“用于土地覆盖分类、采用共享与特定特征学习模型的多模态遥感基准数据集”，《摄影测量与遥感》，第 178 卷，第 68–80 页，2021年。[39] S. K. Roy、A. Deria、D. Hong、B. Rasti、A. Plaza 和 J. Chanussot，“用于遥感图像分类的多模态融合 Transformer”，arXiv 预印本 arXiv:2203.16952，2022年。[40] P. Gader、A. Zare、R. Close、J. Aitken 和 G. Tuell，“MUUFL Gulfport 高光谱与激光雷达机载数据集”，佛罗里达大学，佛罗里达州盖恩斯维尔，美国，技术报告 REP-2013-570，2013年。[41] Z. Liu、H. Mao、C.-Y. Wu、C. Feichtenhofer、T. Darrell 和 S. Xie，“面向 2020 年代的卷积神经网络”，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，2022年，第 11 976–11 986 页。

这张图“fig1.png”提供“png”格式版本，获取地址：

http://arxiv.org/ps/2304.00495v1