ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 4,Jul. 2026,pp. 1194－1211
DOI:10. 16337/j. 1004-9037. 2026. 04. 019

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于短语增强的全局与局部对齐的遥感图像文本检索方法

金开宇，胡　筱，谢　洪，连德富

（认知智能全国重点实验室（中国科学技术大学计算机科学与技术学院），合肥 230088）

摘　要：遥感图文检索现有的方法主要通过对齐配对图像和文本的全局特征来实现检索，但大多方法的注意力都集中在特征提取和特征融合方面，在特征对齐时都只关注于经过汇聚后的全局特征之间的对齐（全局对齐），而忽略掉了图像中细粒度的物体特征与文本中的语义短语特征之间的对齐（局部对齐）。然而在遥感领域，同一个模态内的样本往往高度相似，这会干扰全局对齐。本文提出一种全新的模型——带有短语增强的全局与局部对齐（Global and local alignment with phrase augmentation，GLAPA），在对齐全局特征的同时，引入细粒度的局部特征对齐作为补充。本文通过提取文本中的概念短语，并动态地与图像中的区域特征进行局部对齐，显著提升了检索性能。此外，本文还通过掩码建模增强了单模态内特征之间的关联性学习，进一步提升遥感图像－文本检索的效果。本文在RSICD和RSITMD数据集上进行了大量实验，包括消融实验、超参数敏感性分析和可视化分析等。实验结果表明，GLAPA在RSICD和RSITMD数据集上的表现优于现有的方法，可视化实验也验证了局部对齐的准确性。

关键词：图像-文本检索；遥感；特征对齐；局部对齐；短语增强

中图分类号：TP311.13;TP309　　**文献标志码：**A

引用格式：金开宇，胡筱，谢洪，等．基于短语增强的全局与局部对齐的遥感图像文本检索方法[J]．数据采集与处理，2026，41(4)：1194-1211. JIN Kaiyu，HU Xiao，XIE Hong，et al. Global and local alignment with phrase augmentation for remote sensing image-text retrieval[J]. Journal of Data Acquisition and Processing，2026，41(4)：1194-1211.

引　　言

跨模态检索是近年来出现的一项重要检索任务，旨在从不同模态中自动挖掘给定查询的有价值数据。在遥感领域，跨模态检索最近引起了越来越多的关注，它已被广泛应用于灾害监测和农业管理[1-3]。在这些应用中，遥感图像-文本检索（Remote sensing image-text retrieval，RSITR）起着至关重要的作用，其目的是在给定文本作为查询的情况下，搜索具有相似语义的遥感图像，反之亦然。因此该任务近年来受到广泛关注，并涌现出多种改进方法[4]。

根据使用主干的不同，遥感图像-文本检索的方法可以分为基于卷积神经网络（Convolutional neural network，CNN）＋循环神经网络（Recurrent neural network，RNN）的方法和基于Transformer的方法。早期的方法是通过CNN提取图像特征，RNN提取文本特征。如DBTN[5]通过平均融合策略来组合嵌入空间中的特征，开创了遥感图像-文本检索问题的探索。随后的AMFMN[6]、SWAN[7]、KAM-CL[8]和DOVE[9]采用更加精细的特征融合和提取方法进行改进。近些年来，Transformer架构凭借其强大的全局特征融合能力而受到关注。基于Transformer的视觉和文本表示能够进一步提高检索性能，例如FBCLM[10]使用自注意提取单模态的细粒度特征，使用交叉注意力实现跨模态信息融合；新发布的MTGFE[11]模型，通过多模态融合编码方法，均取得了良好的表现。后续许多基于对比语言-图像预训

基金项目：国家科技创新2030重大项目（2021ZD0111800）。
收稿日期：2025-03-01；**修订日期：**2025-09-08

练模型(Contrastive language-image pre-training model, CLIP)[12]的方法,包括RS5M[13]、PE-RSITR[14]、EBAKER[15]、HarMA[16]和RemoteCLIP[17],通过使用更多遥感领域的数据或者设计更精细的微调适配器,进一步提升了模型的性能。

尽管现有方法,如EBAKER和KAMCL,在遥感图文检索任务中取得了一定成效,前者通过"去除弱相关样本"的策略优化全局对齐过程,后者则借助外部知识增强全局语义判别能力,但它们仍主要侧重于全局特征层面的对齐(全局对齐),缺乏对图像中细粒度的物体特征与文本中的语义短语特征之间的对齐(局部对齐)。然而,遥感图像具有拍摄角度固定、分辨率一致、地物纹理与色调高度相似等特点,导致同一模态内的样本在视觉上呈现出强烈的同质性,严重干扰了全局语义对齐的效果。本文在RSICD数据集上的实证分析表明,不同类别图像的全局特征余弦相似度平均超过0.9,且t-SNE可视化结果显示多个类别在嵌入空间中高度重叠,验证了遥感场景的"同模态相似干扰"现象。因此,如果能够在细粒度级别上对齐图像中的物体和文本内的语义短语,建立特定图像区域与文本概念短语之间的联系,就能够区分出遥感图像和文本中不同的差异,以局部对齐相似度补充全局对齐相似度,从而进一步提高检索能力。

为了克服这个问题,本文提出带有短语增强的全局与局部对齐(Global and local alignment with phrase augmentation, GLAPA)模型。该模型通过联合建模全局相似度与局部相似度,实现更全面的跨模态语义对齐,从而提升检索性能。具体来说,首先使用对比学习对齐视觉和文本的全局特征,然后从文本中提取出概念短语,并将短语特征与图像中进行动态选择,即计算它与图像所有区域块之间的相似度,只选取相似度高于阈值的图像块与token特征进行对齐,在细粒度级别捕捉文本和图像之间的关系,称为局部对齐。此外,还对视觉和语言模态进行掩码建模,使得模型能够更有效地捕获单模态内特征之间的相关性。综上,GLAPA模型能够根据语义内容自适应建立图文之间细粒度对齐关系,其引入的动态局部对齐策略有效弥补了全局对齐在遥感图像高度相似背景下的不足,显著增强了模型的细粒度对齐能力与跨模态判别性能。

1 相关工作

根据使用主干的不同,遥感图像-文本检索的方法可以分为基于CNN+RNN的方法和基于Transformer的方法。

(1)基于CNN+RNN的方法。早期的方法是通过CNN提取图像特征,RNN提取文本特征。DBTN[5]通过平均融合策略来组合嵌入空间中的特征,从而能够学习更稳健的表示,开创了遥感图像-文本检索问题的探索。AMFMN[6]通过多尺度视觉自注意模块提取显著特征,并使用视觉特征指导文本特征的提取,使用动态边界三重损失来缓解类内相似性。GaLR[18]提出了多级信息动态融合(Multi-level information dynamic fusion, MIDF),以融合遥感图像的局部和全局特征来增强图像的语义表达能力。SWAN[7]提出了视觉多尺度融合和场景细粒度感知模块来提升模型的场景理解能力。DOVE[9]利用区域视觉特征自适应地调整视觉和文本嵌入以解决视觉语义冗余问题。RSlight[19]使用知识蒸馏方法将CLIP模型的知识转移到轻量级模型中,以实现高效和准确的遥感图文检索。张永梅等[20]提出基于注意力机制和软匹配的多标签遥感图像检索方法,以充分挖掘多标签遥感图像的高层语义信息。CABIR[21]利用交叉注意力实现跨模态信息交互,并用文本语义引导网络为图像特征分配权重和过滤冗余特征。ITRM[22]从图像中提取出显著性和纹理信息,用于当作自监督信息,增强图像特征的提取。ESFN[23]利用遥感图像和文本中的实体语义来提高对齐程度,提高检索精度。

(2)基于Transformer的方法。Transformer模型凭借其强大的全局特征融合能力,基于Transformer的视觉和文本表示能够进一步提高检索性能。FBCLM[10]使用自注意提取单模态的细粒度特征,使用交叉注意实现跨模态信息融合。PIR[24]提出使用遥感场景识别的先验知识来指导视觉和文本表征的适应性学习,以解决语义噪声问题。SL-SJR[25]通过将遥感图像序列化为图像块来实现图像的布局化建

模来实现遥感对象的远距离依赖和次要目标的保留，以增强遥感图文检索的准确度。随着CLIP[12]在图文检索任务上的成功，许多基于CLIP的方法被提出。RemoteCLIP[17]注释了多个遥感数据集，并比较了不同尺度大小的CLIP模型的性能。HarMA[16]使用InfoNCE损失和自适应三重损失一起微调CLIP模型，在增强均匀对齐的同时保留了预训练的知识。EBAKER[15]通过在对齐前过滤掉弱相关样本对，以减轻与最佳嵌入空间的偏差，并显示推理关键词以区分概念之间微妙的区别。

基于CNN＋RNN的模型采取了精细的特征提取方法，然而，这类方法主要侧重于全局特征的学习，难以有效建模局部短语与图像区域之间的对应关系；基于Transfromer的模型的特征融合能力也相当出色，但是基于Transformer的方法仍然主要依赖全局信息，而对局部短语-图像区域的细粒度对齐研究较少。

除此之外，在遥感领域，现有关于数据增强的工作大多注重于解决遥感图像目标识别研究中样本数量不足的问题[26-27]。为了让模型更关注重要区域，减少对不相关特征的依赖，提高检索的精准度，EBAKER[15]通过引入关键字显式推理（Keyword explicit reasoning，KER）模块，鼓励模型预测遥感图像-文本对局部特征中关键概念的细微差别；KAMCL[8]中知识辅助学习（Knowledge-assisted learning，KAL）框架也对关键概念进行了强化；ESFN[23]利用遥感图像和文本中的实体语义来提高对齐程度；田澍等[28]通过目标语义提示与空间-通道双注意力机制增强了遥感图像-文本检索中对显著局部目标的关注。尽管有对于文本信息中关键短语的关注，但这些工作仍主要利用全局对齐的结果进行检索预测。

与这些方法相比，本文方法关注了完整的局部短语语义，将短语与图像进行细粒度地对齐。同时，本文提出了多模态局部对齐方法，即通过短语增强来辅助局部对齐。与现有研究相比，本文模型更注重局部完整语义的对齐，而不仅仅是利用全局对齐的结果进行检索预测。通过提取文本中的概念短语进行短语增强，并将短语与图像中动态选择出的区域图像块特征进行对齐，本文的GLAPA模型能够在细粒度级别捕捉文本和图像之间的关系。

2 本文方法

如图1所示，本文的方法主要包括3个部分：首先使用CLIP的视觉和文本编码器提取特征，并使用spaCy库进行短语提取，将提取的短语输入文本编码器以获取特征；然后使用对比学习进行全局特征对齐，并使用短语的全局特征与图像的局部特征进行局部特征对齐；引入掩码建模，以建模单模态内部的语义关联。图1中变量含义见本文相关章节。

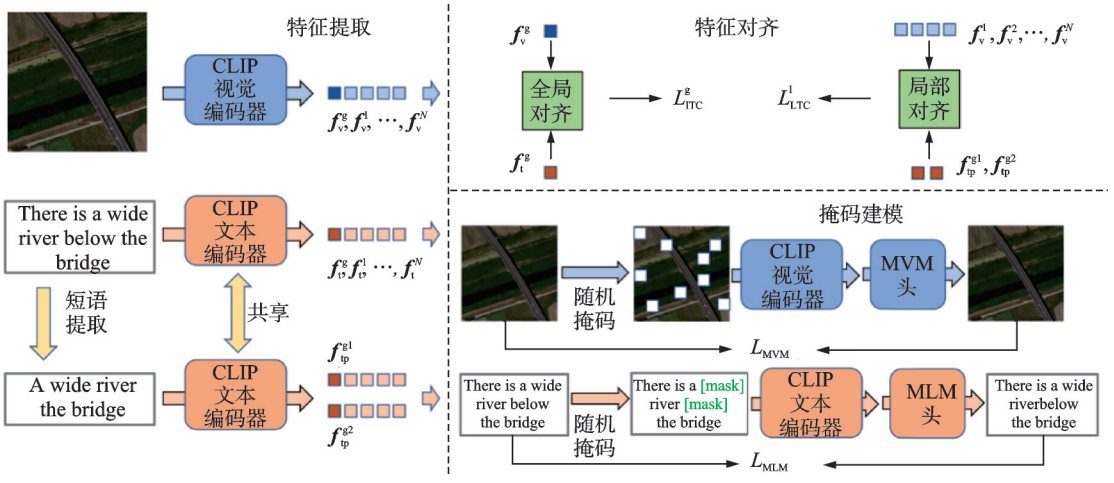


图1 系统模型图

Fig.1 System model diagram

2.1 特征提取

如图1的左边部分所示,特征提取包括视觉编码器和文本编码器,以及短语的特征提取。

(1) 视觉编码器。本文采用ViT架构作为视觉编码器。给定输入图像$\boldsymbol{I}\in\mathbf{R}^{H\times W\times C}$,其中$H$、$W$和$C$分别表示图像的长、宽和通道数目。首先将$\boldsymbol{I}$分割为$N=\frac{H\times W}{P^2}$个图像块,$P$代表图像块的尺寸。然后,所有的图像块通过一个可学习的线性投影映射为一维的token,添加上位置编码和一个可学习的[CLS]token后输入到L层Transformer中。最后,所有的token经过一个线性映射得到D维的输出,其中$\boldsymbol{f}_{\mathrm{v}}^{\mathrm{g}}$表示从[CLS] token汇聚得到的视觉全局特征,$\{\boldsymbol{f}_{\mathrm{v}}^{1},\boldsymbol{f}_{\mathrm{v}}^{2},\cdots,\boldsymbol{f}_{\mathrm{v}}^{N}\}$表示不同图像块的视觉局部特征。上述的过程可以表示为

$$\boldsymbol{f}_{\mathrm{v}}^{\mathrm{g}},\boldsymbol{f}_{\mathrm{v}}^{1},\boldsymbol{f}_{\mathrm{v}}^{2},\cdots,\boldsymbol{f}_{\mathrm{v}}^{N}=\phi(I) \tag{1}$$

式中ϕ表示CLIP的视觉编码器[12]。

(2) 文本编码器。本文使用CLIP的文本编码器作为文本编码器。给定输入文本T,首先使用BPE(Byte pair encoding)进行分词,并在文本的开头和结尾分别插入[SOS]token和[EOS] token。然后,使用嵌入矩阵将token转换为向量表示,并添加位置编码。最后,将所有token嵌入输入到L层Transformer中,并经过一个线性映射,得到每个token的表示$\{\boldsymbol{f}_{\mathrm{t}}^{\mathrm{sos}},\boldsymbol{f}_{\mathrm{t}}^{1},\boldsymbol{f}_{\mathrm{t}}^{2},\cdots,\boldsymbol{f}_{\mathrm{t}}^{M},\boldsymbol{f}_{\mathrm{t}}^{\mathrm{eos}}\}$,其中$\boldsymbol{f}_{\mathrm{t}}^{\mathrm{eos}}$作为文本的全局特征$\boldsymbol{f}_{\mathrm{t}}^{\mathrm{g}}$,$\{\boldsymbol{f}_{\mathrm{t}}^{1},\boldsymbol{f}_{\mathrm{t}}^{2},\cdots,\boldsymbol{f}_{\mathrm{t}}^{M}\}$表示不同token的文本局部特征。上述的过程可以表示为

$$\boldsymbol{f}_{\mathrm{t}}^{\mathrm{g}},\boldsymbol{f}_{\mathrm{t}}^{1},\boldsymbol{f}_{\mathrm{t}}^{2},\cdots,\boldsymbol{f}_{\mathrm{t}}^{M}=\phi(T) \tag{2}$$

式中:ϕ表示CLIP的文本编码器[12],M为文本中的token数量。

(3) 概念短语增强。本文使用著名的自然语言处理(Natural language processing, NLP)处理工具spaCy库,从文本T中提取E个概念短语p,并将每个概念短语单独输入到文本编码器中得到每个概念短语的全局和局部特征,这里只使用每个概念短语的全局特征。例如,针对“an empty port is near some buildings and many green trees。”这句话,提取出{“an empty port”,“some buildings”,“many green trees”}3个概念短语。将这3个概念短语分别送入到文本编码器中,得到它们的全局特征$\{\boldsymbol{f}_{\mathrm{tp}}^{\mathrm{g1}},\boldsymbol{f}_{\mathrm{tp}}^{\mathrm{g2}},\boldsymbol{f}_{\mathrm{tp}}^{\mathrm{g3}}\}$。上述的过程可以表示为

$$\boldsymbol{f}_{\mathrm{tp}}^{\mathrm{g1}},\boldsymbol{f}_{\mathrm{tp}}^{\mathrm{g2}},\cdots,\boldsymbol{f}_{\mathrm{tp}}^{\mathrm{g}E}=\phi(\pi(T)) \tag{3}$$

式中:π表示spaCy库中的概念短语提取工具,E表示文本中的短语数目。

2.2 特征对齐

特征对齐包括全局特征对齐和局部特征对齐,分别用于匹配整体语义和细粒度信息。

(1) 全局特征对齐。与CLIP中的对齐方式类似,本文使用对比学习对齐全局视觉特征和全局文本特征。具体地,记$s^{\mathrm{g}}=\cos(\boldsymbol{f}_{\mathrm{v}}^{\mathrm{g}},\boldsymbol{f}_{\mathrm{t}}^{\mathrm{g}})$为图像和文本之间的全局相似度,$s_{ij}^{\mathrm{g}}$为第$i$张图像与第$j$个文本之间的全局相似度,则全局对齐损失为

$$L_{\mathrm{ITC}}^{\mathrm{g}}=-\frac{1}{B}\sum_{i=1}^{B}\ln\frac{\exp(s_{ii}^{\mathrm{g}}/\tau)}{\sum_{j=1}^{B}\exp(s_{ij}^{\mathrm{g}}/\tau)}-\frac{1}{B}\sum_{i=1}^{B}\ln\frac{\exp(s_{ii}^{\mathrm{g}}/\tau)}{\sum_{j=1}^{B}\exp(s_{ji}^{\mathrm{g}}/\tau)} \tag{4}$$

式中:s_{ii}^{g}代表正配对之间的相似度,B为批量大小,τ为温度超参数。

(2) 局部特征对齐。除了全局对齐外,本文还会对细粒度的局部特征进行对齐。本文把概念短语的全局特征与图像的局部特征进行对齐,即对于每一个概念短语,找到其相似度大于一定阈值的图像块进行对齐。具体来说,记$s_{ne}^{\mathrm{l}}=\cos(\boldsymbol{f}_{\mathrm{v}}^{n},\boldsymbol{f}_{\mathrm{tp}}^{\mathrm{g}e})$为图像的第$n$个图像块与文本的第$e$个概念短语之间的局部相似度,则对于第$e$个概念短语,其与图像之间的相似度为

$$s_e^{\mathrm{l}}=\frac{1}{|A_e|}\sum_{n\in A_e}s_{ne}^{\mathrm{l}}\quad A_e=\{n|s_{ne}^{\mathrm{l}}>\gamma,\ n\in\{1,2,\cdots,N\}\}\tag{5}$$

式中:γ为一个阈值超参数,本文在验证集上对其进行调优,并固定最优取值用于所有实验,A_e表示与第e个概念短语相似度大于γ的局部相似度集合。本文使用一个文本中所有概念短语与图像相似度的平均值作为文本与图像之间的局部相似度,即

$$s^{\mathrm{l}}=\frac{1}{E}\sum_{e=1}^{E}s_e^{\mathrm{l}}\tag{6}$$

记s_{ij}^{l}为第i张图像与第j个文本之间的局部相似度,则局部对齐损失为

$$L_{\mathrm{ITC}}^{\mathrm{l}}=-\frac{1}{B}\sum_{i=1}^{B}\ln\frac{\exp(s_{ii}^{\mathrm{l}}/\tau)}{\sum_{j=1}^{B}\exp(s_{ij}^{\mathrm{l}}/\tau)}-\frac{1}{B}\sum_{i=1}^{B}\ln\frac{\exp(s_{ii}^{\mathrm{l}}/\tau)}{\sum_{j=1}^{B}\exp(s_{ji}^{\mathrm{l}}/\tau)}\tag{7}$$

式中s_{ii}^{l}代表正配对之间的相似度。

2.3 掩码建模

掩码建模用于增强模型的理解能力,通过对图像和文本进行部分屏蔽并重建缺失内容,从而学习更有效的特征表示。

(1) 掩码可视化建模(Masked vision modeling, MVM):掩码可视化建模旨在更好地理解图像中的局部信息,并捕获图像块之间的关系。给定输入图像$\boldsymbol{I}$,随机屏蔽其75%的图像块以获得掩码图像$\boldsymbol{I}^{\mathrm{m}}$。然后将掩码图像$\boldsymbol{I}^{\mathrm{m}}$送入视觉编码器以提取每个图像块的特征。这些特征通过一个MVM头来重建原始图像,重建后的图像记为$\hat{\boldsymbol{I}}$。

本文使用均方误差损失(Mean squared error, MSE)来优化重建图像$\hat{\boldsymbol{I}}$和原始图像$\boldsymbol{I}$之间的掩蔽区域的重建,有

$$L_{\mathrm{MVM}}=\frac{1}{|M_{\mathrm{v}}|}\sum_{i\in M_{\mathrm{v}}}(\boldsymbol{I}_i-\hat{\boldsymbol{I}}_i)^2\tag{8}$$

式中M_{v}表示掩码图像块的指标集。

(2) 掩码语言建模(Masked language modeling, MLM)。掩码语言建模广泛应用于语言模型和多模态训练中,以增强模型的语言理解能力。给定一个文本$\boldsymbol{T}$,用[MASK]标记随机替换其中15%的单词,得到掩码文本$\boldsymbol{T}^{\mathrm{m}}$,然后将$\boldsymbol{T}^{\mathrm{m}}$输入文本编码器以提取每个标记的特征。这些特征通过MLM头部处理来预测单词概率,表示为$\hat{\boldsymbol{T}}\in\mathbf{R}^{M\times V}$,其中$M$是$\boldsymbol{T}$中标记的数量,$V$是词汇表大小。使用交叉熵损失来优化预测概率,有

$$L_{\mathrm{MLM}}=-\frac{1}{|M_{\mathrm{t}}|}\sum_{i\in M_{\mathrm{t}}}\log\hat{T}_{i,y_i}\tag{9}$$

式中:M_{t}表示屏蔽词的索引集,y_i为第i个屏蔽词在词汇表中的索引,$\hat{T}_{i,y_i}$为第i个词为y_i的预测概率。

2.4 训练和推理

本节介绍模型的训练目标与推理方法,通过不同损失函数的加权组合优化模型,并在推理阶段结合全局和局部相似度进行匹配。

(1) 训练。将全局对齐损失、局部对齐损失和掩码建模损失进行加权平均用于训练模型,有

$$L=\lambda_1L_{\mathrm{ITC}}^{\mathrm{g}}+\lambda_2L_{\mathrm{ITC}}^{\mathrm{l}}+\lambda_3L_{\mathrm{MVM}}+\lambda_4L_{\mathrm{MLM}}\tag{10}$$

式中:λ_1、λ_2、λ_3、λ_4为超参数,用于平衡不同损失函数的重要性。具体在实验中,设置$\lambda_1=1$,$\lambda_2=1$,$\lambda_3=1.0$,$\lambda_4=0.5$。

(2) 推理。在推理阶段,使用全局相似度s^g和局部相似度s^l的加权平均作为最终相似度,即

$$s=\alpha s^g+(1-\alpha)s^l \tag{11}$$

式中α为一个超参数,用于平衡全局相似度和局部相似度的贡献。在RSICD数据集[29]上本文默认设置$\alpha=0.7$,即全局相似度占70%,局部相似度占30%;而在RSITMD数据集[6]上,本文默认设置$\alpha=0.5$。在后续的实验分析中本文对α进行了敏感性分析,以评估不同设定对匹配效果的影响。

3 实 验

3.1 实验设置

(1) 数据集。实验中使用了遥感图像-文本检索中最常用的两个基准数据集RSICD[29]和RSITMD[6]来验证本文方法的有效性。RSICD数据集由10 921张分辨率为224像素×224像素的图像组成,每张图像都有5个标题,共计18 190个唯一标题。按照文献[22]中的设置,本文将数据集分成7 862张图像用于训练,1 966张用于验证,1 093张用于测试。RSITMD数据集包含4 743张分辨率为256像素×256像素的图像和23 715个标题,其中21 829个唯一。按照文献[22]中的配置,本文使用3 435张图像进行训练,856张用于验证,452张用于测试。

(2) 实现细节。GLAPA的实现基于HarMA[16]和PIR[24]的代码库,利用OpenCLIP[30]提供的ViT-B-32架构,在自然图像上使用预训练的权值进行初始化。MLM头和MVM头都由一个线性层组成。与在HarMA[16]中一样,本文使用批大小为64的AdamW[31]优化器训练模型,进行40次epochs的训练。此外,本文采用线性预热和线性衰减的学习率调度策略,将学习率设置为$4e^{-4}$,将预热步长比设置为0.2。所有实验都在PyTorch框架下实现,并在两个NVIDIA RTX 3090 GPU上进行训练。

(3) 评估指标。根据之前的遥感图像-文本检索研究,本文使用R@K(Recall at K,$K=1,5,10$)和mR(Mean recall)指标评估方法的有效性。R@K为检索的前K个结果中正确匹配对的比例,而mR表示不同K值下R@K值的平均值,用于提供对检索性能的总体评估。

3.2 主要实验

用于比较的基线方法有:基于CNN+RNN的SWAN[7]、KAMCL[8]、ESFN[23]和DOVE[9],以及基于Transformer的AEFEF[32]、FAAMI[33]、PIR[24]和MTGFE[11]。此外还有基于CLIP(VIT-B-32)预训练模型的PE-RSITR[14]、HarMA[16]、CLIP-FT[30]和RemoteCLIP[17]作为强基线。其中CLIP-FT是在预训练的CLIP模型上使用遥感数据集进行的全量微调。RemoteCLIP则是使用了RET-3训练数据的版本,比其他模型使用了更多的训练数据。为保证结果公平,本文引用原始论文给出的结果用以比较。

本文的方法属于基于CLIP的模型。RSICD和RSITMD数据集的实验结果如表1,2所示。此外,模型使用的视觉和文本主干也展现在表1中。从这些结果中,本文能够得到以下观察结果和结论:

(1) RSICD上的结果。RSICD数据集上的结果展示在表1中(本文中所有表格中加粗数字表示最优结果)。可以看到,使用Transformer作为主干的方法一般要比基于CNN+RNN的方法要好。更进一步,基于CLIP(ViT-B-32)的方法则比普通的Transformer方法表现更好。本文的方法基于CLIP(ViT-B-32),并且添加了额外的细粒度局部对齐和掩码建模,因此在使用了相同的训练数据下,在mR指标上比CLIP-FT提高了5.06%,甚至比使用了更多训练数据用于微调的RemoteCLIP在mR上还要高3.43%。总的来说,本文的结果在RSICD数据集上的各个指标都达到了最先进的效果。

(2) RSITMD上的结果。RSITMD数据集上的结果展示在表2中。对比RSICD,RSITMD有更多样性的文本,并且文本描述更加细致,因此有着更高的检索正确率。基于CLIP(ViT-B-32)的方法在RSITMD上的提升更加明显,远超基于CNN+RNN和普通Transformer方法。本文方法在图查文本的

表1 RSICD数据集上的图像-文本检索结果比较

Table 1 Comparison results of the image-text retrieval on RSICD dataset %

方法	主干	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
SWAN[7]	CNN+RNN	7.41/20.13/30.86	5.56/22.26/37.41	20.61
ESFN[23]	CNN+RNN	8.14/21.04/33.03	6.48/24.08/40.16	22.16
DOVE[9]	CNN+RNN	8.66/22.35/34.95	6.04/23.95/40.35	22.72
KAMCL[8]	CNN+RNN	12.08/27.26/38.70	8.65/27.43/42.51	26.10
AEFEF[32]	Transformer	7.13/16.27/24.95	6.33/23.27/38.85	19.47
FAAMI[33]	Transformer	10.44/22.66/30.89	8.11/25.59/41.37	23.18
PIR[24]	Transformer	9.88/27.26/39.16	6.97/24.56/38.92	24.46
MTGFE[11]	Transformer	15.28/37.05/51.60	8.67/27.56/43.92	30.68
PE-RSITR[14]	CLIP	14.13/31.51/44.78	11.63/33.92/50.73	31.12
CLIP-FT[30]	CLIP	15.89/36.14/47.93	12.21/32.97/48.84	32.33
HarMA[16]	CLIP	16.36/34.48/47.74	12.92/37.17/53.07	33.62
RemoteCLIP[17]	CLIP	17.84/35.96/50.14	13.89/35.15/50.08	33.96
本文模型	CLIP	**18.85/40.35/53.61**	**14.86/40.02/56.65**	**37.39**

表2 RSITMD数据集上的图像-文本检索结果比较

Table 2 Comparison results of the image-text retrieval on RSITMD dataset %

方 法	主干	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
SWAN[7]	CNN+RNN	13.35/32.15/46.90	11.24/40.40/60.60	34.11
ESFN[23]	CNN+RNN	18.81/36.50/52.43	12.52/43.14/62.79	37.70
DOVE[9]	CNN+RNN	16.81/36.80/49.93	12.20/44.13/66.50	37.73
KAMCL[8]	CNN+RNN	16.51/36.28/49.12	13.50/42.15/59.32	36.14
AEFEF[32]	Transformer	13.27/36.06/51.33	11.55/41.42/60.00	35.60
FAAMI[33]	Transformer	16.15/35.62/48.89	12.96/42.39/59.96	35.99
PIR[24]	Transformer	18.14/41.15/52.88	12.17/41.68/63.51	38.24
MTGFE[11]	Transformer	17.92/40.93/53.32	16.59/48.50/67.43	40.78
PE-RSITR[14]	CLIP	23.67/44.07/60.36	20.10/50.63/67.97	44.47
CLIP-FT[30]	CLIP	26.99/46.90/58.85	20.53/52.35/71.15	46.13
HarMA[16]	CLIP	25.81/48.37/60.61	19.92/53.27/71.21	46.53
RemoteCLIP[17]	CLIP	24.78/**50.00/63.27**	22.61/55.27/69.87	47.63
本文模型	CLIP	**29.87**/49.56/62.39	**23.98/57.35/74.03**	**49.53**

R@1上相比于CLIP-FT提高了2.88%,达到了接近30%的准确率。并且在文本查图的检索上,本文方法充分利用了局部对齐分数,全面超过现有的方法。最后,本文方法比RemoteCLIP在mR上提高了1.9%,表明在使用更少训练数据的情况下,本文方法依然能够达到最先进的效果。

3.3 分析实验

本节通过在RSICD和RSITMD数据集上的多个分析实验来验证本文模型的有效性,包括对损失函数的各个部分进行消融实验,以及对训练过程中损失函数超参数λ_1、λ_2、λ_3和λ_4、掩码建模屏蔽比例与相似度分数中的全局和局部分数比例超参数α的敏感性分析。

(1) 消融实验。表3~4展示了本文对损失函数的各个部分进行消融实验的结果。本文使用全局对齐损失,也就是CLIP-FT作为基线。然后分别添加局部对齐、MVM、MLM损失、短语增强以及动态区域选择(+loc表示局部损失、+v表示掩码视觉建模、+l表示掩码语言建模、+aug表示短语增强、+d表示动态区域选择),可以发现mR在RSICD数据集上分别提升了3.01%、2.23%、1.28%、2.65%和3.06%,在RSITMD数据集上分别提升了1.55%、0.80%、0.62%、1.54%和1.97%。其中,局部对齐损失与动态区域选择带来的增益最为显著——前者能够在细粒度层面增强跨模态对齐能力,捕捉并区分遥感图像中的关键细节;后者则通过动态调整过滤阈值,有效剔除无关特征,提升匹配精度进一步的组合实验显示,局部对齐和MVM的组合提升最大,这表明视觉局部特征的重要性。最后将所有的损失都添加到模型中,mR得分在RSICD数据集上达到了37.39%,在RSITMD数据集上达到了49.53%。

表3 RSICD数据集上不同损失的消融实验结果比较

Table 3 Ablation experimental results comparison of different losses on RSICD dataset %

方法	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
基线	15.89/36.14/47.93	12.21/32.97/48.84	32.33
+loc	17.02/38.06/51.51	13.74/37.37/54.33	35.34
+v	16.56/36.41/50.41	13.05/37.31/53.60	34.56
+l	16.93/34.13/47.94	12.72/36.29/53.65	33.61
+aug	18.11/36.86/49.76	13.52/37.29/54.34	34.98
+d	18.38/37.13/49.30	14.70/38.05/54.77	35.39
+loc+v	17.47/38.15/52.42	14.53/38.24/56.10	36.15
+loc+l	18.39/37.79/50.96	14.36/38.83/55.72	36.01
+v+l	16.74/37.33/50.96	13.17/37.69/54.29	35.03
+loc+v+l	**18.85/40.35/53.61**	**14.86/40.02/56.65**	**37.39**

表4 RSITMD数据集上不同损失的消融实验结果比较

Table 4 Ablation experimental results comparison of different losses on RSITMD dataset %

方法	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
基线	26.99/46.90/58.85	20.53/52.35/71.15	46.13
+loc	25.44/49.34/61.50	21.86/55.09/72.83	47.68
+v	25.66/46.46/62.39	20.93/53.10/73.05	46.93
+l	24.78/47.12/61.73	20.31/53.94/72.61	46.75
+aug	25.52/49.19/60.03	22.28/55.78/73.21	47.67
+d	28.21/47.23/63.38	21.88/55.60/72.28	48.10
+loc+v	28.54/49.56/62.61	22.79/55.97/73.50	48.83
+loc+l	28.54/49.34/**62.83**	22.79/55.97/73.41	48.81
+v+l	25.00/**49.78**/61.28	20.88/53.94/72.17	47.18
+loc+v+l	**29.87**/49.56/62.39	**23.98/57.35/74.03**	**49.53**

总而言之,局部对齐损失对性能提升最为显著,MVM和MLM进一步增强了特征建模能力:MVM主要用于提升视觉表示能力,尽管提升不如局部对齐显著,但仍然对检索性能有所帮助;MLM主要增强文本表征能力,在遥感文本描述相对较短的情况下,其影响相对较小。整合所有损失,本文模型得以达到最先进的性能。

(2) 超参数敏感性分析。表5~12分别展示了λ_1、λ_2、λ_3和λ_4的敏感性分析实验,表13~16展示了掩码建模的屏蔽比例的敏感性分析实验,表17~18展示了α的敏感性分析实验,以下是各结果的具体分析。

表5 RSICD数据集上λ_1的敏感性

Table 5 Sensitivity of λ_1 on RSICD dataset %

λ_1	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
0.01	11.63/29.37/41.35	12.46/32.81/50.04	29.61
0.1	16.48/39.62/51.78	15.50/40.47/57.91	36.96
0.5	18.12/40.72/52.15	15.21/40.31/57.54	37.34
1	18.85/40.35/53.61	14.86/40.02/56.65	37.39
5	17.39/38.71/52.97	15.21/40.11/58.11	37.08
10	16.38/39.26/50.96	15.66/40.11/57.30	36.61
100	15.93/37.33/49.58	14.11/36.72/54.78	34.74

表6 RSITMD数据集上λ_1的敏感性

Table 6 Sensitivity of λ_1 on RSITMD dataset %

λ_1	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
0.01	22.12/38.94/54.21	20.35/47.43/65.31	41.39
0.1	26.77/49.11/62.61	22.66/55.93/73.59	48.44
0.5	28.32/51.10/61.28	23.89/56.11/74.91	49.27
1	29.87/49.56/62.39	23.98/57.35/74.03	49.53
5	28.76/49.56/61.28	24.29/57.74/74.16	49.30
10	27.88/48.23/60.84	24.20/57.17/74.60	48.82
100	25.66/48.67/61.95	23.45/54.56/71.95	47.70

表7 RSICD数据集上λ_2的敏感性

Table 7 Sensitivity of λ_2 on RSICD dataset %

λ_2	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
0.01	19.58/39.89/49.86	14.77/39.71/54.55	36.40
0.1	21.05/39.62/50.59	13.83/39.77/53.92	36.47
0.5	20.96/40.17/52.97	14.57/39.60/55.48	37.29
1	18.85/40.35/53.61	14.86/40.02/56.65	37.39
5	20.59/40.81/50.96	14.42/39.86/54.99	36.94
10	18.67/37.88/50.50	14.22/39.16/54.84	35.88
100	15.92/34.04/46.20	12.52/35.63/50.28	32.43

表8 RSITMD数据集上λ_2的敏感性

Table 8 Sensitivity of λ_2 on RSITMD dataset %

λ_2	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
0.01	26.55/52.00/60.62	23.89/56.24/74.12	48.90
0.1	28.54/51.78/59.52	24.51/55.80/75.04	49.19
0.5	28.54/50.01/62.39	25.00/56.33/74.25	49.41
1	29.87/49.56/62.39	23.98/57.35/74.03	49.53
5	27.43/50.23/61.95	23.40/57.04/74.29	49.05
10	25.88/50.67/59.96	25.13/56.20/74.96	48.80
100	24.34/47.13/58.41	24.47/55.04/74.34	47.28

表9 RSICD数据集上λ_3的敏感性

Table 9 Sensitivity of λ_3 on RSICD dataset %

λ_3	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
0.01	18.30/39.89/52.60	14.90/39.60/55.98	36.87
0.1	19.03/39.43/52.88	14.77/40.51/55.37	37.00
0.5	16.74/40.62/53.61	15.01/39.28/56.47	36.95
1	18.85/40.35/53.61	14.86/40.02/56.65	37.39
5	16.74/41.08/52.88	14.97/39.69/56.14	36.91
10	16.47/40.71/53.61	14.48/39.10/55.30	36.61
100	18.21/38.88/50.22	12.17/36.23/51.91	34.60

表10 RSITMD数据集上λ_3的敏感性

Table 10 Sensitivity of λ_3 on RSITMD dataset %

λ_3	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
0.01	27.88/48.01/62.17	24.16/56.60/74.61	48.91
0.1	26.77/48.45/61.06	22.83/58.15/75.32	48.77
0.5	28.76/48.01/62.17	22.83/57.31/74.96	49.01
1	29.87/49.56/62.39	23.98/57.35/74.03	49.53
5	26.11/48.89/61.28	23.81/57.13/73.59	48.47
10	29.43/47.79/59.07	23.63/55.28/73.99	48.20
100	25.67/47.79/60.40	23.41/57.31/74.25	48.14

表11 RSICD数据集上λ_4的敏感性

Table 11 Sensitivity of λ_4 on RSICD dataset %

λ_4	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
0.01	20.41/39.26/51.23	14.39/39.72/55.43	36.74
0.1	18.85/39.53/53.52	14.68/39.39/56.36	37.05
0.5	18.85/40.35/53.61	14.86/40.02/56.65	37.39
1	20.13/38.34/50.50	14.77/39.56/53.84	36.19
5	18.39/36.24/47.48	13.69/37.80/53.95	34.59
10	15.65/30.84/45.92	14.22/37.25/52.70	32.76
100	14.83/29.47/41.44	10.67/31.58/47.01	29.16

表 12 RSITMD数据集上λ_4的敏感性

Table 12 Sensitivity of λ_4 on RSITMD dataset %

λ_4	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
0.01	28.98/46.24/62.17	23.19/55.67/73.06	48.22
0.1	30.31/47.12/62.17	24.96/55.98/72.66	48.86
0.5	29.87/49.56/62.39	23.98/57.35/74.03	49.53
1	29.64/48.01/62.83	22.83/55.36/73.37	48.67
5	27.21/46.24/62.17	22.44/54.34/72.31	47.45
10	25.88/46.68/61.50	21.46/52.79/70.54	46.48
100	25.00/41.37/57.52	18.90/49.47/69.43	43.61

表 13 RSICD数据集上图片掩码比例的敏感性

Table 13 Sensitivity of the MVM ratio on RSICD dataset %

图片掩码比例	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
0.05	17.84/38.97/53.52	14.00/39.09/56.29	36.62
0.15	17.20/39.43/53.79	15.34/38.96/55.46	36.70
0.25	16.65/40.16/53.61	14.53/38.61/56.69	36.71
0.35	17.29/37.69/52.33	14.61/39.75/56.47	36.36
0.45	17.38/38.42/53.70	14.83/39.05/56.45	36.64
0.55	17.20/38.42/52.23	15.03/39.31/56.51	36.45
0.65	17.11/39.25/53.88	14.46/39.29/56.45	36.74
0.75	18.85/40.35/53.61	14.86/40.02/56.65	37.39
0.85	17.93/38.42/52.14	14.70/38.94/57.02	36.53
0.95	18.30/38.24/53.42	14.68/38.30/55.54	36.41

表 14 RSITMD数据集上图片掩码比例的敏感性

Table 14 Sensitivity of the MVM ratio on RSITMD dataset %

图片掩码比例	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
0.05	29.87/47.13/60.40	22.74/55.93/72.26	48.06
0.15	29.21/46.69/60.18	22.17/57.79/72.62	48.11
0.25	30.10/45.14/59.52	22.92/58.32/74.52	48.42
0.35	31.64/47.13/61.73	22.04/56.68/72.66	48.65
0.45	32.97/45.36/60.18	23.36/57.35/73.54	48.79
0.55	28.10/48.01/62.17	24.29/56.11/73.63	48.72
0.65	33.86/46.24/62.17	22.79/57.66/72.53	49.21
0.75	29.87/49.56/62.39	23.98/57.35/74.03	49.53
0.85	31.20/47.13/62.62	23.67/55.84/73.19	48.94
0.95	31.42/46.69/59.74	23.54/57.74/73.32	48.74

表 15 RSICD数据集上文本掩码比例的敏感性

Table 15 Sensitivity of the MLM ratio on RSICD dataset %

文本掩码比例	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
0.05	15.83/40.08/54.53	14.73/39.08/55.09	36.56

续表

文本掩码比例	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
0.15	18.85/40.35/53.61	14.86/40.02/56.65	37.39
0.25	18.12/39.62/52.15	15.04/39.78/56.08	36.80
0.35	17.02/40.54/53.71	14.64/38.88/54.64	36.57
0.45	16.84/40.17/53.80	13.72/39.05/56.08	36.61
0.55	16.65/40.17/54.26	14.33/39.10/54.84	36.56
0.65	18.57/38.89/53.07	14.75/39.14/55.17	36.60
0.75	17.20/39.17/52.88	14.49/39.91/55.61	36.54
0.85	16.29/39.62/52.79	14.00/39.05/55.57	36.22
0.95	16.38/39.62/53.25	13.68/38.42/55.86	36.20

表16 RSITMD数据集上文本掩码比例的敏感性

Table 16 Sensitivity of the MLM ratio on RRSITMD dataset %

文本掩码比例	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
0.05	27.21/50.23/61.95	25.31/55.49/75.00	49.19
0.15	29.87/49.56/62.39	23.98/57.35/74.03	49.53
0.25	25.66/51.78/62.61	24.29/57.30/75.49	49.52
0.35	27.43/50.23/61.51	24.60/56.46/76.20	49.40
0.45	27.43/49.78/62.17	25.31/56.64/75.18	49.41
0.55	26.55/51.55/62.17	24.38/55.27/74.47	49.06
0.65	26.11/50.23/60.84	25.00/56.99/74.65	48.97
0.75	26.33/50.45/61.95	23.01/56.90/74.20	48.80
0.85	26.33/50.45/61.51	23.80/56.11/74.20	48.73
0.95	26.11/49.56/62.17	21.55/56.46/75.35	48.53

表17 RSICD数据集上全局和局部对齐比例的敏感性

Table 17 Sensitivity of the ratio of global and local alignment on RSICD dataset %

方法	α	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
0	0	19.49/37.51/51.88	15.30/**40.60**/**56.83**	36.94
1	0.1	19.12/37.79/52.33	**15.48**/40.64/57.00	37.06
2	0.2	20.04/38.24/52.52	14.97/40.04/56.76	37.09
3	0.3	19.95/39.34/52.42	15.11/40.00/56.76	37.26
4	0.4	20.04/38.98/52.52	15.08/39.93/56.71	37.21
5	0.5	20.31/38.88/52.42	15.17/39.89/57.16	37.31
6	0.6	**20.49**/39.07/52.52	15.11/40.31/56.49	37.33
7	0.7	18.85/**40.35**/53.61	14.86/40.02/56.65	**37.39**
8	0.8	18.94/39.89/**53.89**	14.57/39.91/56.80	37.33
9	0.9	19.03/39.62/53.80	14.77/39.78/56.78	37.29
10	1	19.03/39.71/53.61	14.46/39.74/56.67	37.20

表 18 RSITMD数据集上全局和局部对齐比例的敏感性

Table 18 Sensitivity of the ratio of global and local alignment on RSITMD dataset %

方法	α	图查文本R@1/R@5/R@10	文本查图R@1/R@5/R@10	mR
0	0	28.10/50.66/61.95	22.08/56.37/73.50	48.78
1	0.1	28.54/49.56/62.61	22.79/55.97/73.50	48.83
2	0.2	28.54/49.34/**63.94**	22.83/55.93/73.36	48.99
3	0.3	27.88/50.66/62.39	23.98/56.55/73.98	49.24
4	0.4	28.54/50.22/61.95	24.07/57.30/**74.16**	49.37
5	0.5	**29.87**/49.56/62.39	23.98/57.35/74.03	**49.53**
6	0.6	**29.87**/49.78/62.17	23.94/56.99/74.12	49.48
7	0.7	29.20/50.44/61.50	24.16/57.35/**74.16**	49.47
8	0.8	28.54/50.00/61.28	**24.20**/**57.52**/73.72	49.21
9	0.9	28.32/49.78/63.72	22.57/55.97/73.41	48.96
10	1	28.10/**50.88**/62.39	22.43/55.84/73.54	48.86

对λ_1、λ_2、λ_3、λ_4的敏感性分析实验。在训练阶段，本文将全局对齐损失、局部对齐损失和掩码建模损失进行加权平均用于训练模型，即$L=\lambda_1 L_{\mathrm{ITC}}^{\mathrm{g}}+\lambda_2 L_{\mathrm{ITC}}^{\mathrm{l}}+\lambda_3 L_{\mathrm{MVM}}+\lambda_4 L_{\mathrm{MLM}}$，其中$\lambda_1$、$\lambda_2$、$\lambda_3$和$\lambda_4$为超参数，用于平衡不同损失函数的重要性。由表5～12可以得到，在RSICD和RSITMD数据集上，当$\lambda_1=1$，$\lambda_2=1$，$\lambda_3=1.0$和$\lambda_4=0.5$时，mR分别达到了最大值，验证了在实验中设置$\lambda_1=1$，$\lambda_2=1$，$\lambda_3=1.0$和$\lambda_4=0.5$的合理性。

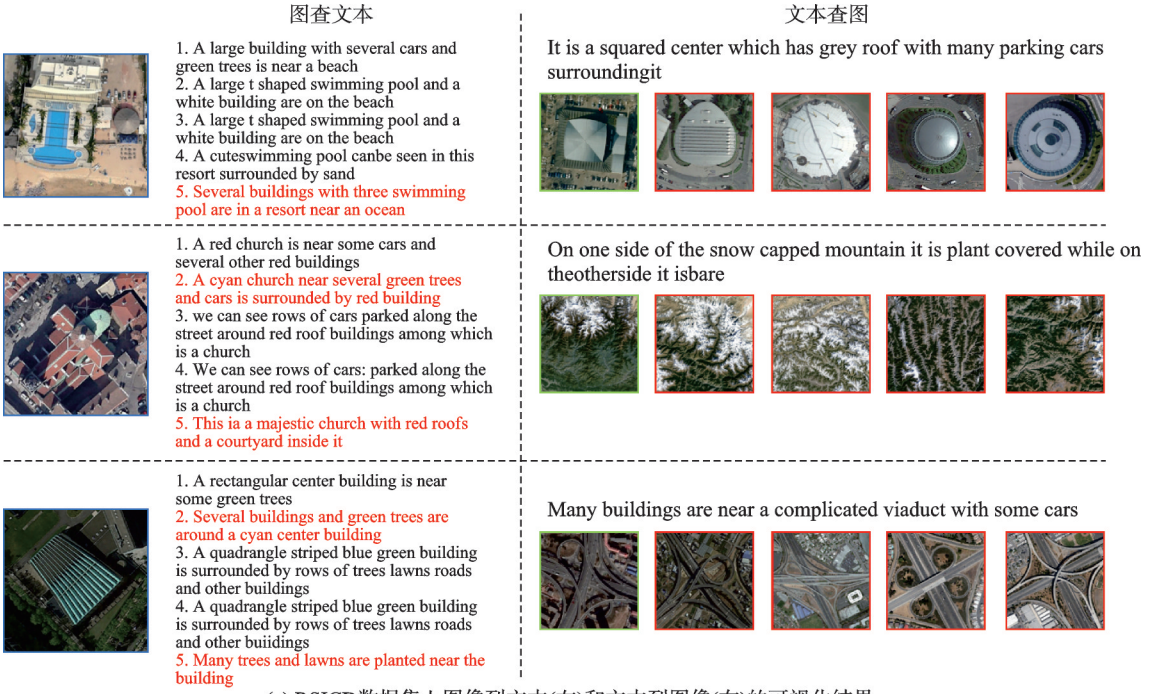
对掩码建模的屏蔽比例的敏感性分析实验。掩码建模用于增强模型的理解能力，通过对图像和文本进行部分屏蔽并重建缺失内容，从而学习更有效的特征表示。由表13～16可以得到，当图片掩码比例为0.75，文本掩码比例为0.15时，mR达到了最大值，可以验证本文实验设置的合理性。同时可以看到，当图片掩码比例达到较大的0.95时，结果变化幅度不大，mR仅较最优值下降1.33%(RSICD)与0.79%(RSITMD)，在文本掩码建模的结果上也可以看到类似的结果，这样的实验结果说明较高的掩码比例并不会导致图像特征过多丢失，模型在掩码建模任务中具有较强鲁棒性。

对α的敏感性分析实验。在推理阶段，本文采用全局相似度s^{g}和局部相似度s^{l}的加权平均作为最终相似度评分，即$=\alpha s^{\mathrm{g}}(1-\alpha)s^{\mathrm{l}}$，其中，$\alpha$控制全局相似度与局部相似度的权重分配，因此其选择对最终检索性能有重要影响。本文将α值从0每步增加0.1到1，可以发现在RSICD数据集上，mR在α为0.7时达到了最大值37.39%，而在RSITMD数据集上，mR在α为0.5时达到了最大值49.53%。且两个数据集上都有在$\alpha=0$时的mR比$\alpha=1$时的mR低，这表明全局相似度在检索中起到了关键的作用，而局部相似度则起到了补充的作用。同时实验结果表明，模型的表现对于α的选择是稳定的，在α的范围变化内性能没有大幅度波动，说明方法具有较强的鲁棒性。

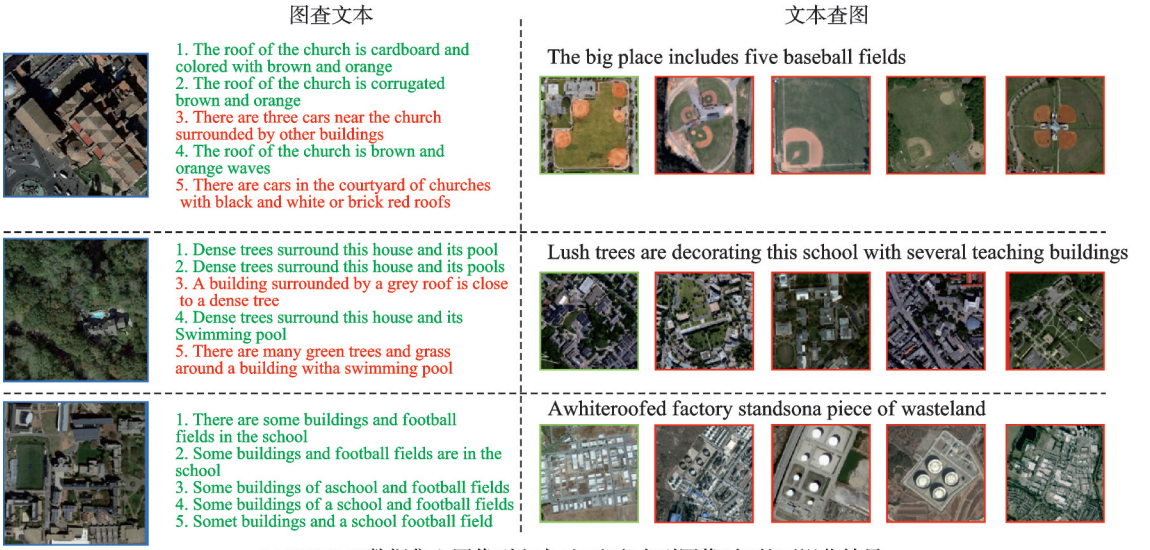
3.4 可视化实验

本节进行两个可视化实验来展示本文方法的有效性。第1个实验是可视化检索的结果，第2个实验本文将短语与图像图像块之间的局部相似度分数进行可视化，表明局部对齐的准确性。

(1) 检索结果可视化。图2显示了前5个检索结果，包括图像查询文本(左)和文本查询图像(右)。查询用蓝色表示，正确的结果用绿色表示，错误的结果用红色表示。检索结果发现，不论是文本还是图像都是非常相似的，这表明遥感图像-文本检索任务十分具有挑战性。同时本文的模型也能够处理其中细微的区别，表明其对全局和局部特征的理解较强。



(a) RSICD数据集上图像到文本(左)和文本到图像(右)的可视化结果
(a) Visualization results on RSICD dataset (image-to-text(left), text-to-image(right))

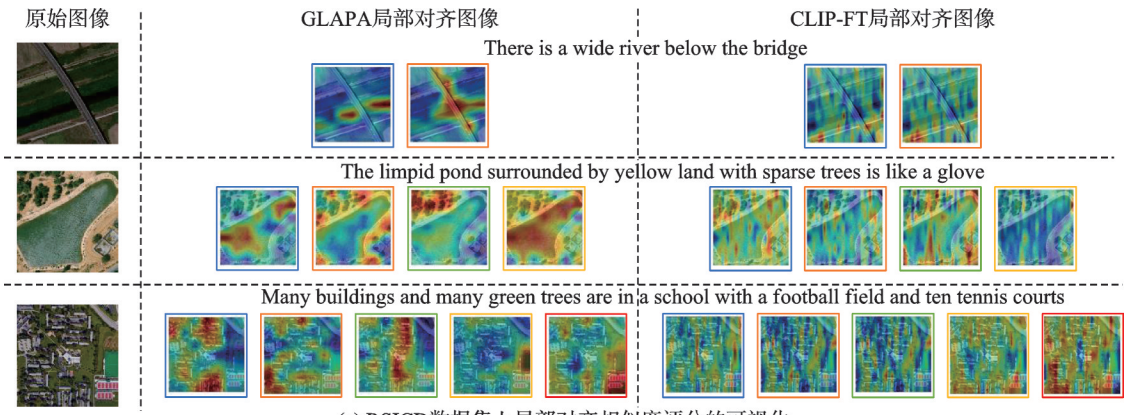


(b) RSITMD数据集上图像到文本(左)和文本到图像(右)的可视化结果
(b) Visualization results on RSITMD dataset (image-to-text(left), text-to-image(right))

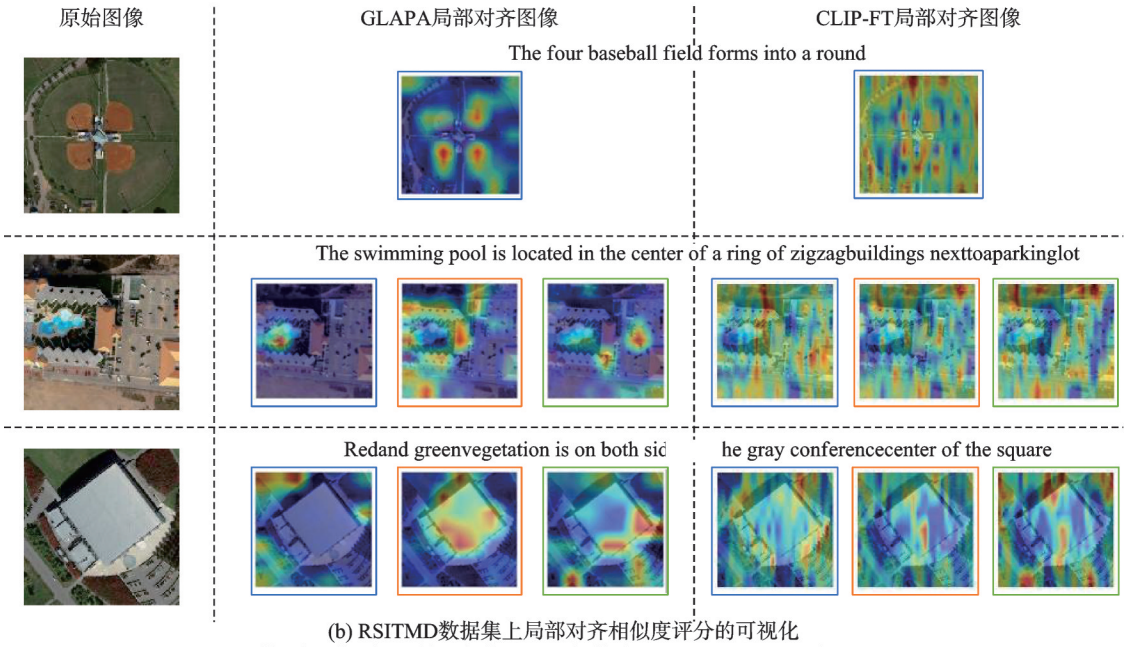
图2 RSICD和RSITMD数据集上图像到文本和文本到图像前5个检索结果的可视化

Fig.2 Visualization of the top 5 image-to-text and text-to-image retrieval results on RSICD and RSITMD datasets

(2)局部相似度可视化。图3展示了短语与图像图像块之间的局部相似度分数的可视化结果。原始图像在左边。右边相同彩色的短语和注意力图一一对应。注意力图中暖色高亮部分表明了短语与其的相似度较高。可以看到,对于每个短语,模型都能较为准确地找到与之对应的图像图像块。表明局部对齐分数能够有效地捕捉短语与图像之间的局部关系,这证明了局部相似度的有效性。



(a) RSICD数据集上局部对齐相似度评分的可视化
(a) Visualization of local alignment similarity score on RSICD dataset



(b) RSITMD数据集上局部对齐相似度评分的可视化
(b) Visualization of local alignment similarity score on RSITMD dataset

图 3 RSICD 和 RSITMD 数据集上局部对齐相似度评分的可视化

Fig.3 Visualization of local alignment similarity score on RSICD and RSITMD datasets

4 结束语

本文提出的GLAPA模型通过全局对齐与局部对齐的结合，有效提升了遥感图像-文本检索的准确性。实验结果证明，GLAPA在多个基准数据集上均超越了现有的最先进方法，尤其在文本细节不足的数据集上表现突出。然而，在部分极端场景下，如单短语文本或多云遮挡图像时，模型仍存在一定局限。此外，当前所用的术语提取方法主要依赖通用语言处理工具如 spaCy，其在遥感专业语境下的适配性仍有待提升。未来的研究将探索融合遥感术语知识库与上下文语义建模的术语提取方法，并进一步优化局部对齐策略，以增强模型在极简文本或遮挡图像等复杂场景下的鲁棒性。

参考文献：

[1] CHI M, PLAZA A, BENEDIKTSSON J A, et al. Big data for remote sensing: Challenges and opportunities[J]. Proceedings

of the IEEE, 2016, 104(11): 2207-2219.

[2] ZHANG W, LI J, LI S, et al. Hypersphere-based remote sensing cross-modal text-image retrieval via curriculum learning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-15.

[3] MA Y, WU H, WANG L, et al. Remote sensing big data computing: Challenges and opportunities[J]. Future Generation Computer Systems, 2015, 51: 47-60.

[4] 李树涛, 马启伟, 王志宇, 等. 多模态对齐在遥感领域的挑战、现状与机遇[J]. 遥感学报, 2025, 29(6): 1551-1565.
LI Shutao, MA Qiwei, WANG Zhiyu, et al. Challenges, current status, and opportunities of multimodal alignment in remote sensing[J]. National Remote Sensing Bulletin, 2025, 29(6): 1551-1565.

[5] ABDULLAH T, BAZI Y, AL RAHHAL M M, et al. TextRS: Deep bidirectional triplet network for matching text to remote sensing images[J]. Remote Sensing, 2020, 12(3): 405.

[6] YUAN Z, ZHANG W, 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, 2021, 60: 1-19.

[7] PAN J, MA Q, BAI C. Reducing semantic confusion: Scene-aware aggregation network for remote sensing cross-modal retrieval[C]//Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. [S.l.]: ACM, 2023: 398-406.

[8] JI Z, MENG C, ZHANG Y, et al. Knowledge-aided momentum contrastive learning for remote-sensing image text retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-13.

[9] MA Q, PAN J, 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.

[10] LI H, XIONG W, CUI Y, et al. A fusion-based contrastive learning model for cross-modal remote sensing retrieval[J]. International Journal of Remote Sensing, 2022, 43(9): 3359-3386.

[11] ZHANG X, LI W, WANG X, et al. A fusion encoder with multi-task guidance for cross-modal text-image retrieval in remote sensing[J]. Remote Sensing, 2023, 15(18): 4637.

[12] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2021: 8748-8763.

[13] ZHANG Z, ZHAO T, GUO Y, et al. RS5M and GeoRSCLIP: A large scale vision-language dataset and a large vision-language model for remote sensing[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-23.

[14] YUAN Y, ZHAN Y, XIONG Z. Parameter-efficient transfer learning for remote sensing image-text retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-14.

[15] JI Z, MENG C, ZHANG Y, et al. Eliminate before align: A remote sensing image-text retrieval framework with keyword explicit reasoning[C]//Proceedings of the 32nd ACM International Conference on Multimedia. [S.l.]: ACM, 2024: 1662-1671.

[16] HUANG T. Efficient remote sensing with harmonized transfer learning and modality alignment[EB/OL].(2024-04-28).https://arxiv.org/abs/2404.18253.

[17] LIU F, CHEN D, GUAN Z, et al. Remoteclip: A vision language foundation model for remote sensing[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-16.

[18] YUAN Z, ZHANG W, TIAN C, et al. Remote sensing cross-modal text-image retrieval based on global and local information [J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-16.

[19] LIAO Y, YANG R, XIE T, et al. A fast and accurate method for remote sensing image-text retrieval based on large model knowledge distillation[C]//Proceedings of IGARSS 2023 IEEE International Geoscience and Remote Sensing Symposium. [S.l.]: IEEE, 2023: 5077-5080.

[20] 张永梅, 徐敏, 李小冬. 基于注意力机制和软匹配的多标签遥感图像检索方法[J]. 计算机应用与软件, 2024, 41(6): 181-185,199.
ZHANG Yongmei, XU Min, LI Xiaodong. A multi-label remote sensing image retrieval method based on attention mechanism and soft matching[J]. Computer Applications and Software, 2024, 41(6): 181-185, 199.

[21] ZHENG F, LI W, WANG X, et al. A cross-attention mechanism based on regional-level semantic features of images for cross-modal text-image retrieval in remote sensing[J]. Applied Sciences, 2022, 12(23): 12221.

[22] YANG R, ZHANG D, GUO Y, et al. A texture and saliency enhanced image learning method for cross-modal remote sensing

image-text retrieval[C]//Proceedings of IGARSS 2023 IEEE International Geoscience and Remote Sensing Symposium. [S.l.]: IEEE, 2023: 4895-4898.

[23] SHUI J, DING S, LI M, et al. Entity semantic feature fusion network for remote sensing image-text retrieval[C]//Proceedings of Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data. [S.l.]: Springer, 2024: 130-145.

[24] PAN J, MA Q, BAI C. A prior instruction representation framework for remote sensing image-text retrieval[C]//Proceedings of the 31st ACM International Conference on Multimedia. [S.l.]: ACM, 2023: 611-620.

[25] 张若愚, 聂婕, 宋宁, 等. 基于布局化-语义联合表征遥感图文检索方法[J]. 北京航空航天大学学报, 2024, 50(2): 671-683.
ZHANG Ruoyu, NIE Jie, SONG Ning, et al. Remote sensing image-text retrieval based on layout semantic joint representation[J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(2): 671-683.

[26] HAO X, LIU L, YANG R, et al. A review of data augmentation methods of remote sensing image target recognition[J]. Remote Sensing, 2023, 15(3): 827.

[27] LALITHA V, LATHA B. A review on remote sensing imagery augmentation using deep learning[J]. Materials Today: Proceedings, 2022, 62: 4772-4778.

[28] 田澍, 张秉熙, 曹林, 等. 基于目标语义提示与双注意力感知的遥感图像文本检索方法[J]. 电子与信息学报, 2025, 47(6): 1734-1746.
TIAN Shu, ZHANG Bingxi, CAO Lin, et al. Remote sensing image text retrieval method based on object semantic prompt and dual-attention perception[J]. Journal of Electronics & Information Technology, 2025, 47(6): 1734-1746.

[29] LU X, WANG B, ZHENG X, 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.

[30] CHERTI M, BEAUMONT R, WIGHTMAN R, et al. Reproducible scaling laws for contrastive language-image learning [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 2818-2829.

[31] LOSHCHILOV I. Decoupled weight decay regularization[EB/OL]. (2017-11-14). https://arxiv.org/abs/1711.05101.

[32] ZHANG J, WANG L, ZHENG F, et al. An enhanced feature extraction framework for cross-modal image-text retrieval[J]. Remote Sensing, 2024, 16(12): 2201.

[33] ZHENG F, WANG X, WANG L, et al. A fine-grained semantic alignment method specific to aggregate multi-scale information for cross-modal remote sensing image retrieval[J]. Sensors, 2023, 23(20): 8437.

作者简介:

金开宇(2001-),男,硕士研究生,研究方向:因果表征学习、领域泛化和计算机视觉,E-mail: proton00@mail.ustc.edu.cn。

胡筱(2002-),女,硕士研究生,研究方向:大模型智能体理论,E-mail: xh0421@mail.ustc.edu.cn。

谢洪(1987-),男,特任教授,研究方向:线学习算法及其应用。

连德富(1985-),**通信作者**,男,教授,研究方向:大模型检索增强、大模型智能体、可信人工智能、推荐系统以及时序大模型等,E-mail: liandefu@ustc.edu.cn。

(编辑:刘彦东)

Global and Local Alignment with Phrase Augmentation for Remote Sensing Image-Text Retrieval

JIN Kaiyu, HU Xiao, XIE Hong, LIAN Defu*

(Laboratory of Cognitive Intelligence (School of Computer Science and Technology, University of Science and Technology of China), Hefei 230088, China)

Abstract: Remote sensing image-text retrieval has received increasing attention in recent years. Most existing methods rely on aligning global features between images and texts to perform retrieval, with a strong focus on feature extraction and fusion. However, they typically overlook the alignment between fine-grained object regions in images and semantic phrases in texts (i.e., local alignment). This limitation is particularly problematic in remote sensing, where intra-modality samples often exhibit high visual similarity, confusing global alignment. To address this, we propose GLAPA (Global and local alignment with phrase augmentation), a novel framework that complements global alignment with fine-grained local alignment. Specifically, GLAPA extracts conceptual phrases from textual descriptions and dynamically aligns them with relevant image patches, enhancing the model's ability to capture detailed cross-modal semantics. In addition, we incorporate masked modeling to strengthen intra-modal feature learning, further improving retrieval performance. Extensive experiments on the RSICD and RSITMD datasets—overing ablation studies, hyperparameter analysis, and visualization—demonstrate that GLAPA significantly outperforms state-of-the-art methods. Visualization results also confirm the effectiveness of our local alignment strategy.

Highlights:

1. A phrase-enhanced global and local alignment framework is proposed for remote sensing image-text retrieval.
2. Dynamic phrase-region alignment improves fine-grained cross-modal matching.
3. Masked vision and language modeling enhance intra-modal representation learning.
4. State-of-the-art performance is achieved on RSICD and RSITMD datasets.

Key words: image-text retrieval; remote sensing; feature alignment; local alignment; phrase augmentation

Foundation item: National Key R&D Program of China (No.2021ZD0111800).

Received: 2025-03-01; **Revised**: 2025-09-08

***Corresponding author**, **E-mail**: liandefu@ustc.edu.cn.