DOI:10.19651/j.cnki.emt.2519142

全局-局部特征融合的多尺度遥感检索算法*

周晨菡 许晓阳 魏 伟
(西安科技大学人工智能与计算机学院 西安 710054)

摘 要:针对遥感图像在跨模态检索任务中存在图像冗余信息干扰、多尺度信息提取不足、全局与局部信息无法有效融合导致的检索精度较低等问题,提出一种适用于多尺度任务的遥感图文检索模型 IGMR。首先,设计多维感知增强卷积模块 MFE,提取局部信息的同时过滤冗余特征,并通过融合多注意力模块来关注图像高频信息,提升特征表达能力。其次,设计多尺度分块注意力网络 RFPA,捕获不同尺度的上下文信息。随后,构建自适应特征融合模块 AFFM,将提取到的全局与局部特征进行动态融合,增强对高质量信息的关注。在公开数据集 RSICD 和 RSITMD 上的实验结果表明,提出的 IGMR 方法在遥感跨模态检索任务中,平均召回率 mR 分别提高了 1.83%、3.21%,检索精度达到了 19.73%和 31.83%,总体检索性能显著提升。

关键词:遥感图像;跨模态检索;多维感知增强卷积;多尺度空洞卷积;自适应特征融合;多注意力机制
中图分类号:TP391.3;TP18;TN919.8 **文献标识码**:A **国家标准学科分类代码**:520.2040

Multi-scale remote sensing retrieval algorithm with global-local feature fusion

Zhou Chenhan Xu Xiaoyang Wei Wei
(School of Artificial Intelligence and Computer Science, Xi'an University of Science and Technology, Xi'an 710054, China)

Abstract: Aiming at the issues of interference from redundant information in images, insufficient multi-scale information extraction, and low retrieval accuracy caused by the ineffective integration of global and local information in cross-modal retrieval tasks for remote sensing images, this paper proposes a network model of multi-scale cross-modal remote sensing image retrieval (IGMR) suitable for multi-scale tasks. Firstly, a multi-dimensional perception enhanced convolution module (MFE) is designed to extract local information while filtering redundant features. It also integrates a multi-attention module to focus on the high-frequency information of images, thereby enhancing feature expression ability. Secondly, a multi-scale patch attention network (RFPA) is developed to capture contextual information at different scales. Subsequently, an adaptive feature fusion module (AFFM) is constructed to dynamically fuse the extracted global and local features, strengthening attention to high-quality information. Experimental results on the public datasets RSICD and RSITMD demonstrate that the proposed IGMR method increases the average recall rate (mR) by 1.83% and 3.21% respectively in remote sensing cross-modal retrieval tasks, with retrieval accuracies reaching 19.73% and 31.83%. The overall retrieval performance is significantly improved.

Keywords: remote sensing image; cross-modal retrieval; multidimensional perception enhanced convolution; multi-scale dilated convolution; adaptive feature fusion; multi-attention mechanism

0 引 言

自 20 世纪以来遥感卫星和传感器技术的飞速发展催生了海量遥感图像。系统、高效利用这些影像在环境评估[1]、军事侦察[2]、灾难预测[3]和资源勘探[4]等领域至关重要。无人机通过红外设备对战场局势进行侦察和记录成为了获取战场情报的关键工具。然而,遥感图像存在数量庞大、纹理模糊、信息价值低等问题,严重制约了战场高效决策与情报分析。因此,如何从大量图像中准确且快速检索与文本匹配的图像,或根据描述定位对应图像成为亟待解决的问题。

跨模态检索[5-6]作为多模态的重要分支,旨在实现不同

收稿日期:2025-06-19
* 基金项目:国家自然科学基金(12071367)、西安科技大学优秀青年基金(2024YQ2-08)项目资助

模态数据之间的关联检索。基于双塔匹配网络[7]的跨模态模型已成为当前主流结构。跨模态检索融合了计算机视觉[8]和自然语言处理[9]两个领域,通过联合建模视觉与语言模态的深层语义关联,实现跨模态语义对齐,从而达到异构数据间的精准匹配与检索。

传统的跨模态图文检索主要分为人工方法和跨模态特征学习。人工方法依赖领域专家进行图像语义标注与关键区域标定,但其受限于专家经验和先验知识。跨模态特征学习又分为基于子空间的方法[10]和基于主题模型的方法[11]。基于子空间的方法如典型相关分析[12](canonical correlation analysis,CCA)、偏最小二乘法[13](partial least squares,PLS)和双线性模型[14](bilinear model,BLM)等主要通过将异构模态映射至共享子空间以提升相似性,但难以根本解决模态异构问题。该类型方法将不同模态映射到同一主题空间以消除差异及提升语义一致性,但传统主题模型多为浅层,难以挖掘深层关联,限制了检索性能。

随着深度学习,特别是卷积神经网络[15]在图像处理中的广泛应用,跨模态图文检索算法的性能显著提升。基于深度学习的方法在检索速度、准确性、泛化性和鲁棒性方面均表现出优势。近年来,相关研究主要分为两类,第1类是基于图像描述的方法,通过生成式方法生成用以描述输入图像的关键词标签,再进行文本检索。第2类方法是基于文本-视觉双模态表征方法,使用独立编码器分别提取图像和文本特征,并映射到统一语义空间中进行匹配。

Krizhevsky等[16]提出的AlexNet证明了深度卷积神经网络在图像识别中的潜力,不仅推动图像分类的发展,也对语音识别、自然语言处理等领域产生深远影响。Faghri等[17]提出了新的视觉语义嵌入模型(visual semantic embeddings,VSE++)。它通过优化嵌入空间和困难负样本处理提升匹配效果。但其泛化性较差且对负样本选择依赖较强。Lee等[18]提出了堆叠交叉注意力模型(stacked cross attention,SCAN),通过堆叠交叉注意力实现图像区域与文本词之间的细粒度对齐,提高了匹配的可解释性,但对高质量标注依赖较大,在复杂图像内容上表现受限。Wang等[19]提出的跨模态自适应消息传递模型(cross-modal adaptive message passing,CAMP)采用最难负样本二元交叉熵损失提升训练效果,但在多尺度特征处理上仍存在不足。Yuan等[20]提出了非对称多模态特征匹配网络通过视觉特征引导文本生成,并提出新的三元组损失函数根据样本对的相似性调整边距。但无法过滤冗余目标而。张宏图等[21]通过图卷积网络整合高低阶邻居信息学习一致性表征,但难以捕捉所有模态间的复杂交互。Liu等[22]利用跨模态Transformer架构进行联合特征学习,支持多任务跨模态推理。Ma等[23]提出方向导向的视觉语义嵌入模型,利用区域视觉特征指导最终的视觉和文本表示。但对细节和特殊场景的处理不足。Yang等[24]提出隐式-显示关系推理方法减少语义混淆,也改善了细粒度的局部特征对齐和全局一致性问题。但掩码策略导致部分信息丢失。Chen等[25]提出上下文感知局部-全局语义对齐方法增强了复杂场景下的匹配能力,但对于遥感噪声处理不够充分。Zhang等[26]通过空间特征重构强化空间关系捕捉,但过渡关注通道特征而忽略了空间特征的精细提取。

上述方法在特定任务上表现优异,但仍存在局限。由于遥感图像中目标尺度大,现有方法难以同时捕捉局部细节与全局上下文信息,导致小目标或大范围表达不充分。且遥感图像背景复杂、信息冗余,上述方法的特征提取机制缺乏有效的过滤机制。尽管上述部分方法尝试融合多尺度特征,但是未能实现自适应特征融合,限制了模型对复杂场景的表达能力。受上述相关研究的启发,针对以上跨模态检索算法出现问题,同时考虑到遥感图像自身具有多尺度和小目标的特性,本文提出了一种多尺度跨模态遥感图像图文检索的网络模型(network model of multi-scale cross-modal remote sensing image retrieval,IGMR)。主要贡献包括:

1)通过设计多维感知增强卷积模块(multi-dimensional perception enhanced convolution module,MFE)来对跨模态检索模型中图像的局部特征进行提取和信息过滤,以获得更关键的局部信息。

2)提出多尺度分块注意力网络(multi-scale patch attention network,RFPA)模块通过多尺度空洞卷积捕捉不同层次的图像特征,解决在特征提取过程中细节或全局特征丢失的问题。

3)提出自适应特征融合模块(adaptive feature fusion module,AFFM)使局部和全局特征进行自适应地融合。这两个层面的信息相互补充,实现对高质量信息的专注。

4)在文本的处理反面,引入双向门控循环单元(bidirectional gate recurrent unit,Bi-GRU)[27]来对文本进行特征提取。将得到的图片特征和文本特征进行相似度计算,此外,结合多元重排序算法[28]对相似性结构进行二次排序,以进一步提升检索性能。

经过一系列实验验证,本文提出的IGMR模型在公开数据集RSICD和RSITMD上,平均召回率mR分别提高了1.83%、3.21%,检索精度达到了19.73%和31.83%。在遥感图像的图文检索方面的检索效果突出。

1 方法概述

1.1 整体网络结构

本文提出的跨模态检索模型采用双分支架构,如图1所示,由图像处理分支与文本编码分支组成,通过联合特征空间映射实现图文相似性度量。首先通过网络的图像处理分支进行特征提取和融合,得到图像特征$F^{(0)}$,其中卷积神经网络采用ResNet-18架构,其输入层接收3通道RGB图像,首层卷积核尺寸为7×7,步长为2;池化层卷积核大小为3×3。残差块内卷积核尺寸为3×3。输入通道数从3

开始，经过卷积层、池化层和 4 个残差块逐渐增至 512。最后经过全连接层将通道数映射到 1 000。文本由 Bi-GRU 分支获取对应的文本特征 $T^{(0)}$。每条文本语句对应一个词级特征集合，该集合包含每一个单词所对应的单词特征。将图像特征 $F^{(0)}$ 和文本特征 $T^{(0)}$ 进行融合后计算其相似度，再结合双向检索优化相似性矩阵的排名信息。

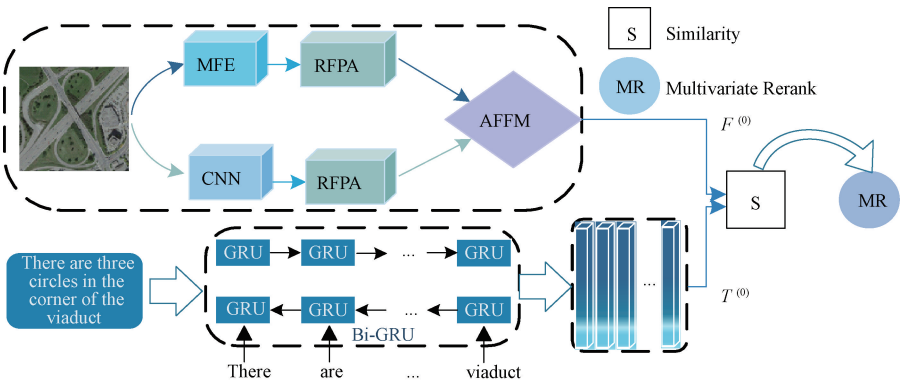


图 1　总体网络结构

Fig. 1　Overall network structure

1.2　图像特征提取

1）多维感知增强卷积模块

传统的局部特征提取方法通常使用图卷积建模节点间拓扑关系[29]，但图卷积本身存在对局部目标关注度不足而导致部分重要信息丢失，以及无法对遥感图像中的冗余信息进行充分过滤得问题。为解决这些问题，本文提出残差图卷积与串联的通道-空间注意力机制的多维感知增强卷积模块 MFE，该模块通过改进图卷积神经网络的结构设计，来捕捉图结构中各个节点与其邻居节点之复杂的连接关系，利用各边权重与邻居节点的特征更新自身节点，实现了对图结构中节点间复杂关系的高效建模。在每一层之间引入残差连接，增强特征信息之间的流动性，保证了最终的每个节点至少保留初始输入信息。改进图卷积通过权重修正有效解决了模型加深网络带来的过平滑问题。同时，残差连接能够解决梯度消失和信息瓶颈问题，保障信息的有效传递和梯度的顺畅反向传播。

多维感知卷积模块是由 3 层的隐藏层和多注意力机制组成。每层对节点进行卷积操作，每个节点的特征向量与其邻居节点的特征向量进行聚合和组合，实现信息处理。在每一层之间均添加残差连接，使得网络可以学习到恒等映射来保留重要信息，从而保持网络的稳定性。在每一层之间均使用 PReLU 激活函数进行激活。这可以让神经元在负输入时仍然进行更新，避免使用 ReLU 激活函数造成网络的神经元“死亡”，同时学习到更加灵活的非线性映射。通过三层改进图卷积能够在不同的目标范围内进行加权，帮助模型聚焦于图片局部的关键信息，忽略冗余特征，减轻信息丢失。在输出层融合多注意力机制模块（fusion of multi-attention mechanism module，FMA），使模块能够强化关键特征通道和空间区域的权重分配，实现了对冗余信息的过滤。MFE 模块结构如图 2 所示。

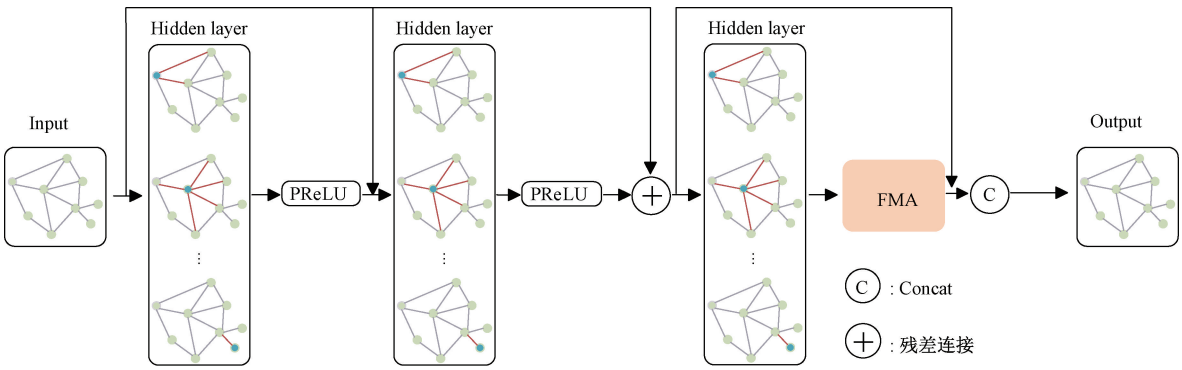


图 2　多维感知增强卷积模块

Fig. 2　Multidimensional perception enhanced convolution module

首先，将模型初始输入的特征矩阵 $\boldsymbol{M}^{(0)}$ 和经过图卷积操作后的特征矩阵 $\boldsymbol{M}^{(l)}$ 进行加权融合。并且通过预设的权重系数 α 来控制映射来对图像进行特征提取，以保留关键信息。三层改进图卷积网络模型定义如式(1)所示。

$$\boldsymbol{G}^{(l)}=\sigma(\tilde{\boldsymbol{D}}\boldsymbol{G}^{(l-1)}\boldsymbol{E}^{(l)}+\alpha_l\boldsymbol{I})+\boldsymbol{G}^{(l-1)} \tag{1}$$

其中，$l=1,2,3$ 代表图卷积的层数，$\alpha_l=0.1$ 为确定

的超参数，$\tilde{\boldsymbol{D}}$ 是经过标准化的邻接矩阵，$\boldsymbol{E}^{(l)}$ 是第 l 层图卷积的权重矩阵，$\boldsymbol{G}^{(l)}$ 为第 l 层图卷积的输入，σ 是 PReLU 激活函数。

为了进一步强化局部信息和全局信息互补作用，需要通过空间和通道充分突出关键特征以消除冗余特征对有效信息提取的影响。因此，在经过三层改进图卷积处理得到特征图后引入融合多注意力机制模块 FMA。其执行流程如图 3 所示。具体过程可以表示为：初始特征 $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$ 特征图经过全局平均池化和全局最大池化操作对整合输入特征图的空间信息，生成两个不同的空间上下文描述符 $\boldsymbol{s} \in \mathbb{R}^{C\times 1\times 1}$ 来分别表示各个通道。之后经过卷积层和激活函数 ReLU 学习关键的特征。最后通过 sigmoid 激活函数归一化权重至[0,1]，生成通道注意力图 $\boldsymbol{F}_c$ 计算公式为式(2)。

$$\boldsymbol{F}_c(\boldsymbol{F}) = \delta(\boldsymbol{W}_1\sigma(\boldsymbol{W}_0\boldsymbol{s})) \tag{2}$$

其中，δ 为 sigmoid 非线性激活函数，σ 为 ReLU 非线性激活函数，$\boldsymbol{W}_1$，$\boldsymbol{W}_0$ 为权重。随后将得到的特征权重 $\boldsymbol{F}_c$ 与初始特征图 $\boldsymbol{F}$ 执行逐个元素的乘法，生成通道注意力图 $\boldsymbol{F}_1$，计算公式如式(3)所示。

$$\boldsymbol{F}_1 = \boldsymbol{F}_c(\boldsymbol{F}) \times \boldsymbol{F} \tag{3}$$

通道注意力对各通道赋予不同的权重来区分各通道的重要程度，在降低学习成本的情况下使模型在通道上可以区分特征的重要性。在通道注意力的基础上加入空间注意力，在各个通道上提供重点关注区域，对通道注意力进行补充。通过 FMA 来捕捉丰富的特征信息，对重要的信息进行学习的同时对冗余信息进行过滤。

将通道注意力输出的特征图作为空间注意力输入的特征 $\boldsymbol{F}_1$，接着对输入特征图进行基于通道的全局最大池化和全局平均池化操作，得到两个 $\boldsymbol{f}_{\text{avg}} \in \mathbb{R}^{H\times W\times 1}$ 和 $\boldsymbol{f}_{\max} \in \mathbb{R}^{H\times W\times 1}$ 的特征图。将这两个特征图结合形成有效的特征描述符，再经过 7×7 卷积核进行卷积运算融合特征，经过 sigmoid 操作生成空间注意力特征 $\boldsymbol{F}_s$。空间注意力计算公式如式(4)所示。

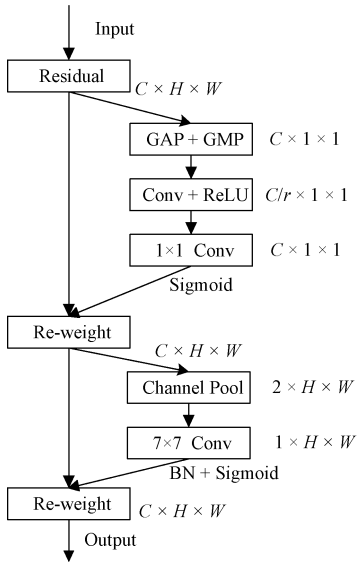


图 3 融合多注意力机制模块 FMA

Fig. 3 Fusion of multi-attention mechanism module FMA

$$\boldsymbol{F}_s(\boldsymbol{F}_1) = \delta(f^{7\times 7}([\boldsymbol{f}_{\text{avg}}, \boldsymbol{f}_{\max}])) \tag{4}$$

其中，$f^{7\times 7}$ 表示卷积核为 7×7 的卷积运算，δ 为 sigmoid 激活函数。最后将 $\boldsymbol{F}_s(\boldsymbol{F}_1)$ 与输入特征 $\boldsymbol{F}_1$ 相乘得到注意力融合模块的最终输出 $\boldsymbol{F}''$，计算公式如式(5)所示。

$$\boldsymbol{F}'' = \boldsymbol{F}_s(\boldsymbol{F}_1) \times \boldsymbol{F}_1 \tag{5}$$

2)多尺度分块注意力模块

低层次特征虽包含丰富的细节信息，但对高层次语义的表达能力较弱；高层次特征虽有较强的语义表达能力，但缺乏空间细节特征。针对遥感图像多尺度表达的复杂性以及局部细节与全局语义关联复杂的问题，本文构建了多尺度分块注意力模块 RFPA，如图 4 所示。

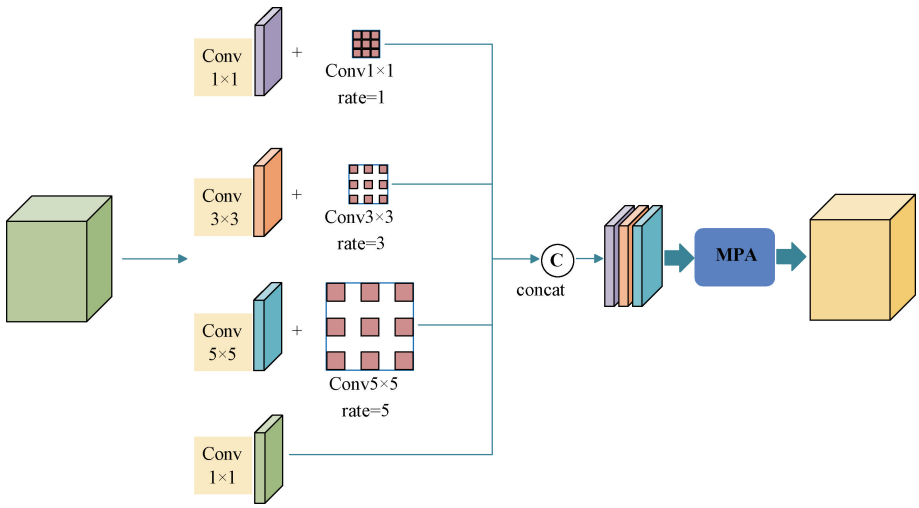


图 4 RFPA 模块

Fig. 4 RFPA module

RFPA 模块在使用不同大小卷积核的基础上采用不同的空洞率，在不改变图像的基础上增大感受野，因此可以通过调整不同的空洞率来实现对遥感图像多尺度的特征提取。首先将输入特征经过 4 个分支进行特征提取，分别是 3 个不同卷积核空洞卷积和一个恒等映射分支。其中，3 个不同卷积核的空洞卷积分支分别使用不同的空洞率来提取不同尺度的感受野的特征信息。不同空洞率的卷积分支用于提取全局信息来提升全局视野。1×1 卷积直接传递原始特征的恒等映射分支，它直接传递原始特征以避免信息丢失。再将得到的特征信息进行拼接合并操作，使模型能够选择最相关的特征用与检索任务。然后将合并得到的特征通过多路径分块注意力（multi-path block attention，MPA）增强对目标区域的关注度。

由于空洞率为 1 可以保持原始的局部感受野大小，适合捕捉小尺度局部特征；空洞率为 3 能够增加感受野，捕捉中等尺度信息，能够提供上下文信息；进一步扩大感受野，使用空洞率为 5，能够有效捕捉大范围信息，提升对大尺度信息的感知能力，同时仍保持较高的计算效率。因此本文使用空洞率为 1、3、5 的空洞卷积并联组成空洞卷积金字塔模块。将 MFE 和卷积神经网络得到的特征图分别输入到多尺度空洞卷积金字塔 RFPA 中，经过不同的空洞率的卷积核可以获得具有不同感受野的特征图，从而得到多尺度信息。为了使提取到的多尺度信息更加有效，将特征图通过多路径分块注意力赋予不同的权重，捕获多尺度中最有效的检索目标特征信息。

将图像和文本映射到公共的特征空间进行匹配过程中，考虑到描述图像的文本出现同时涉及局部目标和全局场景的情况，因此需要从图像中提取不同尺度的信息以便对齐与文本描述相符的特征。不同于其他方法，本文方法在获取不同感受野特征图的信息时，有不同的感受野还会关注中心区域的重要性。在将不同尺度特征进行融合时，本文将多尺度特征沿着通道维度进行联级操作，从而实现多尺度特征融合。该过程可以表示为式(6)。

$$\boldsymbol{F}' = W' \otimes [\boldsymbol{O}_1, \boldsymbol{O}_2, \boldsymbol{O}_3, \boldsymbol{O}_4] \tag{6}$$

其中，$\boldsymbol{O}_i$ 表示第 i 个卷积后的输出，$[\boldsymbol{O}_1, \boldsymbol{O}_2, \boldsymbol{O}_3, \boldsymbol{O}_4]$ 表示各个输出之间沿着通道维度进行拼接操作。

多路径分块注意力如图 5 所示。该模块通过上下两条 PatchAware 分支进行多尺度补丁和全局特征互补，能够捕捉到更丰富的上下文信息，特别是在复杂的遥感图像中。MPA 模块通过优化局部特征的加权方式，有效地增强细节信息地区分度。中间分支进行 3 次连续卷积来处理局部特征，能够逐步提取更深层次的特征信息，有助于捕捉小物体的细节，同时避免了直接使用大卷积核带来的计算负担和信息丢失。每一层卷积都在不同尺度上捕捉局部信息，得到更丰富的特征表示。

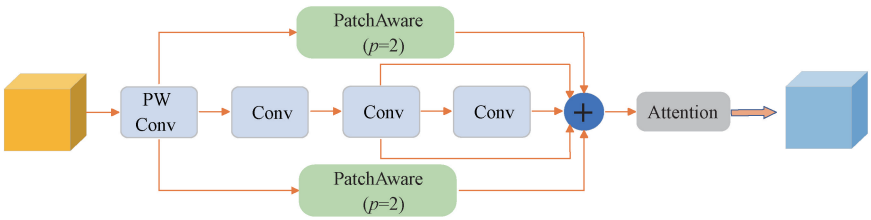


图 5　多路径分块注意力

Fig. 5　Multi-path block attention

将不同感受野在通道维度进行拼接后的特征图作为初始特征 $\boldsymbol{F} \in \mathbb{R}^{H\times W\times C'}$，将 $\boldsymbol{F} \in \mathbb{R}^{H\times W\times C'}$ 通过点卷积调整输入特征维度为 $\boldsymbol{F} \in \mathbb{R}^{H\times W\times C}$。然后通过 3 条并行分支进行处理。通过上下两个 PatchAware 分支将特征 $\boldsymbol{F} \in \mathbb{R}^{H\times W\times C}$ 分割为 2×2 的非重叠块，每个分块为 $P^{4,H/2,W/2,C}$，并对每个分块通过前馈神经网络执行通道均值计算来捕捉不同尺度的局部信息。最后，生成压缩后的特征表示 $P^{4,H/2,W/2}$。通过这种方式模型不仅能够处理图像的全局信息，还能深入分析局部区域的细节。然后使用线性变换和 softmax 激活函数计算每个分块在空间维度上的注意力权重。再基于这些权重对每分块通过元素级乘法进行重新加权，生成权重化的特征表示，如式(7)所示。

$$\begin{aligned} \boldsymbol{F}'_1 &= \sigma(MPL(LN(MLP(\boldsymbol{P}^{4,H/2,W/2})))) \\ \boldsymbol{F}' &= \boldsymbol{F}'_1 \odot \boldsymbol{P}^{4,H/2,W/2} \end{aligned} \tag{7}$$

其中，σ 为 softmax 激活函数，$\odot$ 为逐元素相乘。

之后使用特定的任务嵌入 $\boldsymbol{\xi} \in \mathbb{R}^C$ 和余弦相似度函数 $sim(\cdot,\cdot)$ 选择最相关的特征，每个 Token $\boldsymbol{t}_i$ 重新加权过程如式(8)所示。

$$\hat{\boldsymbol{t}}_i = \boldsymbol{L} \cdot sim(t_i, \boldsymbol{\xi}) \cdot \boldsymbol{t}_i \tag{8}$$

其中，$\boldsymbol{L}$ 是通道选择的线性变换矩阵。最后对进行不同路径的分块融合和特征提取后的特征 $\tilde{\boldsymbol{F}}$ 进行通道和空间注意力加权，再经过激活函数和批归一化操作生成最终输出 $\boldsymbol{F}'' \in \mathbb{R}^{H\times W\times C}$。多路径分块注意力模块通过并行路径实现特征的分块加权处理，不同局部补丁根据在图中的重要性被赋予不同权重，根据权重大小来关注检索过程中最有用的特征。在目标相似度高的场景中做出更精确地区分，对多尺度提取到的信息进行有效性聚焦和整合，从而可以得到更加准确的检索结果。

3）自适应特征融合模块

为了有效融合全局和局部特征，本文提出了一种自适

应特征融和模块 AFFM,如图 6 所示。遥感图像中目标具有多样化的特点。与自然图像相比,遥感图像包含更丰富的语义信息。然而,随着卷积网络数量的增加,卷积网络通过多次下采样扩大感受野,以捕捉全局上下文信息。但经过多次的下采样会导致小目标空间特征被严重稀释。且多次卷积和激活函数叠加会导致过平滑现象,这会削弱目标区分度,导致目标特征被背景噪声淹没,最终表现为在遥感图像中的目标信息逐渐减少甚至消失,因此局部信息和全局信息的融合仍然是亟待解决的问题。为了解决上述问题,本文提出了 AFFM 使局部特征和全局信息进行特征融合。这两个层面的信息需要相互补充,通常需要使用局部信息来修正包含冗余信息的全局表示,以及利用丰富的全局特征作补充局部信息。并且自适应地采用密集分支和稀疏分支实现特征动态筛选,达到对高质量信息的专注。

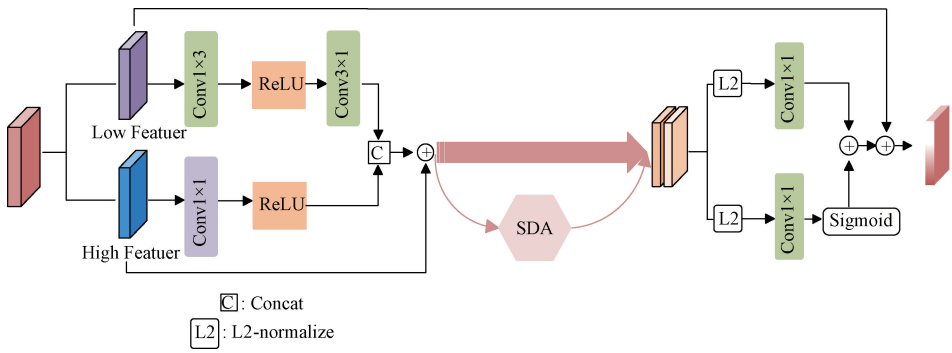


图 6 自适应特征融合模块 AFFM

Fig. 6 Adaptive feature fusion module AFFM

图像特征提取网络所获得的多尺度特征维度不统一,因此对全局和局部特征分别采用 3×3 卷积和 1×1 卷积,其步长分别为 4、1。为了降低计算量与参数量,此处将 3×3 卷积拆分为串联的 1×3 和 3×1 的卷积,其步长变为(1,4)、(4,1)。为了增强网络的非线性表达能力,降维后的特征进行修正线性单元处理,以增强网络的非线性表达能力。最终将得到的信息进行拼接得到初步的融合特征 V_f。为了专注于高质量信息,将特征图通过自适应的稀疏和密集分支进行筛选,从而捕获高质量特征。将捕获的高质量特征分别进行不同路径的 1×1 卷积,对特征进行精细化调整并融合,进一步强化重要特征的学习。使用残差连接再分别将原始局部特征直接传递至网络末尾进行相加操作,将原始的全局特征进行直接传递到中部进行相加,为网络提供更多的信息来源。

为了关注重要特征而过滤冗余特征,本文采用在初步的融合特征 V_f 之后采用自适应稀疏-密集注意力模块(adaptive sparse-dense attention module,SDA),如图 7 所示。采用平方激活函数的自注意来过滤低查询键匹配分数的负面影响特征。同时引入密集分支,采用 softmax 层来帮助保留关键信息。采用自适应方式融合两个分支信息。给定一个归一化特征映射 $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$,将其划分为大小为 $M\times M$ 的非重叠窗口,从第 k 个窗口得到平坦表示 $\boldsymbol{X}^k \in \mathbb{R}^{M^2\times C}$。然后从 $\boldsymbol{X}$ 生成查询矩阵 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$,其公式如式(9)所示。

$$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}_Q,\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}_K,\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}_V \tag{9}$$

查询 $\boldsymbol{W}_Q$、键 $\boldsymbol{W}_K$ 和值 $\boldsymbol{W}_V \in \mathbb{R}^{C\times d}$ 的线性投影矩阵在所在窗口之间共享,注意力计算如式(10)所示。

$$\boldsymbol{A}=f(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d}+\boldsymbol{B})\boldsymbol{V} \tag{10}$$

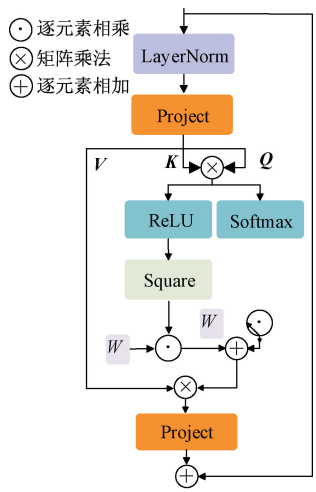


图 7 SDA 注意力

Fig. 7 SDA attention

其中,$\boldsymbol{A}$ 表示估计的注意力,$\boldsymbol{B}$ 表示可学习的相对位置偏差,$f(\cdot)$ 是一个评分函数。是对查询和键地维度 d 进行缩放,避免数值不稳定。在此过程中,并行计算了不同头部的权重,不同的头部被连接后通过线性投射融合。然后采用密集型注意机制通过 softmax 层,考虑所有查询关键字对来获得注意力得分,如式(11)所示。

$$DSA=Softmax(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d}+\boldsymbol{B}) \tag{11}$$

而稀疏自注意力机制选择有用的交互来增强特征聚合。使用基于平方激活函数的层来实现注意力的稀疏性,它消除了与负分数的相似性,并向前传播最有用的信息流,如式(12)所示。

$$SSA=\mathrm{ReLU}^2(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d}+\boldsymbol{B}) \tag{12}$$

单独采用稀疏自注意力分支学习到不充分信息不足以用于后续过程，而单独采用密集自注意力分支会在不相关区域引入噪声。因此采用双分支来自适应的使获得的特征信息进行过滤和学习。在双分支结合后可更新为式(13)。

$$\boldsymbol{A}=(w_1\times \boldsymbol{SSA}+w_2\boldsymbol{DSA})\boldsymbol{V} \tag{13}$$

其中，$w_1,w_2\in\mathbb{R}^1$ 是自适应调制两个分支归一化权重。其可通过式(14)计算得到，式中 a_i 是可学习参数。SDA 通过双分支来过滤掉不相关区域的噪声交互与可利用的信息特征之间更好的权衡。

$$w_i=\frac{e^{a_i}}{\sum_{n=1}^{N}e^{a_n}},\ i=\{1,2\} \tag{14}$$

1.3　文本特征表示

文本特征与图片特征同等重要，为实现跨模态特征对齐，本文引入 Bi-GRU，它由前向 GRU 和反向 GRU 组成[30]，分别从向前、向后两个方向提取句子和单词级别的特征，获取句子的上下文相关信息。其具体过程如式(15)、(16)所示。

$$\overrightarrow{\boldsymbol{h}_t}=\overrightarrow{f_{GRU}}(\boldsymbol{e}_t,\overrightarrow{\boldsymbol{h}_{t-1}}) \tag{15}$$

$$\overleftarrow{\boldsymbol{h}_t}=\overleftarrow{f_{GRU}}(\boldsymbol{e}_t,\overleftarrow{\boldsymbol{h}_{t-1}}) \tag{16}$$

式中：$\overrightarrow{\boldsymbol{h}_i}$ 和 $\overleftarrow{\boldsymbol{h}_i}$ 分别表示正向 GRU 和反向 GRU 的隐藏状态。最终，将双向 GRU 的结果输入到支持向量机生成句子级特征向量，如式(17)所示。

$$\boldsymbol{s}=MLP(\frac{1}{n}\sum_{i=1}^{n}\frac{\overrightarrow{\boldsymbol{h}_i}+\overleftarrow{\boldsymbol{h}_i}}{2}) \tag{17}$$

其中，$\boldsymbol{s}$ 即为文本的句子级特征。

1.4　损失函数

本文采用多模态特征匹配领域的三元组损失 Triplet Loss[31] 损失函数，如式(18)所示。

$$L(\boldsymbol{I},\boldsymbol{T})=\sum\dot{\boldsymbol{T}}[\varepsilon-\cos(\boldsymbol{I},\boldsymbol{T})+\cos(\boldsymbol{I},\dot{\boldsymbol{T}})]^{+}+\sum\dot{\boldsymbol{I}}[\varepsilon-\cos(\boldsymbol{I},\boldsymbol{T})+\cos(\dot{\boldsymbol{I}},\boldsymbol{T})]^{+} \tag{18}$$

其中，$\boldsymbol{I}$ 表示图像，$\boldsymbol{T}$ 表示文本，$\dot{\boldsymbol{T}}$ 表示与图像 $\boldsymbol{I}$ 不匹配的文本，$\dot{\boldsymbol{I}}$ 表示与文本 $\boldsymbol{T}$ 不匹配的图像，$\cos(x,y)$ 表示向量 x 和 y 之间的余弦相似度，ε 为正样本与负样本之间的最小距离，$[x]^{+}$ 表示 ReLU 激活函数且确保损失值大于 0。

该损失函数的优化目标是通过调整图像和文本的嵌入向量，使得匹配的图像与文本之间的相似度更高，即它们之间的余弦相似度最大。而不匹配的图像与文本之间的相似度小于正样本对的相似度，并且与正样本对的相似度差至少为 ε。Triplet Loss 通过最小化正样本对和负样本对之间的相似度差来优化模型的表示，使得跨模态检索任务中的图像与文本的匹配更加精确。

2　实验结果与分析

2.1　数据集

为验证本文方法有效性并评估本文模型的性能，本文在两个公开的遥感数据集 RSICD[32] 和 RSITMD[33] 上分别进行了实验。RSICD 数据集共有 31 个不同的类别，一共包含 10 921 张遥感图像。每张图像大小均为 224 pixel×224 pixel。每张遥感图像都对应 5 个不同的描述语句。RSITMD 数据集部分图像源自 RSICD 数据集，其余图像来源于 Goole Earth，总共包含 4 743 张图片，以及 23 715 个相应的文本描述。本文按照 8∶1∶1的原则[34]来将两个数据集划分训练集、测试集和验证集，进而展开模型的训练与评估。

2.2　实验环境及实验参数

本文基于 Pytorch 深度学习框架构造网络模型，处理器采用 Inter(R)Xeon(R)W-2150B CPU@3.00 GHz，实验环境为 Python3.8，CUDA12.2。模型训练和测试使用 Nvidia GeForce RTX3090Ti 24 GB 显卡。

本文设置 Bi-GRU 的隐藏节点数设置为 512，词向量维度设置为 300，图像和文本嵌入空间均为 512。Triplet Loss 边缘阈值 δ 限制为 0.2。使用带有三元组损失的 Adam 优化器[35]来训练网络 70 个 epoch，批次大小为 100，初始学习率 ε 为 1×10^{-4}，每 20 个 epoch 衰减 0.7。这种策略有助于训练过程中逐步提高模型性能，同时避免过拟合现象发生。

2.3　消融实验

本节设计了 4 组实验分别验证多维感知增强卷积模块、融合多尺度空洞卷积与多路径分块注意力模块和自适应特征融合模块的有效性。

1)IGMR 网络各模块有效性验证

为了验证 MFE 模块和 RFPA 模块以及 AFFM 模块的有效性，本节进行了 4 组对比实验，在 RSICD 和 RSITMD 数据集上进行。分别是不添加 3 种模块的 Base 组、只添加 MFE、只添加 MFE 和 RFPA、只添加 MFE 和 AFFM 以及同时添加 3 种模块 5 种情况，实验结果如表 1 和 2 所示。

由表 1 和 2 实验结果可见，IGMR 网络在 RSICD 和 RSITMD 数据集上的平均召回率 mR 均优于基础架构，分别提升了 1.83%和 3.21%，验证了各模块的有效性。3 个模块均加入网络时，在 RSICD 数据集上，文本检索图像和图像检索文本任务上指标均高于 Base 组。在 RSITMD 数据集上，文本检索图像任务上的检索结果均高于 Base 组；在图像检索文本任务中 R@1 也同样高于 Base 组。

在加入 RFPA 模块后，模型性能显著提升。对遥感图像进行不同尺度的特征提取，获取更多重要的特征信息。再通过多路径分块注意力机制对高相似度的场景中的特征做出精确区分，使多尺度提取到的信息进行有效性聚焦得到更加准确的检索结果，这表明多尺度特征提取对遥感图像是有效的。MFE 采用残差图卷积和融合多注意力模块，使局部图像信息充分被提取的同时过滤冗余信息，也实现了良好的效果。融合多注意力机制弥补了在提取图像高频信息时的不足。AFFM 模块通过对全局-局部特征

表1 不同模块在RSICD数据集上有效性验证结果

Table 1 Validation results of different modules on RSICD dataset %

对照项	模型设置			Image-to-Sentence			Sentence-to-Image			mR
	MFE	RFPA	AFFM	R@1	R@5	R@10	R@1	R@5	R@10	
Base				6.61	18.17	28.08	4.75	18.60	31.18	17.90
1	√			6.50	17.84	28.64	4.96	17.83	29.69	17.58
2	√	√		6.89	19.12	29.70	4.98	19.61	32.56	18.76
3	√		√	7.13	17.74	28.91	5.03	19.14	32.10	18.34
4	√	√	√	**7.56**	**19.81**	**30.28**	**5.18**	**20.96**	**34.59**	**19.73**

表2 不同模块在RSITMD数据集上有效性验证结果

Table 2 Validation results of different modules on RSITMD dataset %

对照项	模型设置			Image-to-Sentence			Sentence-to-Image			mR
	MFE	RFPA	AFFM	R@1	R@5	R@10	R@1	R@5	R@10	
Base				11.28	26.32	38.05	10.13	35.22	50.73	28.62
1	√			12.38	**30.75**	**42.92**	10.75	35.44	53.18	30.90
2	√	√		12.83	28.54	42.04	10.34	35.62	50.66	30.01
3	√		√	12.16	28.09	42.69	11.15	32.46	49.60	29.69
4	√	√	√	**13.15**	29.65	42.86	**11.37**	**42.32**	**51.65**	**31.83**

的提取和融合,以获取更全面的信息。然后通过自适应的稀疏-密集注意力模块来关注更重要的信息,将信息进行有效融合。本文模型在两个数据集上的召回率平均值均有明显上升,这表明结合MFE模块、RFPA模块和AFFM模块能获得更全面、更精确的遥感图像特征。再将特征进行更有效地融合,从而提升模型检索的精度,同时也证明了3个模块在本文模型中的有效性和重要性。根据表2的结果可知,在图像检索文本任务中,单独添加MFE模块时,R@5和R@10指标最优。由于MFE模块较为简单,能够更好地聚焦于图像关键特征,使图像匹配到的文本更精确。然而,仅添加MFE模块时网络会忽略图像非关键特征,对局部和全局融合不充分,在处理复杂的图像和文本时具有局限性。因此导致除上述指标之外,其余指标均为最低。当3个模块同时加入网络后,不仅能够关注重要特征,还对不同层次的特征进行有效聚焦,并充分融合,从而提升图像与文本的匹配能力。尤其在处理复杂数据集时,模型整体检索性能最佳。

2)MFE模块有效性验证

本节将MFE模块分成3组在RSICD数据集上进行实验验证,分别是改进的图卷积层数为1的MFE、改进的图卷积层数为4的MFE和改进的图卷积层数为3的MFE模块。实验结果如图8所示,其中图像检索文本任务的实验结果如图8(a)所示;文本检索图像的实验结果如图8(b)所示。由实验结果可知,在其余模块均不改变的情况下改进的图卷积为3层时性能最好。与改进的单层图卷积和改进的四层图卷积相比,无论在图像检索文本任务中还是文本检索图像任务中,本文构建的改进三层图卷积的MFE模块的召回率均为最优且平均召回率mR分别提升了1.13%和1.22%。因为采用单层图卷积的MFE模块仅能聚合直接邻居的信息感受野有限,难以捕捉图像中长距离的语义关联,会导致提取的局部特征无法与文本的全局描述有效匹配。使用四层图卷积的MFE模块,在网络加深后会出现过平滑的问题,导致丢失关键细节信息。在MFE模块中使用三层改进图卷积有较大的感受野的同时通过引入残差连接使每层图卷积能够保留初始特征,有效避免了深层网络的信息丢失和梯度消失,从而维持特征表达能力。且三层图卷积逐步聚合邻域信息,分层扩展感受野并结合PReLU激活函数逐步进行非线性变换,增强了特征层次与表达能力,有效避免出现特征的表达能力削弱的情况。同时通过权重修正来使MFE可以有效地聚焦于关键的局部细节。并且通过FMA注意力进一步增强了网络对通道和空间位置上高频细节信息的关注能力。由此可以证明,本文提出的MEF模块对遥感图像的局部关键信息提取更加充分,且对冗余信息进行充分过滤提高了跨模态检索的准确率。

2.4 对比实验

为了验证本文方法的有效性,本文选取VSE++、SCAN、CAMP、CMFM-Net[36]、AMFMN、SIRS[37]和SSJDN[38]跨模态图文检索方法与本文方法在不同数据集上进行对比。实验结果如表3和4所示。

根据表3和4的数据可以得出,本文所提得IGMR模型在RSICD和RSITMD两个遥感数据集上均取得了优越得性能。如表3所示,IGMR模型在图像检索文本任务中,R@1,R@5,R@10分别达到了7.56%、19.81%、30.28%。

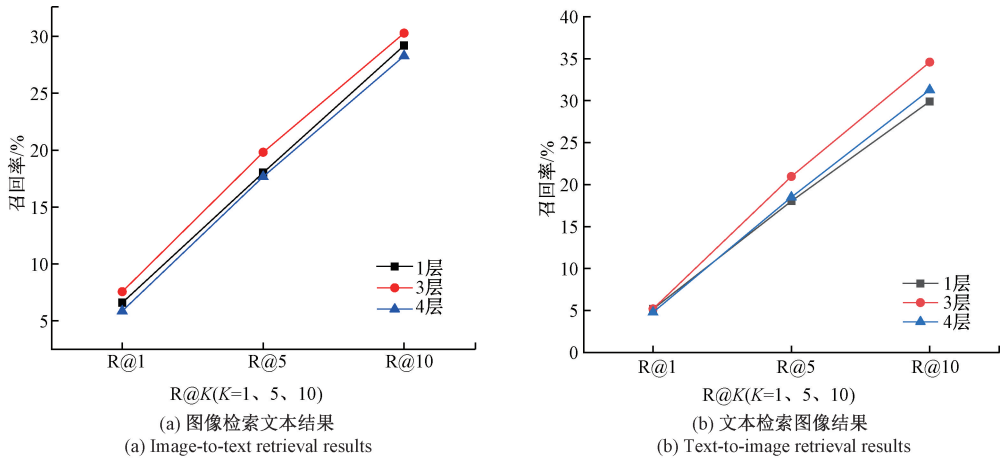


图 8　不同层数的 MFE 模块图文检索结果可视化

Fig. 8　Visualization of image-text retrieval results of MFE modules with different layers

表 3　RSICD 数据集对比实验结果

Table 3　RSICD data set comparison experimental results　%

方法	Image-to-Sentence			Sentence-to-Image			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++	4.56	16.73	22.94	4.37	15.37	25.35	14.89
SCANt2i	4.39	10.90	17.64	3.91	16.20	26.49	13.25
SCANi2t	5.85	12.89	19.84	3.71	16.40	26.73	14.23
CAMP	5.12	12.89	21.12	4.15	15.23	27.81	14.39
CMFM-Net	5.40	18.66	28.55	5.31	18.57	30.03	17.75
AMFMN-sim	**7.68**	17.01	27.72	4.08	19.35	31.07	17.82
AMFMN-soft	6.64	18.29	28.70	4.90	19.02	31.39	18.16
AMFMN-fusion	6.19	18.32	29.21	4.79	19.51	32.61	18.44
SSJDN	6.5	19.70	30.10	4.90	20.20	**36.50**	19.70
SIRS	7.21	19.26	29.10	**5.37**	19.52	32.13	18.76
Ours	7.56	**19.81**	**30.28**	5.18	**20.96**	34.59	**19.73**

表 4　RSITMD 数据集对比实验结果

Table 4　RSITMD data set comparison experimental results　%

方法	Image-to-Sentence			Sentence-to-Image			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++	9.07	21.61	31.78	7.73	27.80	41.00	23.17
SCANt2i	10.18	28.53	38.49	10.10	28.98	43.53	26.64
SCANi2t	11.06	25.88	39.38	9.82	29.38	42.12	26.28
CAMP	11.73	26.99	38.05	8.27	27.79	44.34	26.20
CMFM-Net	10.84	28.76	40.04	10.00	32.83	47.21	28.28
AMFMN-sim	12.17	26.33	38.50	9.38	33.10	49.12	28.10
AMFMN-soft	11.95	27.88	40.71	10.04	32.43	51.33	29.06
AMFMN-fusion	11.73	26.32	40.49	9.20	32.39	49.03	28.19
SSJDN	12.20	29.40	**44.20**	10.80	42.20	**68.90**	**34.60**
SIRS	13.11	29.35	41.32	10.59	35.31	51.28	30.16
Ours	**13.15**	**29.65**	42.86	**11.37**	**42.32**	51.65	31.83

在多个指标上超越当前主流模型,表现出较强的图文匹配能力;在文本检索图像任务中其性能亦处于较领先水平,平均检索精度 mR 达到 19.73%,同样优于现有的多数方法。这一结果表明,IGMR 方法无论对于局部信息还是全局信息都实现充分的提取。由于遥感图像自身存在较多的冗余特征,该方法还能够充分过滤不重要的信息而专注于重要特征,并将这些重要特征进行有效融合。

此外,通过 Bi-GRU 对文本上下文信息进行建模,从而提升了图文特征之间得对齐精度。尽管 AMFMN-soft 模型采用视觉引导注意力机制以获得多样化的特征表达;SIRS 模型专注于图像中与文本描述的关键实体更为一致的重要区域;SSJDN 模型通过解耦结构更好地捕捉图像细粒度特征,因此这些模型在某些指标上检索效果突出。但是上述 3 个模型没有将提取到的特征与其他特征进行有效融合,对非重点区域图像信息关注度低,从而导致模型仅在某个指标最优无法做到整体检索效果最好。而本文算法将提取到的重要及非重要的图像特征进行充分融合,同时也注重两种模态间的有效融合使模型能得到更多的图像信息,在与文本特征匹配时更有优势,模型整体检索结果更准确。

从表 4 可看出,IGMR 模型在 RSITMD 数据集上的图像检索文本以及文本检索图像的子任务上 R@1、R@5 均达到最优,分别达到 13.15%、29.65%、11.37%、42.32%。mR 的值为 31.83%,为次优结果。IGMR 模型通过 MFE 模块对局部关键信息进行关注并有效地过滤冗余特征,其中融合多注意力机制 FMA 提升了网络捕获高频细节的能力。然后,通过 RFPA 模型进行多尺度特征提取,并通过多路径分块注意力获取不同尺度下最有效的信息使网络获取更加丰富的特征信息,从而有效解决了传统网络提取过程中的信息丢失问题。最后 IGMR 通过 AFFM 模块自适应地融和全局和局部特征信息,并利用密集分支和稀疏分支实现动态选择。该方法充分考虑到全局特征和局部特征的依赖关系,实现了全局和局部特征相互指导。同时,模型还对文本特征进行提取,以实现图像和文本的准确匹配。但是,在图像检索文本以及文本检索图像任务中 R@10 和 mR 上略低于 SSJDN 算法,可能是由于返回结果中包含了较多冗余信息,遥感图像中存在多种地物与复杂描述,导致无法很好区分相关信息。虽然检索效果有一定的损失,但是通过其他返回结果可以发现,检索结果整体上仍为最优。

以上实验结果表明,本文方法 IGMR 在对齐细粒度特征的同时提升了全局语义理解能力,显著缩短了查询样本与目标样本之间的语义距离,并且表现出更强的鲁棒性和泛化性。

2.5 边缘阈值影响分析

为探究 Triplet Loss 不同边界值对模型性能的影响,本节基于 RSICD 数据集进行实验。实验过程中仅通过修改三元组损失函数的边缘阈值 δ 来测试其对 mR 指标的影响,实验结果如图 9 所示。

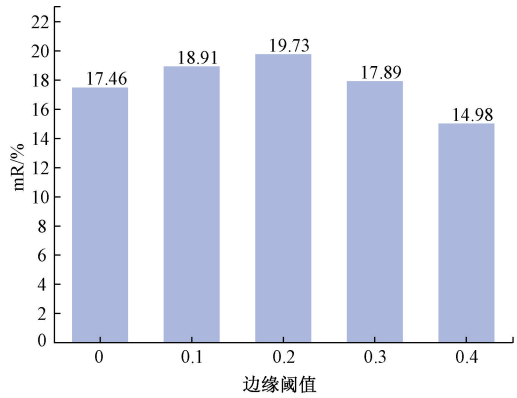


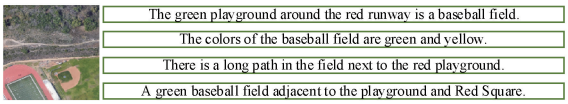
图 9 不同边缘阈值在 mR 指标上的结果

Fig. 9 The results of different edge thresholds in mR

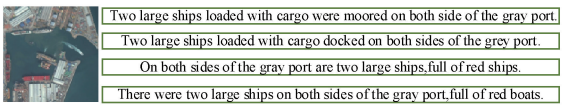
在跨模态检索任务中,三元组损失用于衡量输入样本之间的相对相似性。边缘阈值在此机制中起强化作用,通过确保正样本和锚点之间的距离尽可能缩小的同时负样本与锚点之间的距离至少维持在一个界定的阈值上,从而增强模型学习能力。根据实验结果分析,随着界值增加,本文提出的 IGMR 模型的学习性能呈现先升后降的趋势。当 $\delta=0.2$ 时,模型达到最佳性能。当其继续增大,学习性能呈下降趋势。因此本文将损失函数边界阈值设为 0.2,以此优化模型的整体性能。

2.6 可视化展示

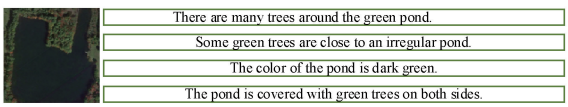
经过以上实验,得到如图 10 和 11 所示的可视化结果。图 10 从上至下为检索结果排名,图 11 从左至右为检索结果排名。图 10(a)~(c)所示,均展示了图像检索文本的示例。该任务使用本文的 IGMR 模型,对待检索文本与所查询的遥感图像进行相似性度量,选取相似度最高的 4 个文本作为检索结果。本文模型得到的文本均与原遥感图像的文本描述一致。因为 IGMR 在对图片处理时通过 MFE 进行全局和局部特征的充分提取并且实现对冗余信息的有效过滤,再利用 FMA 注意力机制对空间和通道信息进行特别关注来聚焦重要特征。然后,RPFA 模块在提取不同尺度的特征时,通过多路径注意力机制分别关注各尺度特征中的关键信息,高效整合了多尺度信息。最后,AFFM 模块将这些特征信息充分融合,并与文本特征进行相似度计算得到的检索结果。如图 11(a)~(c)所示,展示了使用 IGMR 模型进行文本检索图像的可视化结果,同样选取相似度最高的 4 个图像作为检索的最终结果。可视化结果充分证明了本文模型的有效性。但是分析示例 3 检索结果可知,可能是由于数据集中的遥感图像过于相似,且文本特征更加复杂,无法获得更加准确的文本上下文信息,因此也无法将图片特征和文本特征进行有效的对齐,从而导致检索结果出现误差。



(a) 示例1
(a) Example 1



(b) 示例2
(b) Example 2



(c) 示例3
(c) Example 3

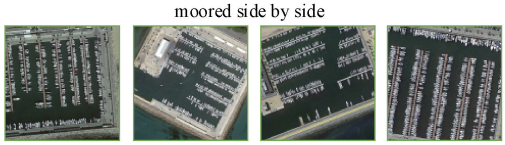
图 10　图像检索文本示例

Fig. 10　Image retrieval text example

输入语句：There are some vertical texture on the meadow

(a) 示例1
(a) Example 1

输入语句：We can see a square harbour basin on the coast with boats moored side by side



(b) 示例2
(b) Example 2

输入语句：A baseball field is surrounded by many green trees and several buildings

(c) 示例3
(c) Example 3

图 11　文本检索图像示例

Fig. 11　Text retrieval image example

2.7　鲁棒性实验

1)动态响应测试

学习率(learning rate,lr)是控制模型参数更新速度的延时配置。较大的学习率意味着系统对梯度信号的响应更快,但可能导致振荡;较小的学习率意味着响应更平缓、更稳定,但收敛速度慢。动态调整学习率就是在动态地重新配置系统的响应特性。针对不同学习率在 RSICD 数据集上来测试模型的性能,且学习率每 20 个 epoch 衰减 0.7 来动态测试模型检索性能。其实验结果如表 5 所示。由测试结果可知,当初始 lr=0.000 2 时,平均检索精度最高。也证明了本文模型选择在 lr 为 0.000 2 训练过程中学习率的动态衰减的条件下,平均检索结果仍较为出色。

表 5　不同学习率下的平均检索精度

Table 5　Average retrieval accuracy under different learning rates

lr	mR/%
0.000 1	19.11
0.000 2	**19.73**
0.000 3	18.61

2)不同温度下算法鲁棒性验证

为进一步验证模型的鲁棒性,本文在不同环境温度条件下对进行了温度控制的模拟测试,其中 25℃ 为基础温度。评估延时控制系统的精度变化。实验过程中,保持模型参数和数据集不变,仅通过外部温控设备调节环境温度。实验结果如表 6 所示。由实验结果可知,在不同温度环境下 IGMR 模型的检索精度变化幅度均小于 1%。表明本文模型具有良好的温度适应性和系统稳定性,能够在较宽的温度范围内保持较高的检索性能。

表 6　不同温度下的平均检索精度变化

Table 6　Retrieval accuracy under different temperatures

数据集	温度/℃	mR/%
RSICD	10	19.45
	25	**19.73**
	35	19.42
RSITMD	10	31.44
	25	**31.83**
	35	31.17

3　结　　论

本文提出了一种适用多尺度任务的遥感图像跨模态检索模型 IGMR。此模型通过 MFE 模块充分提取遥感图像的局部特征,有效解决特征提取过程中信息丢失问题,并减少冗余信息。同时,通过多注意力 FMA 关注通道和空间上的细节特征,显著提升了 MFE 模块的细节特征提取能力。RFPA 模块对图片进行多尺度的特征提取,得到更丰富的图像信息,通过多路径分块注意力对场景做出区分,将得到的丰富的图像信息进行有效聚焦。AFFM 模块进一步有效融合不同尺度的图像特征,并采用自适应稀疏-密集注意力来突出遥感图像中的重要目标信息。此外,本文模型还采用 Bi-GRU 提取文本特征,实现图像和文本的准确匹配。本文在两个公开数据集上进行对比试验和消融实验,实验结果表明本文方法在遥感图像跨模态检索任务上达到了最佳性能。然而,值得注意的是,本文方法尚未充分考虑图像和文本之间的语义关系,以及文本中单词的

不同含义导致的检索精度降低问题。在未来工作中,将着重解决单词语义歧义问题,以进一步提升检索的准确性。

参考文献

[1] LIU Y J, LI X F, REN Y B. A deep learning model for oceanic mesoscale eddy detection based on multi-source remote sensing imagery[C]. IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2020: 6762-6765.

[2] SHI T CH, LIU M Y, NIU Y, et al. Detection of small ships in remote sensing images based on deep convolutional neural network[C]. Global Oceans 2020: Singapore-US Gulf Coast. IEEE, 2020: 1-5.

[3] NOGUEIRA K, FADEL S G, DOURADO Í C, et al. Exploiting convnet diversity for flooding identification[J]. IEEE Geoscience and Remote Sensing Letters, 2018, 15(9): 1446-1450.

[4] BOOYSEN R, GLOAGUEN R, LORENZ S, et al. The potential of multi-sensor remote sensing mineral exploration: Examples from southern Africa[C]. IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2019: 6027-6030.

[5] MI L, DAI X J, CASTILLO-NAVARRO J, et al. Knowledge-aware text-image retrieval for remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024. DOI: 10.1109/TGRS.2024.3486977.

[6] 刘华咏,黄聪,金汉均.注意力增强的视觉Transformer图像检索算法[J].电子测量技术,2023, 46(23): 50-55.
LIU H Y, HUANG C, JIN H J. Image retrieval method with attention-enhanced visual Transformer[J]. Electronic Measurement Technology, 2023, 46(23): 50-55.

[7] 王欢,宋丽娟,杜方.基于多模态知识图谱的中文跨模态实体对齐方法[J].计算机工程, 2023, 49(12): 88-95.
WANG H, SONG L J, DU F. Chinese cross-modal entity alignment method based on multi-modal knowledge graph[J]. Computer Engineering, 2023, 49(12):88-95.

[8] ZHAO X, WANG L M, ZHANG Y F, et al. A review of convolutional neural networks in computer vision[J]. Artificial Intelligence Review, 2024, 57(4): 99.

[9] OTTER D W, MEDINA J R, KALITA J K. A survey of the usages of deep learning for natural language processing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(2): 604-624.

[10] HE R, ZHANG M, WANG L, et al. Cross-modal subspace learning via pairwise constraints[J]. IEEE Transactions on Image Processing, 2015, 24(12): 5543-5556.

[11] JIA Y Q, SALZMANN M, DARRELL T. Learning cross-modality similarity for multinomial data[C]. 2011 International Conference on Computer Vision. IEEE, 2011: 2407-2414.

[12] YANG X H, LIU W F, LIU W, et al. A survey on canonical correlation analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(6): 2349-2368.

[13] ESPOSITO V V, RUSSOLILLO G. Partial least squares algorithms and methods[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2013, 5(1): 1-19.

[14] LIN T Y, ROYCHOWDHUEY A, MAJI S. Bilinear CNN models for fine-grained visual recognition[C]. IEEE International Conference on Computer Vision, 2015: 1449-1457.

[15] CONG SH, ZHOU Y. A review of convolutional neural network architectures and their optimizations[J]. Artificial Intelligence Review, 2023, 56(3): 1905-1969.

[16] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[17] FAGHRI F, FLEET D J, KIROS J R, et al. Vse++: Improving visual-semantic embeddings with hard negatives[C]. British Machine Vision Conference. London: British Machine Vision Association, 2017: 1707-1717.

[18] LEE K H, CHEN X, HUA G, et al. Stacked cross attention for image-text matching[C]. European Conference on Computer Vision (ECCV), 2018: 201-216.

[19] WANG Z H, LIU X H, LI H SH, et al. Camp: Cross-modal adaptive message passing for text-image retrieval[C]. IEEE/CVF International Conference on Computer Vision, 2019: 5764-5773.

[20] YUAN ZH Q, ZHANG W K, FU K, et al. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-19.

[21] 张宏图,化春键,蒋毅,等.融合改进图卷积的跨模态检索[J].计算机工程与应用,2024, 60(11):95-104.
ZHANG H T, HUA CH J, JIANG Y, et al. Cross-modal retrieval with improved graph convolution[J]. Journal of Computer Engineering & Applications,

2024，60(11)：95-104.

[22] LIU J S，BATRA D，PARIKH D，et al. Vilbert：Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[J]. Advances in Neural Information Processing Systems，2019，32，DOI：10.48550/arxiv.1908.02265.

[23] MA Q，PAN J CH，BAI C. Direction-oriented visual-semantic embedding model for remote sensing image-text retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing，2024，62：1-14.

[24] YANG L L，ZHOU T Q，MA W T，et al. Remote sensing image-text retrieval with implicit-explicit relation reasoning [J]. IEEE Transactions on Geoscience and Remote Sensing，2024，62：1-11.

[25] CHEN X M，ZHENG X T，LU X Q. Context-aware local-global semantic alignment for remote sensing image-text retrieval [J]. IEEE Transactions on Geoscience and Remote Sensing，2025，63：1-12.

[26] ZHANG W H，CHEN J L，ZHANG W K，et al. Efficient reconstruction of spatial features for remote sensing image-text retrieval [J]. Transactions of Nanjing University of Aeronautics & Astronautics，2025(1)：101-111.

[27] 刘述喜，刘科，黄思源，等. 基于多源融合图与 SE-BiGRU-ResNet 模型的 MMC 子模块开路故障诊断[J]. 仪器仪表学报，2024，45(11)：322-337.
LIU SH X，LIU K，HUANG S Y，et al. Open-circuit fault diagnosis of MMC sub-module based on multi-sourcefusion graph and SE-BiGRU-ResNet model[J]. Chinese Journal of Scientific Instrument，2024，45(11)：322-337.

[28] KHAMNUANSIN D，CHALOTHORN T，CHUANGSUWANICH E. MrRank：Improving question answering retrieval system through multi-result ranking model [J]. ArXiv preprint arXiv：2406.05733，2024.

[29] 刘罡，李小雨，吴烨，等. 基于跨模态协同感知的双流融合动作识别模型[J]. 电子测量技术，2025，48(21)：87-97.
LIU G，LI X Y，WU Y，et al. A dual-stream fusion action recognition model based on cross-modal co-sensing [J]. Electronic Measurement Technology，2025，48(21)：87-97.

[30] ZHAO T W，WANG J，CHE J X，et al. Performance degradation prediction of proton exchange membrane fuel cell based on CEEMDAN-KPCA and DA-GRU networks[J]. Instrumentation，2024，11(1)：51-61.

[31] 范馨月，张阔，张干，等. 细微特征增强的多级联合聚类跨模态行人重识别算法[J]. 电子测量与仪器学报，2024，38(3)：94-103.
FAN X Y，ZHANG K，ZHANG G，et al. Cross-modal person re-identification algorithm based on multi-level joint clustering with subtle feature enhancement[J]. Journal of Electronic Measurement and Instrumentation，2024，38(3)：94-103.

[32] LU X Q，WANG B Q，ZHENG X T，et al. Exploring models and data for remote sensing image caption generation[J]. IEEE Transactions on Geoscience and Remote Sensing，2017，56(4)：2183-2195.

[33] YANG R，WANG SH，HAN Y P，et al. Transcending fusion：A multi-scale alignment method for remote sensing image-text retrieval [J]. IEEE Transactions on Geoscience and Remote Sensing，2024，62：1-17.

[34] 杨钰雪，何甜，樊京杭，等. 基于交叉注意力与特征聚合的跨模态图文检索研究[J]. 计算机工程，2026，52(2)：311-321.
YANG Y X，HE T，FAN J H，et al. Research on cross-modal image-text retrieval based on cross attention and feature aggregation [J]. Computer Engineering，2026，52(2)：311-321.

[35] REYAD M，SARHAN A M，ARAFA M. A modified ADAM algorithm for deep neural network optimization[J]. Neural Computing and Applications，2023，35(23)：17095-17112.

[36] YU H F，YAO F L，LU W X，et al. Text-image matching for cross-modal remote sensing image retrieval via graph neural network[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing，2023，16：812-824.

[37] ZHU Z C，KANG J，DIAO W H，et al. SIRS：Multi-task joint learning for remote sensing foreground-entity image-text retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing，2024，62：1-15.

[38] ZHENG CH Y，SONG N，ZHANG R Y，et al. Scale-semantic joint decoupling network for image-text retrieval in remote sensing[J]. ACM Transactions on Multimedia Computing，Communications and Applications，2023，20(1)：1-2.

作者简介

周晨菡，硕士研究生，主要研究方向为深度学习、跨模态检索。

E-mail：1483275902@qq.com

许晓阳(通信作者)，博士，教授，博士生导师，主要研究方向为深度学习。

E-mail：xiaoyang_xu@qq.com

魏伟，硕士研究生，主要研究方向为深度学习、目标检测。

E-mail：weiwei_xust@qq.com