基于 MFF-SFE 的遥感图文跨模态检索方法*

钟金彦[1,2]，陈俊[1,3,4]，李宇[1†]，吴业炜[1]，葛小青[1]

(1 中国科学院空天信息创新研究院，北京，100094；2 中国科学院大学电子电气与通信工程学院，北京，100049；3 中国科学院计算机网络信息中心，北京，100083；4 中国科学院大学计算机科学与技术学院，北京，100049)

（2024 年 1 月 8 日收稿；2024 年 4 月 17 日收修改稿）

钟金彦，陈俊，李宇，等. 基于 MFF-SFE 的遥感图文跨模态检索方法[J]. 中国科学院大学学报，DOI:10. 7523/j.ucas. 2024. 025.

摘　要 遥感图文跨模态检索技术能够从海量的遥感数据中快速获取有价值的信息，但现有遥感图文检索方法对遥感图像中的多尺度信息利用不足、目标信息识别效果不佳，检索精度相对较低。为此，本文提出了一种新的遥感图文跨模态检索方法。该方法主要包括一个多尺度特征融合（multi-scale feature fusion, MFF）模块和一个显著特征增强（salient feature enhancement，SFE）模块，分别用于融合遥感图像的多尺度信息、加强对遥感图像目标信息的表达能力，从而提高遥感图文跨模态检索精度。本文在两个公开的遥感图像文本数据集上进行了实验验证，实验结果表明本文提出的方法在遥感图文跨模态检索任务中大部分评价指标都优于其他方法，具有最佳的总体检索性能。

关键词 跨模态检索；遥感图像；深度学习；多尺度特征

中图分类号：TP751.1　　**文献标志码：**A　　**DOI:**10. 7523/j.ucas. 2024. 025

Cross-modal retrieval method based on MFF-SFE for remote sensing image-text

ZHONG Jinyan[1,2], CHEN Jun[1,3,4], LI Yu [1], WU Yewei[1], GE Xiaoqing[1]

(1 *Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing* 100094, *China*; 2 *School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing* 100049, *China;* 3 *Computer Network Information Center, Chinese Academy of Sciences, Beijing* 100083, *China;* 4 *School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract Remote sensing image-text cross-modal retrieval technology can quickly obtain valuable information from massive remote sensing data. However, existing remote sensing image-text retrieval methods have limitations in utilizing multi-scale information within remote sensing images, and the weak recognition of target information leads to relatively low retrieval accuracy. To address these issues, this paper proposes a new method for remote sensing image-text cross-modal

* 中科院青年促进会（E0331804）资助

† 通信作者，E-mail: liyu202615@aircas.ac.cn

retrieval. This method mainly comprises a multi-scale feature fusion (MFF) module and a salient feature enhancement (SFE) module, which are designed to integrate multi-scale information of remote sensing images and enhance the expression of target information in remote sensing images, so as to improve the precision of remote sensing image-text cross-modal retrieval. Experimental validation was conducted on two publicly available remote sensing image-text datasets. The results demonstrate that the proposed method outperforms other methods across most evaluation metrics in the remote sensing image-text cross-modal retrieval task, and exhibits the best overall retrieval performance.

Keywords cross-modal retrieval; remote sensing images; deep learning; multi-scale feature

随着对地观测能力的飞速发展，当前获取的遥感数据量呈指数增长，遥感数据呈现多元化、海量化趋势，遥感对地观测进入了大数据时代（NASA 的地球观测卫星每天向地球传输 TB 级的数据，Sentinel 系列卫星在观察期间每天接收超过 10 TB 规模的数据）。遥感大数据的“数据海量、信息淹没”问题[1]日益突出，如何设计一种有效的遥感图像的检索方法来从海量的遥感数据中快速获得需要的数据、提高遥感数据的管理和使用效率，已经成为越来越多研究者关注的焦点。

根据查询数据模态的差异，遥感图像检索技术可分为单模态检索和跨模态检索[2]，这两者的主要区别在于进行查询的数据的模态和待检索的数据的模态是否相同，如果相同则为单模态检索，反之则为跨模态检索。对于遥感图像检索而言，单模态检索是指查询和检索的遥感数据都是同一种类型的遥感图像，例如，Liu 等[3]使用深度对抗哈希在光学遥感图像中进行检索，Ye 等[4]利用卷积神经网络（convolutional neural network，CNN）来学习合成孔径雷达图像和光学遥感图像之间的域不变特征，从而对合成孔径雷达图像进行检索。在实际应用中，由于数据模态存在多样性，往往需要利用其他模态的数据（如文本、语音等）而不仅仅是同一种类型的遥感图像，来检索遥感图像，例如 Guo 等[5]提出了一种深度视觉-音频网络，使用音频直接检索遥感图像，该网络基于预训练的 CNN 和深度音频网络，并使用神经网络来对语音特征和遥感图像特征进行融合与分类。与遥感图像的单模态检索相比，由于查询与检索数据分布空间不同导致的“异构鸿沟”问题，遥感图像的跨模态检索，尤其是遥感图像与文本间的跨模态检索，仍面临不少挑战。

早期的遥感图像文本检索是以人工标注文本或关键词的方式进行的，即由人工预先对每幅遥感图像进行文本注释，然后通过比较预定义文本注释与输入文本之间的相似性进行检索。这种方式需要花费大量人力进行人工标注，检索效率有限，且对标注人员的专业性有一定要求，无法应对遥感图像数量的快速增长带来的挑战。因此越来越多的研究者开始研究遥感图像描述的生成，例如 Shi 和 Zhou[6]采用全卷积网络构建遥感图像描述框架，对遥感图像生成相应的文本描述。这种基于生成遥感图像描述的方法成功解决了人工标注的成本问题，然而，这种将遥感图像与文本分别进行处理的检索方式仍会受到“异构鸿沟”的影响，导致检索精度较低。如何构建不同模态信息间的相似性度量模型，以解决两种模态信息之间相似性难以直接度量的问题，是实现遥感图像文本跨模态检索的关键。

近些年来，随着多模态数据的快速增长以及深度学习技术的持续发展，自然图像领域对跨模态检索问题的研究已经取得了丰富的成果[7–9]。同时，在遥感领域，也有越来越多的研究者开始探索基于深度学习的遥感图文跨模态检索问题[10–13]，其主要可以分为基于语义对齐的方法和基于多尺度信息增强的方法两类。

基于语义对齐的方法着力于挖掘遥感图像与文本之间的潜在对应关系，通过语义对齐将图像信息与文本信息相对应，加强遥感图像和文本之间的语义关系，从而提升跨模态检索的精度。Cheng 等[14]设计了一个深度语义对齐网络，采用注意力机制来增强图像文本间的对应关系，并通过门函数来过滤不必要的信息，获得具有辨别力的视觉特征。Zheng 等[15]采用交叉注意力机制组合语句级文本信息和区域级图像信息，实现跨模态信息的交互。Tang 等[16]提出了一种交互增强特征 Transformer，使用特征嵌入模块同时处理视觉特征和文本特征来减少两种模态的语义不一致性，并通过信息交互增强模块来进行跨模态信息交互。尽管基于语义对齐的方法能充分挖掘图像与文本之间的深层关系，提高模态间的相似性，但由于遥感图像覆盖范围广、目标信息不突出，在进行语义对齐时往往会受到冗余信息的影响，进而影响检索的准确性。

基于多尺度信息增强的方法考虑到遥感图像的多尺度特性，主要关注如何更好地提取遥感图像的多尺度信息，通过更精准的遥感图像特征来提高检索效果。Yuan 等[17]提出了一种非对称多模态特征匹配网络，利用多尺度视觉自注意力模块提取遥感图像的显著特征。为了减小模型的参数量，Wang 等[18]设计了一个轻量化的多尺度探索模块，将深度卷积和扩展卷积相结合，以较小的成本挖掘多尺度信息。张若愚等[19]则针对遥感图像目标远距离建模困难的问题，基于 Transformer 编码器进行视觉的空间布局化建模，构建了主导语义监督下的布局化视觉特征提取模块来提取遥感图像中的显著目标。

尽管当前已存在不少基于多尺度信息增强的遥感图文跨模态检索方法的研究，但仍存在以下问题：

（1）现有的遥感图文跨模态检索方法对遥感图像中的多尺度信息利用不足，仍无法充分提取这些信息，在一定程度上影响了遥感图像-文本检索任务的精度。

（2）现有方法对遥感图像中的目标信息识别效果不佳。当前的方法大多聚焦于如何在遥感图像和文本间进行信息的交互，而忽略了对遥感图像中冗余特征的过滤，难以提取显著性特征，阻碍了其对遥感图像内容的理解。

针对现有方法多尺度信息利用不足的问题，本文设计了一个更有效的多尺度特征融合（multi-scale feature fusion，MFF）模块，融合遥感图像的低、中、高层特征，充分利用遥感图像的多尺度信息，提升检索精度；同时，为了解决现有方法对遥感图像中的目标信息识别效果不佳问题，本文构建了一个显著特征增强（salient feature enhancement，SFE）模块，使用多尺度特征中的低层视觉特征对 MFF 模块提取的、具有多尺度信息的融合特征进行增强，加强对遥感图像目标信息的表达能力，同时通过多尺度信息尽可能地提高模型对较小目标的识别能力；最终提出了一种基于 MFF-SFE 的遥感图文跨模态检索方法，通过 MFF 模块和 SFE 模块相结合来得到更加精确和全面的遥感图像特征，从而提高模型的检索准确性。

1 基于 MFF-SFE 的遥感图文跨模态检索网络

本文提出的基于 MFF-SFE 的遥感图文跨模态检索网络采用 Inception Resnet V2[20]和双向编码表示变换器（bidirectional encoder representations from transformers，BERT）[21]分别提取遥感图像特征和文本特征，并通过 MFF 模块和 SFE 模块增强遥感图像特征，最后使用余弦相似度函数计算遥感图像特征与文本特征之间的相似度。网络主要包括四个部分：遥感图像/文本特征提取、MFF 模块、SFE 模块和相似性度量，整体结构如图 1 所示。

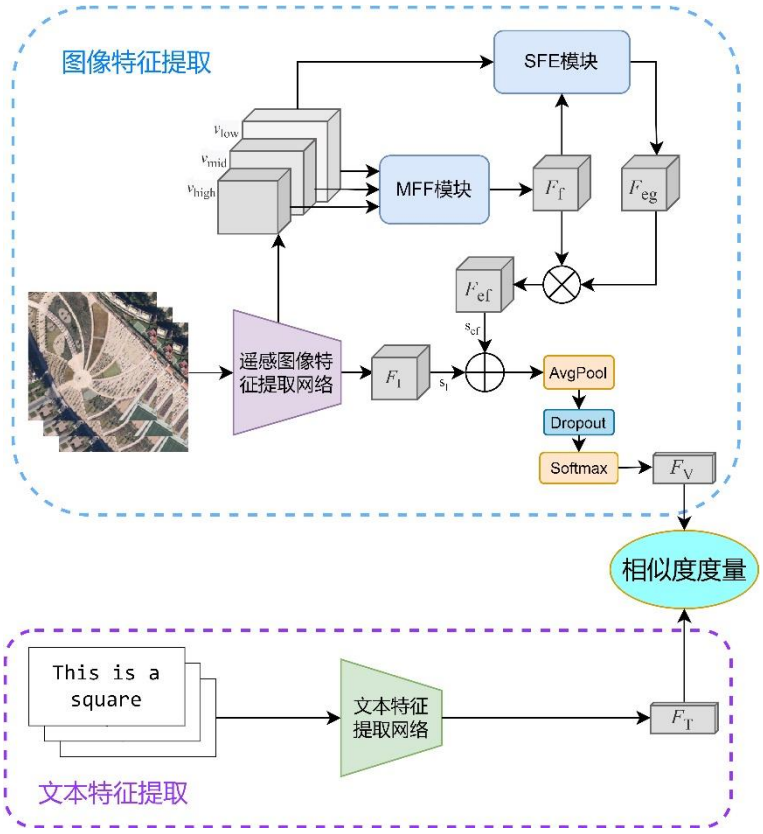

图 1 基于 MFF-SFE 的遥感图文跨模态检索网络结构

Fig. 1 Network structure for remote sensing image-text cross-modal retrieval based on MFF-SFE

首先，在遥感图像特征提取部分，本文采用 Inception Resnet V2 作为遥感图像特征提取网络，获得遥感图像特征向量 F_{I}，同时提取遥感图像的多尺度特征 v_{low}、v_{mid}、v_{high}，并将多尺度特征与低层视觉特征 v_{low} 分别输入 MFF 模块和 SFE 模块；再将增强后的融合特征 F_{ef} 与 F_{I} 相加并降维，得到最终的图像特征 F_{V}。其次，在文本特征提取部分，本文使用 BERT 作为文本特征提取网络来获得指定维度的文本特征 F_{T}，此处每条文本语句都对应一个文本特征。

同时，为了充分利用遥感图像中的多尺度信息、解决遥感图像的目标信息不突出问题，本文分别设计了 MFF 模块与 SFE 模块。MFF 模块使用遥感图像特征提取网络提取的多尺度特征，并分别进行卷积操作再相连，然后采用通道注意力机制自适应地优化各个通道的权重，得到具有多尺度信息的融合特征 F_{f}。SFE 模块将 MFF 模块获得的融合特征 F_{f} 作为基准特征，并将低层视觉特征 v_{low} 作为强化特征，通过卷积等处理得到显著信息特征门向量 F_{eg}，从而进一步优化融合特征 F_{f}，获得能够突出遥感图像目标信息的显著信息特征 F_{ef}。

最后，将显著信息特征 F_{ef} 与遥感图像特征提取网络的输出 F_{I} 相加，得到最终的图像特征 F_{V}，并对图像特征 F_{V} 和文本特征 F_{T} 进行相似性度量，最终实现更为精准的遥感图文跨模态检索。

1.1 图像与文本特征提取

本小节的特征提取环节包含两个部分，分别是图像特征提取和文本特征提取。下面将分别对这两种特征提取进行详细介绍。

（1）图像特征提取

定义输入遥感图像为 $I \in \mathbb{R}^{H \times W \times C}$，通过遥感图像特征提取网络提取遥感图像的多尺度特征与图像特征

$F_{\mathrm{I}} \in \mathbb{R}^{H_h \times W_h \times C_h}$，其中，多尺度特征包括低层视觉特征 $v_{\mathrm{low}} \in \mathbb{R}^{H_l \times W_l \times C_l}$、中层视觉特征 $v_{\mathrm{mid}} \in \mathbb{R}^{H_m \times W_m \times C_m}$ 和高层视觉特征 $v_{\mathrm{high}} \in \mathbb{R}^{H_h \times W_h \times C_h}$。然后将其输出特征 F_{I} 与增强后的显著信息特征 F_{ef} 分别乘以可变系数 s_{I}、s_{ef}，通过平均池化层、失活率为 0.8 的 Dropout 层和 Softmax 层后，得到最终的图像特征 F_{V}，如图 1 的图像特征提取部分所示。

本文使用 Inception Resnet V2 作为模型的遥感图像特征提取网络，并分别选取其中 Stem 块、Reduction-A 块和 Reduction-B 块的输出作为多尺度特征中的低层、中层和高层视觉特征 v_{low}、v_{mid} 和 v_{high}，其结构如图 2 所示。

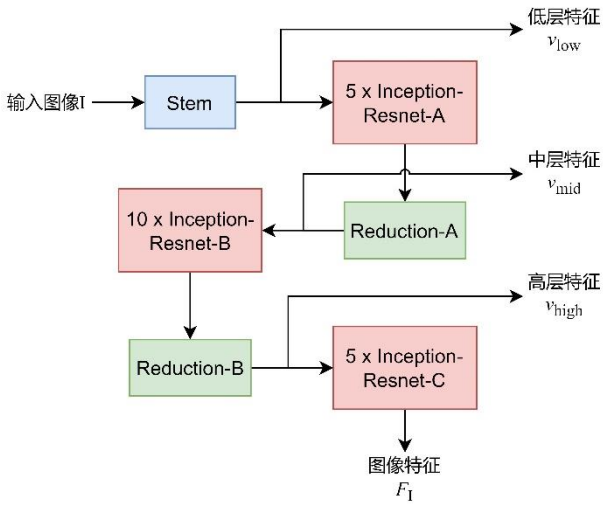

图 2 遥感图像特征提取网络结构

Fig. 2 Structure of the remote sensing image feature extraction network

（2）文本特征提取

对于输入的描述语句 S，本文使用 BERT 预训练的文本编码器进行分词和词向量编码，得到对应词块的词编码向量 T_{emd}，再与编码器提取的掩码向量一并输入 BERT 预训练模型，获得文本特征。考虑到 BERT 模型生成的文本特征都是 768 维的向量，本文通过全连接层将这些文本特征转换为与图像特征 F_{V} 相同维度的向量，从而得到最终的文本特征向量 F_{T}，整个文本特征提取网络的结构如图 3 所示。

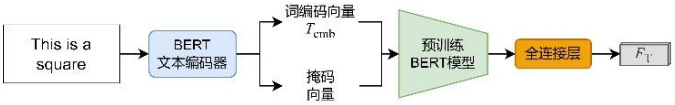

图 3 文本特征提取网络结构

Fig. 3 Structure of the text feature extraction network

1.2 MFF 模块

与自然图像相比，遥感图像往往包含了丰富的目标信息。对于较大的目标，可以使用全局特征来表达。然而，对于较小的目标，随着卷积网络层数的增加，其在遥感图像特征中的信息可能会逐渐减少甚至消失。为了更好地提取遥感图像中的多尺度信息，并保留其中较小目标的特征信息，本文构建了如图 4 所示的 MFF 模块。该模块对多尺度特征进行融合，并采用通道注意力机制来增强多尺度信息处理能力。

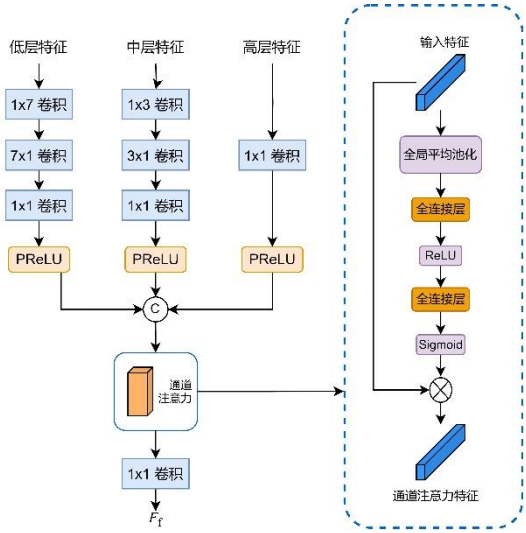

图 4 MFF 模块结构

Fig. 4 Structure of the MFF module

由于遥感图像特征提取网络所获得的多尺度特征 v_{low}、v_{mid} 和 v_{high} 维度不统一，模块先对 v_{low} 和 v_{mid} 分别进行一次 7×7 和 3×3 的卷积（步长分别为 4、2）。为了降低计算量与参数量，此处将 7×7 和 3×3 卷积分别拆分为 1×7、7×1 卷积的串联和 1×3、3×1 卷积的串联，其步长也分别变为（1，4）、（4，1）和（1，2）、（2，1）。然后，为了增强网络的非线性表达能力，模块对 v_{high} 及降维后的 v_{low}、v_{mid} 都进行 1×1 卷积和参数化修正线性单元（图 4 中的 PReLU）处理。最后将三者拼接在一起，

得到初步的融合特征v_{f}。

由于v_{f}中不同通道对于多尺度信息的表达能力不同，本文对初步融合特征v_{f}使用了通道注意力，自适应地为每个通道分配权重，从而提高 MFF 模块对多尺度信息的利用能力。同时，为了降低特征维度，也便于后续处理，本文使用 1×1 卷积把通道数降低到与v_{high}相同，最终得到融合特征F_{f}。

1.3 SFE 模块

由于遥感图像中目标的复杂性，融合特征F_{f}中仍然存在冗余。这导致目标信息无法突出，因此需要进一步优化融合特征，过滤其中的无关和冗余信息，从而更好地突出遥感图像中的目标信息。

为此，本文设计了如图 5 所示的 SFE 模块。考虑到高层视觉特征往往难以保留遥感图像中较小目标的信息，本文将低层视觉特征v_{low}作为强化特征，并将 MFF 模块提取的融合特征F_{f}作为基准特征，使用处理过的强化特征对基准特征进行元素相乘，生成含有目标信息的显著信息特征门向量。然后，通过卷积和最大池化来增强显著性特征并降低空间维度，得到最终的显著信息特征门向量F_{eg}。

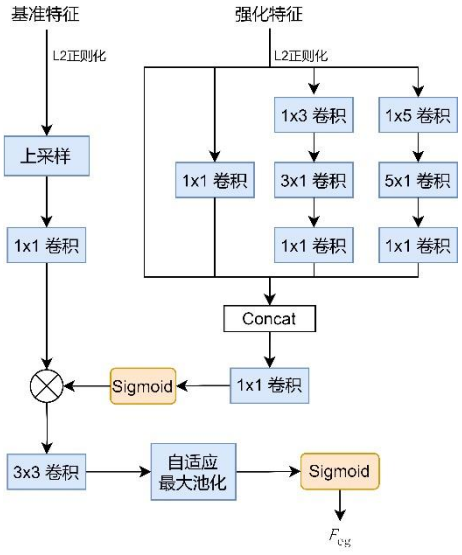

图 5 SFE 模块结构

Fig. 5 Structure of the SFE module

具体来讲，为了提升强化特征对基准特征的增强效果，SFE 模块首先通过四个不同的卷积分支处理强化特征，然后将这四个分支的输出进行拼接，并使用 1×1 卷积降维，从而得到具有更多信息的强化特征，再使用 Sigmoid 激活函数抑制强化特征中的无用信息。另一方面，为了统一特征维度，模块对基准特征F_{f}进行上采样和 1×1 卷积，再将其与处理过的强化特征相乘，从而得到具有目标信息的显著信息特征门向量。最后，模块使用 3×3 卷积和自适应最大池化来保证显著信息特征门向量与基准特征F_{f}的维度一致，并进一步增强特征中的显著信息并过滤冗余信息，从而得到最终的显著信息特征门向量F_{eg}。

1.4 相似性度量与损失函数

为了在统一的特征空间里度量不同模态的信息，本文使用余弦相似度来计算遥感图像特征与文本特征的相似性，并采用交叉熵损失函数优化训练检索模型。

本文将每张遥感图像设为一个独立类别，并将与之对应的文本归为同一类别，即认为图像及其对应的文本在共同的特征空间中应具有相同的特征表示。余弦相似度和交叉熵损失函数的计算公式如下所示：

$$\cos(F_{\mathrm{V}},F_{\mathrm{T}})=\frac{F_{\mathrm{V}}F_{\mathrm{T}}}{\|F_{\mathrm{V}}\|_2\|F_{\mathrm{T}}\|_2}, \tag{1}$$

$$L=-\frac{1}{N}\sum_{i=1}^{N}\ln\frac{e^{x_{i,y_i}}}{\sum_{j=1}^{n}e^{x_{i,j}}}, \tag{2}$$

上式中：$\|\bullet\|_2$表示计算向量的 2-范数，i表示特征索引，N表示批次大小，j表示类别索引，n表示类别总数，$x_{i,j}$表示第i个样本的与第j个类别的相似度，y_i表示样本i的真实类别。

当图像特征F_{V}和文本特征F_{T}完全相同时，即$F_{\mathrm{V}}=F_{\mathrm{T}}$时，由式（1）可得，此时$F_{\mathrm{V}}$和$F_{\mathrm{T}}$余弦相似度等于 1；相反，若$F_{\mathrm{V}}$和$F_{\mathrm{T}}$相差越大，则两者的余弦相似度将越趋近于-1。

将同一批次内所有图像特征与文本特征间的余弦相似度作为$x_{i,j}$输入交叉熵损失函数中，结合每个样本的实际类别y计算第i个样本与其真实类别y_i之间的相似度x_{i,y_i}。由式（2）可见，在第i个样本时，若样本越能正确地与类别相匹配，也就是与真实类别的相似度x_{i,y_i}越高、与其他错误类别的相似度$x_{i,j(j\neq y_i)}$越低，则这

个样本的负对数似然越小；若在批次内正确匹配的样本数越多，则所有样本的负对数似然的平均数越小，也就是交叉熵损失 L 越小（趋近于 0），反之 L 越大。

本文将图像和文本间的相似度矩阵作为图像-文本相似度，计算图像检索文本任务的交叉熵损失；同时，将以上相似度矩阵的转置作为文本-图像的相似度，计算文本检索图像任务的交叉熵损失；最后取两个损失的均值作为本文模型的损失函数。

2 实验结果及分析

2.1 数据集

为验证本文方法的有效性并评估本文模型的性能，本文在两个公开的遥感图像文本数据集 UCM-Captions[22]和 Sydney-Captions[22]上分别进行了对比实验与消融实验。两个数据集中每张图像都对应 5 句相关的文本描述。其中，UCM-Captions 数据集由 21 类场景的 2100 幅遥感图像和 10500 句文本描述组成，每幅图像的大小为 256×256，图像分辨率为 1 英寸；Sydney-Captions 数据集中包含 613 幅 500×500 的遥感图像和 3065 句文本描述，图像分辨率为 0.5m。两个数据集的部分样例如图 6 所示。

(1) It is a medium residential area .

(2) A medium residential area with houses sorrounded by plants .

(3) A medium residential area with a narrow road goes through this area .

(4) Some houses with plants surrounded in the medium residential area .

(5) This is a medium residential area with houses and plants and a narrow road goes through .

（a）UCM-Captions

(1) A wide river with deep green waters and some boats on it .

(2) There are some small boats on the wide river with some houses on the bank .

(3) Some small boats on the river with some houses on the bank .

(4) There are some white boats on the river with some houses on the bank .

(5) A wide river with deep green waters and some boats on it .

（b）Sydney-Captions

图 6 数据集中遥感图像与描述

Fig. 6 Remote sensing images and descriptions in dataset

2.2 实验环境及评价指标

本文所有实验都是基于 Python3.7 及深度学习框架 PyTorch1.8.0 和 CUDA11.2 实现，并在配备 Inter(R) Core(TM) i9-10920X、GeForce RTX 3090 和 128G RAM 的工作站上运行，操作系统为 CentOS 7.9。

在训练过程中，本文将遥感图像先缩放成 278×278 像素，旋转 90° 后再进行中心裁剪成 256×256 像素，从而增强训练样本；测试时则将遥感图像统一缩放成 256×256 像素。

本文以 8：1：1 的比例将数据集划分为训练集、测试集和验证集，批次大小设置为 64，使用 Adam 优化器进行网络训练，迭代次数为 100，初始学习率设为 0.0001，beta1 和 beta2 分别设为 0.9 和 0.98，权重衰减设为 0.2。在训练过程中，每 20 次迭代进行一次学习率衰减，衰减因子为 0.7。同时，将图像特征维度和文本特征维度都设置成 512。

为了充分验证本文方法的有效性并评估模型的性能，本文采用召回率 R@K（K=1，5，10）和平均召回率 mR 作为评价指标。R@K 表示在返回的检索结果中，按相似度从大到小排序，前 K 个排序结果中出现正确样本的概率；mR 则代表图像检索文本和文本检索图像两个任务中所有 R@K 的平均值。R@K 和 mR 的值越高表

明模型的检索效果越好，反之越差，其计算公式如下：

$$\mathrm{RP_K} = \bigvee_{i=1}^{\mathrm{K}} \mathrm{R}_i , \tag{3}$$

$$\mathrm{R@K} = \frac{1}{M}\sum^{M} \mathrm{RP_K} , \tag{4}$$

$$\mathrm{mR} = \frac{1}{N}\sum_{K=1}^{N} \mathrm{R@K} , \tag{5}$$

上式中：R_i 表示相似度最高的前 K 项检索结果中，第 i 个检索是否正确，正确则为 1，反之为 0；$\bigvee^{\mathrm{K}}$ 表示对 K 个 0/1 的项进行逻辑或。$\mathrm{RP_K}$ 则代表前 K 项检索中是否存在正确的检索结果，当 $\mathrm{RP_K}=1$ 时，前 K 个返回结果中存在与查询样本相关的结果，反之，当 $\mathrm{RP_K}=0$ 时则不存在。M 表示总共进行了 M 次检索；N 表示两个跨模态检索任务的 R@K 指标总数。

2.3 实验结果分析

2.3.1 对比实验

为了评估本文提出的基于 MFF-SFE 的遥感图文跨模态检索模型的检索性能，本文选取了以下六种跨模态图文检索方法进行对比：

（1）VSE++[23]：VSE++模型是自然图像领域图像文本检索的先驱。该方法使用卷积网络和循环网络将图像信息和文本信息嵌入到同一空间中，并提出了三元组损失来训练图像-文本匹配模型。

（2）SCAN[24]：SCAN 模型在 VSE++的基础上，利用 Faster RCNN[25]提取图像特征，并尝试将图像中的目标与文本中的目标对齐。

（3）CAMP[26]：CAMP 模型提出了一种自适应信息传递方法，该方法能自适应地控制跨模态信息传递的信息流，并使用融合特征计算图像和文本的相似度。

（4）MTFN[27]：MTFN 模型采用秩分解的方式设计了多模态融合网络，以此计算嵌入特征的距离。

（5）CLIP[9]：CLIP 模型通过对大量图文数据进行对比学习，使用余弦相似度将图像与文本映射至统一特征空间内，其在众多领域中的零样本跨模态任务上表现出了显著优势。

（6）AMFMN[17]：AMFMN 模型设计了一个多尺度视觉自注意力模块来提取遥感图像的显著特征，并定义了一个动态可变边界的损失函数来解决样本对匹配边界问题。

本文所提方法与上述方法在两个数据集的两个检索任务上分别计算了 R@1、R@5、R@10，并计算了 R@K 的平均值 mR 共七个评价指标。其中，由于遥感图像中存在着较小的目标，并为了降低计算复杂度，本文选择 CLIP 预训练模型中的 ViT-B/32 进行对比实验。实验结果如表 1 和表 2 所示，其中加粗部分表示该列下最好的结果。

表 1 UCM-Captions 数据集对比实验效果

Table 1 Results of comparative experiment on UCM-Captions

方法	图像检索文本			文本检索图像			mR%
	R@1%	R@5%	R@10%	R@1%	R@5%	R@10%	
VSE++	12.38	44.76	65.71	10.10	31.80	56.85	36.93
SCAN t2i	14.29	45.71	67.62	12.76	50.38	77.24	44.67
SCAN i2t	12.85	47.14	69.52	12.48	46.86	71.71	43.43
CAMP	14.76	46.19	67.62	11.71	47.24	76.00	43.92
MTFN	10.47	47.62	64.29	14.19	52.38	78.95	44.65
CLIP	23.33	49.81	61.71	23.77	61.39	77.90	49.65
AMFMN	16.67	45.71	68.57	12.86	53.24	79.43	46.08
本文	**27.62**	**55.24**	**73.33**	**25.81**	**63.62**	**81.43**	**54.51**

表 2 Sydney-Captions 数据集对比试验效果

Table 2 Results of comparative experiment on Sydney-Captions

方法	图像检索文本			文本检索图像			mR%
	R@1%	R@5%	R@10%	R@1%	R@5%	R@10%	
VSE++	24.14	53.45	67.24	6.21	33.56	51.03	39.27
SCAN t2i	18.97	51.72	74.14	17.59	56.90	76.21	49.26
SCAN i2t	20.69	**55.17**	67.24	15.52	57.59	76.21	48.74
CAMP	15.52	51.72	72.41	11.38	51.72	76.21	46.49
MTFN	20.69	51.72	68.97	13.79	55.51	77.59	48.05
CLIP	16.13	46.77	67.74	18.71	59.68	78.39	47.90
AMFMN	24.14	51.72	**75.86**	14.83	56.55	77.89	50.17
本文	**24.19**	51.61	72.58	**20.00**	**65.48**	**84.19**	**53.01**

从表 1 可以得出，本文方法在 UCM-Captions 数据集上所有评价指标都取得了最好的检索结果。与对比方法中总体性能最好的 CLIP 相比，本文方法在图像检索文本任务中 R@1、R@5 和 R@10 分别提升了 4.29%、5.43%和 11.62%，在文本检索图像任务中召回率分别提升了 2.04%、2.23%和 3.53%；两个任务所有指标的平均值 mR 总体提升 4.86%。CLIP 的 ViT-B/32 预训练模型先将图像分割成 32×32 的小块，再通过 ViT（Vision Transformer）进行处理，能够有效捕获图像中的上下文信息与细节信息，但仍然无法有效地过滤遥感图像中的冗余特征，因此难以提取其中的显著性特征。而本文采用 Inception Resnet V2 进行特征提取，并通过 SFE 模块，使用低层视觉特征对融合特征进行增强，尽可能地保留遥感图的显著特征，同时过滤冗余特征，因而取得了较好的效果。

由表 2 可见，本文方法在 Sydney-Captions 数据集上的大部分评价指标都优于其他方法，在文本检索图像任务中三个召回率分别比对比方法中最好的 AMFMN 方法提升了 5.17%、8.93%和 6.3%。不过，在图像检索文本任务中，本文提出的模型在 R@1 上几乎与 AMFMN 相同，在 R@5 和 R@10 略低于 AMFMN 和 SCAN i2t。这可能与 Sydney-Captions 数据集自身数据类别分布不均衡存在着一定的关系。但在总体性能上，本文的方法仍有比较明显的优势，图像检索文本和文本检索图像两个任务的平均召回率 mR 对比现有方法中表现最好的 AMFMN 提升了 2.84%。总体而言，本文的方法与现有方法相比，在遥感图文跨模态检索任务上达到了最佳的检索性能。

2.3.2 消融实验

为了评估和分析本文所提出的 MFF 模块和 SFE 模块对检索性能的作用和贡献，本文在两个数据集上分别进行了消融实验，主要包括以下几个部分：

（1）CNN+BERT 表示最基本的跨模态检索架构。在实验时，使用 Inception Resnet V2 作为模型的遥感图像特征提取网络，并将其输出的遥感图像特征不经过 MFF 模块和 SFE 模块的处理，直接通过平均池化、Dropout 和 Softmax 层进行降维，得到最终的图像特征。

（2）在 CNN+BERT 基础检索架构上添加 MFF 模块：使用 MFF 模块对多尺度特征进行融合，并将融合特征与原单一尺度的图像特征分别乘以可变系数后相加，再进行降维得到具有多尺度信息的图像特征。该模型用 MFF+BERT 表示。

（3）在 CNN+BERT 的技术上增加 SFE 模块：将原 Inception-Resnet-C 块输出的图像特征 F_1 作为 SFE 模块的基准特征，并使用低层视觉特征作为强化特征对其进行显著性特征增强，最后将增强后的特征进行降维得到最终的图像特征。该模型用 SFE+BERT 表示。

（4）最后将 MFF 模块与 SFE 模块相结合，在融合

多尺度信息的同时增强显著性特征，即本文提出的基于MFF-SFE的遥感图文跨模态检索模型，用MFF+SFE+BERT表示。

本文在两个数据集上的消融实验结果如表3和表4所示，其中最佳的结果用粗体表示。

表3 UCM-Captions数据集消融实验效果

Table 3 Results of ablation experiment on UCM-Captions

方法	模块		图像检索文本			文本检索图像			mR%
	MFF	SFE	R@1%	R@5%	R@10%	R@1%	R@5%	R@10%	
CNN+BERT			23.33	50.95	63.81	22.66	60.66	77.90	49.89
	√		20.95	53.81	**73.33**	21.33	62.00	81.14	52.09
		√	21.43	50.47	68.57	18.28	60.76	**82.48**	50.33
	√	√	**27.62**	**55.24**	**73.33**	**25.81**	**63.62**	81.43	**54.51**

表4 Sydney-Captions数据集消融实验效果

Table 4 Results of ablation experiment on Sydney-Captions

方法	模块		图像检索文本			文本检索图像			mR%
	MFF	SFE	R@1%	R@5%	R@10%	R@1%	R@5%	R@10%	
CNN+BERT			**24.19**	46.77	61.29	19.03	53.22	72.26	46.13
	√		20.97	50.00	61.29	19.67	**65.81**	83.55	50.22
		√	22.58	**53.23**	64.52	**21.93**	56.13	77.10	49.25
	√	√	**24.19**	51.61	**72.58**	20.00	65.48	**84.19**	**53.01**

由以上两表可见，MFF+BERT模型在UCM-Captions数据集和Sydney-Captions数据集上的平均召回率mR均优于基础的CNN+BERT架构，分别提升了2.2%和4.09%。这证明了MFF模块的有效性，表明MFF模块能够更好地利用遥感图像中的多尺度信息。

其次，SFE+BERT模型在两个数据集上的大部分指标都高于基础的CNN+BERT架构，例如，在Sydney-Captions数据集的文本检索图像任务中，与CNN+BERT模型相比，R@1、R@5和R@10分别提升了2.9%、2.91%、4.84%。这说明SFE模块能够有效提取遥感图像中的目标信息，证明了SFE模块的有效性。同时，SFE+BERT模型的整体效果略逊于MFF+BERT模型。这可能是因为SFE+BERT模型使用F_1作为SFE模块的基准特征，缺少了多尺度信息，导致模块对图像特征中较小目标的特征信息增强不足，降低了检索模型对目标信息的识别能力。

最后，完整的MFF+SFE+BERT模型在两个数据集上，召回率平均值mR有着明显提升。这表明，通过结合运用MFF模块和SFE模块，本文能够获得更全面、更精确的遥感图像特征，从而提升检索模型的精度，同时也证实了MFF和SFE两个模块在本文模型中的有效性和重要性。

2.3.3 可视化展示

通过上述实验，本文得到如图7和图8所示的结果，其中，正确的检索结果使用绿色字体或绿色边框标记，错误结果则用红色字体或边框标记。图7为本文模型进行图像检索文本任务的三个示例，该任务使用在训练集上训练好的模型，对待检索文本与所查询的遥感图像进行相似性度量，并按相似度由高到低选取其中最高的四个文本作为图像检索文本任务的最终检索结果。

由图 7（a）和图 7（b）可见，本文模型检索得到的文本均与原遥感图像的文本描述一致，充分证明了本文模型的有效性。从图 7（c）可见，相似度排名第四位的文本与该遥感图像的内容并不完全相同。经分析，应该是由于这句文本与正确文本（图 7（c）中第二句文本）的文本特征过于相似所导致的，后续将会针对文本语句中的语义信息展开进一步研究。

TOP 1: A river with a road on one bank and plants on the other side .

TOP 2: Green plants flourish on one bank of a river with a road on the other side .

TOP 3: Lots of green plants flourish on one bank of a river with a road on the other side .

TOP 4: It is a river with black water and a road on one bank of the river .

（a）示例一

TOP 1: A house with plants surrounded and a road beside in the sparse residential area .

TOP 2: A house with orange roofs is surrounded by lawn and withered trees in the sparse residential area .

TOP 3: This is a sparse residential area with a villa surrounded by lawn and withered trees .

TOP 4: A house with lawn and withered trees surrounded is in the sparse residential area .

（b）示例二

TOP 1: Lots of boats docked in lines at the harbor and the boats are closed to each other .

TOP 2: Many boats docked in lines at the harbor and the water is blue .

TOP 3: Lots of boats docked at the harbor and the water is blue .

TOP 4: Many boats docked in lines at the harbor and the water is black .

（c）示例三

图 7 图像检索文本结果的部分示例

Fig. 7 Selected examples of text results for image retrieval

如下所示，图 8 为本文模型进行文本检索图像任务的三个示例。该任务通过训练好的模型计算文本与待检索遥感图像间的相似度，并按相似度大小选取四张遥感图像作为文本检索图像任务的最终结果。

由图 8（a）和图 8（b）可见，模型检索出的遥感图像均与查询文本相匹配，充分展示了本文模型的有效性。从图 8（c）可见，相似度排名前二的遥感图像并非查询文本所对应的遥感图像。经分析，可能是由于数据集中某些遥感图像过于相似，从而导致检索结果产生较大误差，后续将对此问题展开进一步研究。

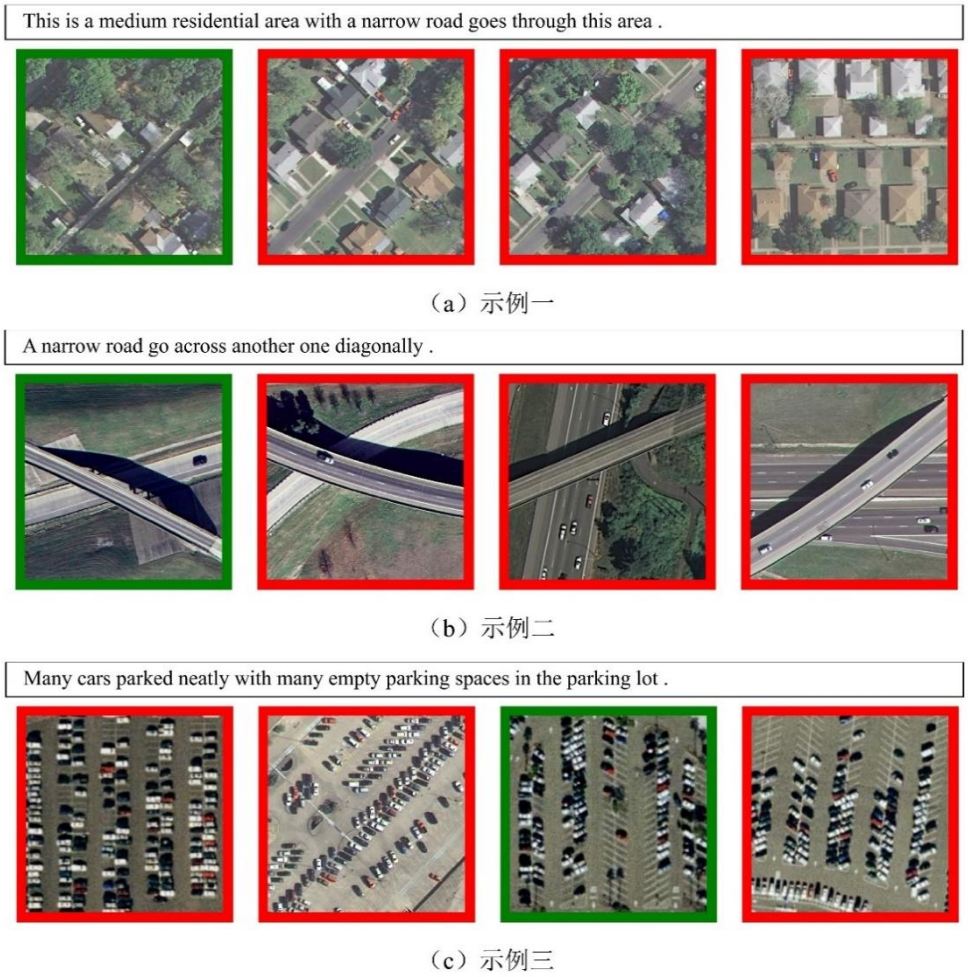

图 8 文本检索图像结果的部分示例

Fig. 8 Selected examples of image results for text retrieval

3 结论

本文提出了一种基于 MFF-SFE 的遥感图文跨模态检索方法，使用 Inception Resnet V2 和 BERT 分别提取遥感图像和文本的特征，并采用 MFF 模块和 SFE 模块来融合并增强遥感图像的图像特征。其中，MFF 模块对提取的遥感图像多尺度特征进行融合，并采用通道注意力机制增强多尺度信息处理能力；而 SFE 模块则使用低层视觉特征对 MFF 模块输出的融合特征进行显著性特征增强，并通过自适应最大池化来过滤冗余特征，从而突出遥感图像的目标信息。

本文在两个公开数据集上进行了对比实验和消融实验，实验结果表明本文方法在遥感图文跨模态检索任务上达到了最佳的检索性能，并证明了本文所提出的 MFF 模块和 SFE 模块的有效性。

不过本文方法尚未考虑图像与文本之间的语义关系，所用的相似性度量比较简单，且模型规模相对较大，后续将考虑在图像特征与文本特征之间加入双向注意力机制、并在相似性度量处引入度量学习，进一步提升模型的准确性，同时采用轻量化进一步优化网络模型。

参考文献

[1] Chi M 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. DOI: 10.1109/JPROC.2016.2598228.

[2] Kaur P, Pannu H S, Malhi A K. Comparative analysis on cross-modal information retrieval: A review[J]. Computer Science Review, 2021, 39: 100336. DOI: 10.1016/j.cosrev.2020.100336.

[3] Chen C, Zou H X, Shao N Y, et al. Deep semantic hashing retrieval of remotec sensing images[C]//IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. Valencia, Spain. IEEE, 2018: 1124-1127. DOI: 10.1109/IGARSS.2018.8519276

[4] Ye F M, Luo W, Dong M, et al. SAR image retrieval based on unsupervised domain adaptation and clustering[J]. IEEE Geoscience and Remote Sensing Letters, 2019, 16(9): 1482-1486. DOI: 10.1109/LGRS.2019.2896948.

[5] Guo M, Zhou C H, Liu J H. Jointly learning of visual and auditory: A new approach for RS image and audio cross-modal retrieval[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, 12(11): 4644-4654. DOI: 10.1109/JSTARS.2019.2949220.

[6] Shi Z W, Zou Z X. Can a machine generate humanlike language descriptions for a remote sensing image?[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(6): 3623-3634. DOI: 10.1109/TGRS.2017.2677464.

[7] Wang G A, Hu Q H, Cheng J, et al. Semi-supervised generative adversarial hashing for image retrieval[C]//European Conference on Computer Vision. Cham: Springer, 2018: 491-507.10.1007/978-3-030-01267-0_29

[8] Lu J S, Batra D, Parikh D, et al. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[EB/OL]. 2019: arXiv: 1908.02265. http://arxiv.org/abs/1908.02265

[9] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[EB/OL]. 2021: arXiv: 2103.00020. (2021-02-26)[2024-04-01]. http://arxiv.org/abs/2103.00020.

[10] 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. DOI: 10.3390/rs12030405.

[11] Lv Y F, Xiong W, Zhang X H, et al. Fusion-based correlation learning model for cross-modal remote sensing image retrieval[J]. IEEE Geoscience and Remote Sensing Letters, 2021, 19: 6503205. DOI: 10.1109/LGRS.2021.3131592.

[12] Mikriukov G, Ravanbakhsh M, Demir B. Deep unsupervised contrastive hashing for large-scale cross-modal text-image retrieval in remote sensing[EB/OL]. 2022: arXiv: 2201.08125. http://arxiv.org/abs/2201.08125

[13] Yuan Z Q, Zhang W K, Tian C Y, 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: 5620616. DOI: 10.1109/TGRS.2022.3163706.

[14] Cheng Q M, Zhou Y Z, Fu P, et al. A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 4284-4297. DOI: 10.1109/JSTARS.2021.3070872.

[15] Zheng F Z, Li W P, 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. DOI: 10.3390/app122312221.

[16] Tang X, Wang Y J, Ma J J, et al. Interacting-enhancing

feature transformer for cross-modal remote-sensing image and text retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5611715. DOI: 10.1109/TGRS.2023.3280546.

[17] Yuan Z 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: 4404119. DOI: 10.1109/TGRS.2021.3078451.

[18] Wang Y J, Ma J J, Li M T, et al. Multi-scale interactive transformer for remote sensing cross-modal image-text retrieval[C]//IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium. Kuala Lumpur, Malaysia. IEEE, 2022: 839-842. DOI: 10.1109/IGARSS46834.2022.9883252.

[19] 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. DOI:10.13700/j.bh.1001-5965.2022.0527.

[20] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. February 4 - 9, 2017, San Francisco, California, USA. ACM, 2017: 4278 – 4284. DOI: 10.5555/3298023.3298188.

[21 Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. 2018: arXiv: 1810.04805. (2018-10-11)[2024-04-01]. http://arxiv.org/abs/1810.04805.pdf.

[22] Qu B, Li X L, Tao D C, et al. Deep semantic understanding of high resolution remote sensing image[C]//2016 International Conference on Computer, Information and Telecommunication Systems (CITS). Kunming, China. IEEE, 2016: 1-5. DOI: 10.1109/CITS.2016.7546397.

[23] Faghri F, Fleet D J, Kiros J R, et al. VSE++: improving visual-semantic embeddings with hard negatives[EB/OL]. 2017: arXiv: 1707.05612. (2017-07-18)[2024-04-01]. http://arxiv.org/abs/1707.05612.

[24] Lee K H, Chen X, Hua G, et al. Stacked cross attention for image-text matching[C]//European Conference on Computer Vision. Cham: Springer, 2018: 212-228.10.1007/978-3-030-01225-0_13

[25] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149. DOI: 10.1109/TPAMI.2016.2577031.

[26] Wang Z H, Liu X H, Li H S, et al. CAMP: cross-modal adaptive message passing for text-image retrieval[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South). IEEE, 2019: 5763-5772. DOI: 10.1109/ICCV.2019.00586.

[27] Wang T, Xu X, Yang Y, et al. Matching images and text with multi-modal tensor fusion and re-ranking[C]//Proceedings of the 27th ACM International Conference on Multimedia. October 21 - 25, 2019, Nice, France. ACM, 2019: 12 – 20. DOI: 10.1145/3343031.3350875.