中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-15

论文引用格式:Zhou Kaili, Wang Peng, Chen Jian. Remote sensing image caption generation method based on multi-scale and multi-semantic fusion and collaboration[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.250591. (周凯立, 王鹏, 程剑. 多尺度多语义融合协同的遥感图像字幕生成方法[J/OL]. 中国图象图形学报,XXXX:1-15. DOI: 10.11834/jig.250591.)[DOI:10.11834/jig.250591]

多尺度多语义融合协同的遥感图像字幕生成方法

周凯立[1,2,3],王鹏[1,4],程剑[1]

1. 南京航空航天大学电子信息工程学院, 南京　210016; 2. 复杂系统先进控制与智能自动化湖北省重点实验室, 武汉　430074; 3. 地球探测智能化技术教育部工程研究中心, 武汉　430074; 4. 南京航空航天大学深圳研究院,深圳　518110.

摘　要:**目的**　遥感图像字幕生成(remote sensing image captioning, RSIC)作为融合计算机视觉与自然语言处理的跨模态任务,能将遥感图像中的关键地物、场景及关联关系转化为通俗文本,在灾害应急、农业监测等领域具有重要应用价值。然而,现有RSIC方法存在特征提取不全面如忽视多尺度纹理特征或单一语义特征依赖、跨模态信息对齐不足的问题。**方法**　本文提出一种集成多模块的RSIC新方法。在特征提取阶段,设计多尺度特征融合模块(multi-scale feature fusion module, MSFFM)与混合语义融合模块(hybrid semantic integration module, HSIM):MSFFM通过结合空间-通道注意力及频域增强,强化纹理与结构信息;HSIM利用局部特征、全局语义特征,经特征维度对齐与自注意力融合,丰富图像语义表达。在文本生成阶段,构建跨模态注意力模块,融合LSTM隐状态与单词嵌入向量生成引导信号,实现视觉-文本特征深度对齐;同时设计跨模态特征对齐损失,以温度缩放交叉熵最小化图像与文本特征语义差异。**结果**　实验在UCM-Captions、RSICD及自建SAR-Captions数据集上验证,所提方法在BLEU1-BLEU4、METEOR、ROUGE、CIDEr指标上均优于对比模型。其中,在SAR-Captions数据集上,BLEU4值达到0.8570、CIDEr值达到4.2606,较基线模型提升显著;消融实验证明MSFFM、HSIM及跨模态对齐模块对性能提升的关键作用。**结论**　本文所提出的RSIC模型,能有效整合多源特征、优化跨模态对齐,为遥感图像(尤其是SAR图像)的语义解读提供高效解决方案。

关键词:遥感图像字幕生成;多尺度;混合语义;注意力机制;特征提取;跨模态对齐

Remote sensing image caption generation method based on multi-scale and multi-semantic fusion and collaboration

Zhou Kaili[1,2,3], Wang Peng[1,4], Chen Jian[1]

1. College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China; 2. Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074 Hubei, China; 3. Engineering Research Center of Intelligent Technology for Geo-Exploration, Ministry of Education, Wuhan 430074 Hubei, China; 4. Shenzhen Research Institute, Nanjing University of Aeronautics and Astronautics, Shenzhen 518110.

Abstract: **Objective** Remote sensing image captioning (RSIC) emerges as a pivotal cross-modal task that bridges computer vision and natural language processing, aiming to translate complex visual information in remote sensing images, including key ground objects, spatial scenes, inter-object relationships, and environmental contexts into human-readable

收稿日期:2025-11-24;**修回日期**:2026-03-17

基金项目:国家自然科学基金项目(92464204);高等学校学科创新引智计划项目资助(B17040);广东省基础与应用基础研究基金项目(2025A1515010258);深圳市科技计划项目(JCYJ20240813180005007).

Supported by: National Natural Science Foundation of China (92464204); 111 project (B17040); Guangdong Basic and Applied Basic Research Fund (2025A1515010258); Shenzhen Science and Technology Program (JCYJ20240813180005007).

textual descriptions. This technology plays an irreplaceable role in a wide range of practical applications: in disaster emergency response, it enables rapid interpretation of affected areas (e. g. , identifying collapsed buildings or flooded regions from post-disaster remote sensing images) to support real-time decision-making; in agricultural monitoring, it facilitates automated analysis of crop distribution, growth status, and land use changes; in environmental change analysis, it helps track deforestation, glacial melting, and coastal erosion over long periods. Unlike single-modal tasks such as image classification or object detection, RSIC delivers more comprehensive and coherent semantic expression, aligning with human natural understanding patterns and significantly lowering the threshold for non-professional users to utilize remote sensing data. However, existing RSIC methods face two critical challenges that hinder their performance and applicability. First, feature extraction is often incomplete: many approaches either fail to fully capture multi-scale texture features of remote sensing images (resulting in loss of fine-grained details like surface textures of ground objects) or over-rely on a single type of semantic feature (limiting the ability to represent complex scene information). Second, cross-modal information alignment is insufficient: most models focus more on processing image features but lack effective mechanisms to bridge the semantic gap between visual features and textual information, leading to inconsistencies between generated captions and image content. **Method** To address the aforementioned challenges, this paper proposes a novel RSIC method integrated with multiple functional modules. In the feature extraction stage, two core modules are designed: the Multi-Scale Feature Fusion Module (MSFFM) and the Hybrid Semantic Integration Module (HSIM). MSFFM adopts ResNet50 as the backbone, extracting shallow details features, middle main features, and deep global features from its last three convolutional layers. Shallow details features undergoes depthwise separable convolution for fine details and reduced computation plus spatial attention focusing on key textures to optimize main features. Deep global features is enhanced via dilated convolution expanded receptive field for global structure and channel attention to strengthen critical global feature transmission. After weighted fusion, frequency domain enhancement uses FFT and learnable parameters to adaptively select useful bands, boosting texture and structural representation. The HSIM module enriches image semantic expression by integrating multi-source features. Specifically, Faster R-CNN is used to extract object-level local features, while the CLIP pre-trained model advantaged by large-scale image-text pair pre-training is employed to extract global semantic features, which helps alleviate the problem of scarce remote sensing data. Before fusion, the global features extracted by CLIP are passed through a fully connected layer to align their dimension with local features and texture features (output from MSFFM). The three types of features (texture features, local object features, and global semantic features) are concatenated, and self-attention is utilized to optimize the final visual feature representation—this self-attention mechanism abandons learnable parameters, instead calculating similarity in spatial and channel dimensions through matrix transposition and multiplication, and generating attention weights via softmax to highlight key semantic information. In the text generation stage, a Cross-Modal Alignment Module (CMAM) is constructed to achieve deep alignment between visual and textual features. First, the hidden state of the LSTM (containing rich temporal context information) and word embedding vectors are concatenated to form a comprehensive textual semantic representation, which serves as a guiding signal to enhance the pertinence of image feature selection. This textual representation is then used to compute feature similarity with visual features through dot product operation, and the results are fed into a fully connected layer to generate channel-wise attention weights for image features. Using the textual semantic representation as queries, and channel-enhanced visual features as keys and values, multi-head attention is applied to achieve precise focus on visual features, ensuring deep alignment between visual content and textual semantics. To further minimize the semantic gap between image and text features, a cross-modal feature alignment loss is designed based on temperature-scaled cross-entropy. This loss function calculates the cosine similarity matrix of all image-text pairs in a batch using image features extracted by the encoder and global text features extracted by an additional Transformer encoder. Cross-entropy loss is computed along both image and text dimensions to optimize bidirectional alignment, and a temperature parameter τ is introduced to adjust the probability distribution, controlling the "smoothness" of predictions and improving the stability of model training. **Result** Experiments are conducted on UCM-Captions, RSICD, and SAR-Captions datasets, comparing with mainstream models. Evaluation metrics include BLEU1-BLEU4, METEOR, ROUGE, and CIDEr. The proposed method outperforms all comparative models across datasets. On UCM-Captions, BLEU4 reaches 0. 7130 and CIDEr 3. 3543, with a 17. 9% BLEU1-BLEU4 attenuation rate (lower than mlat's 18. 7%),

indicating superior long-sentence coherence. On RSICD, BLEU4 is 0.3277 and CIDEr 0.9223, outperforming HC and aoa. On SAR-Captions, BLEU4 hits 0.8570 and CIDEr 4.2606, verifying adaptability to SAR images and dataset effectiveness. Ablation experiments show that the cross-modal alignment module boosts BLEU4 and CIDEr significantly; MSFFM+HSIM improves all metrics; the full model (integrating all modules) achieves optimal performance, confirming multi-module synergy. **Conclusion** This paper proposes a novel RSIC method integrated with multi-scale feature fusion, hybrid semantic integration, and cross-modal alignment modules, along with a dedicated SAR image captioning dataset. The MSFFM and HSIM modules effectively address the problem of incomplete feature extraction by integrating multi-scale texture features, local object features, and global semantic features. The CMAM module optimize the alignment of cross-modal information, reducing the semantic gap between images and texts.

Key words: Remote sensing image captioning; multi-scale; hybrid semantics; attention mechanism; feature extraction; cross-modal alignment

0 引　言

伴随着卫星及成像技术的高速发展，遥感图像的数量激增。如何从这些图像中精准提取出具有实际意义的语义信息，是当下的研究热点。遥感图像字幕超脱于图像字幕任务（杜海骏等，2020；谭云兰等，2021；周峻宇等，2025），是一项融合计算机视觉与自然语言处理的跨模态任务，旨在以通俗的文字信息解读丰富的遥感图像内容，精准捕捉图像中的关键地物、场景及它们之间的关系（Stefanini等，Ren等，2022；刘健等，2025）。与其他单模态图像任务如图像分类（Lin等，石争浩等，2023；Chen等，2024；Gao等，2025）、目标检测（Zhang等，2023；Yao等，2024；余凌霄等，2025）、语义分割（Xiao等，2023；侯志强等，2025）不同，遥感图像字幕任务信息表达更全面、更连贯，能捕捉更复杂的语义关联，更贴近人类自然理解习惯。因此探索有效的遥感图像字幕生成方法，在灾害应急响应、农业监测、环境变化分析等领域具有重要应用价值，有效降低了遥感数据的使用门槛（Cheng等，2022）。

合成孔径雷达（synthetic aperture radar，SAR）图像是遥感图像的重要分支，其通过微波成像，具有全天候、全天时、可穿透地表覆盖物的独特优势，在军事侦察、海洋监测等领域不可替代（王蓉芳等，2023；Meena等，Li等，2025）。SAR图像字幕因SAR数据的独特价值与解读难度，是一个极具潜力的SAR图像理解方案。SAR和光学遥感图像均具有尺度跨度大的特点：单幅遥感图像可能同时包含不同尺度的地物，如舰船与岛屿，字幕需平衡细节与宏观分布的描述。其次，遥感图像字幕数据标注的具有稀缺性，现有公开数据集（如RSICD、UCM-Captions）规模远小于自然图像字幕数据集，SAR图像字幕目前少有公开的专用数据集。

基于深度学习的遥感图像字幕生成方法通常使用编码器-解码器架构，可分为特征提取阶段和文本生成阶段（Dhinesh等，2023）。早期的研究（Qu等，2016）（Zhang等，2017）使用卷积神经网络提取图像特征，并使用循环神经网络生成字幕。为了更好地关注遥感图像特征中必要的部分，Lu等人（2017）首先将注意力机制引入RSIC，显著提高了模型性能。此后，注意力机制在RSIC大放异彩。为了更好地关注图像中的各种目标，自下而上的注意（Anderson等，2018）证明对象级特征在改进跨模态任务中效果良好，在遥感图像领域也有一定的优势。为了利用语义内容的结构化空间关系，Zhao等人（2021）提出了结构化注意，并利用像素级区域图像分割信息。Zhang等人（2021）提出了RSIC的全局视觉特征引导注意（global visual feature-guided attention，GVFGA）机制和语言状态引导注意（linguistic state guided attention，LSGA）机制。前者用于从融合后的图像特征中滤除冗余的特征分量，使视觉特征更加突出。后者增强了视觉和文字特征的融合，消除了不相关的信息。同时，遥感图像的多尺度信息引起广泛关注，注意力机制常与多尺度特征融合方法结合，Li等人（2021）提出了一种循环注意（recurrent attention，RA）机制，从编码和非视觉特征中提取高级注意图，帮助解码器识别有效信息。Zhang等人（2019）提出了一种基于属性关注机制的遥感图像描述生成模型，均利用了多层卷积神经网络的输出特征。Zhang等人（2023）提出多源交互阶梯注意力机制，将注意力权重分为核心、周边等三级，以更贴近人类描述图

像的思维过程。Wang等人(2024)将目标掩码和多目标分类与字幕生成任务相结合,结合跨任务提取的图像特征实现了语义与空间特征的高效融合。目前的RSIC研究大多集中于特征提取阶段,虽然这些方法有效地提高了RSIC模型性能,但仍存在特征提取不全面的问题,仅利用了单种语义特征或不能有效挖掘图像的多尺度纹理特征。在文本生成阶段,目前大多研究使用基础的递归神经网络或Transformer(Vaswani等,2017),MLAT(Liu等,2022)利用长短期记忆网络(long short-term memory,LSTM)聚合Transformer编码层特征优化多尺度信息利用,Zhao等人(2024)提出协同连接Transformer使用交叉注意处理对象级特征和网格特征,促使解码器关注图像局部和全局特征。Meng等人(2024)引入对比语言-图像预训练模型(contrastive language-image pre-training,CLIP)(Radford等,2021)提取图像嵌入作为全局特征,结合全局分组注意力增强局部建模,同样使用交叉注意的Transformer用于生成字幕。但这些方法仍注重图像特征的处理,对跨模态信息的对齐有限,未能很好利用上下文信息。并且,遥感图像字幕数据集规模小仍是限制遥感图像字幕生成模型发展的一个难点。

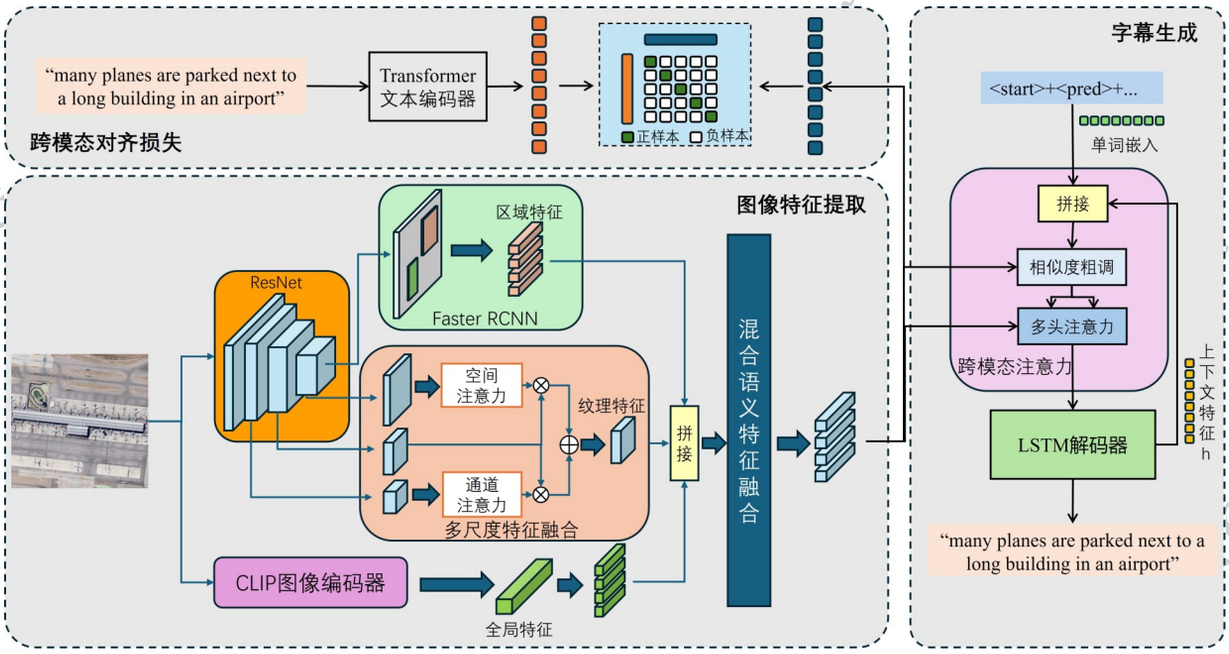

图1　总体网络架构

Fig. 1　Overall network architecture

为了应对上述问题,本文提出一种新的RSIC方法,该方法利用多尺度特征融合模块(multi-scale feature fusion module, MSFFM)和混合语义融合模块(hybrid semantic integration module, HSIM)提取遥感图像特征,旨在获取全面有效的视觉特征。首先,我们利用resnet50作为主干网络,在不同的分辨率层上提取多尺度特征,多尺度特征融合模块基于自注意力和频域增强的方法挖掘遥感图像的纹理特征。同时,我们利用Faster RCNN(Ren等,2016)和CLIP分别提取图像的对象级特征和全局特征,CLIP是通过大规模的图像-文本对数据预训练的跨模态模型,有助于缓解遥感图像数据稀缺的问题。混合语义融合模块将多种特征融合,以丰富纹理和语义信息。跨模态对齐模块(cross-modal alignment module, CMAM)利用跨模态注意力模块用于实现视觉特征与文本信息的交互,同时设计跨模态损失函数约束全局视觉特征。MSFFM、HSIM与CMAM三个模块呈"特征提取-语义丰富-模态对齐"的协同关系,MSFFM模块提取多尺度视觉纹理与结构特征,提供精细化视觉特征基础;HSIM整合局部与全局语义特征,丰富视觉特征的语义信息;MSFFM与HSIM的组合实现全面的视觉-语义特征提取;最后CMAM模块搭建跨模态桥梁。三者形成完整技术链路,互补解决了图像到字幕的匹配。最后,我们还构建了一个人工标注的SAR图像字幕数据集。

文本的贡献如下:1)提出了多尺度特征融合模

块(MSFFM)和混合语义融合模块(HSIM)。MSFFM可以融合丰富的多尺度特征,突出浅层特征的纹理信息和深层特征的语义信息。HSIM模块利用对象级特征、纹理特征和全局特征充分丰富图像特征,二者结合实现视觉-语义特征的全面提取。2)提出了跨模态对齐模块(CMAM)。使用注意力机制旨在将全面的图像特征与不同模态特征之间深度对齐,从而使模型能够生成准确的隐藏状态,实现视觉特征精准筛选,提升跨模态匹配效率。此外,我们引入了一种跨模态特征对齐损失来最小化图像和文本特征之间的差异。3)我们在两个公开的自然遥感数据集上验证了所提出方法的优越性能。并构建了一个人工标注的SAR-Captions图像字幕数据集,使用多种方法验证了其有效性。

1 本文方法

1.1 总体框架

所提出的方法的总体框架如图1所示。它由多尺度特征融合模块、混合语义融合模块、跨模态注意力模块和双层LSTM组成,并以跨模态特征对齐损失约束图像与字幕信息的差异。首先,本文通过多尺度特征融合策略,利用ResNet50主干网络的最后3层卷积层输出特征,以空间、通道注意力机制整合纹理等多维度特征,并结合频域增强方法进一步突出纹理信息;再通过多语义融合策略,利用卷积网络、目标检测网络和CLIP预训练大模型,整合纹理信息、局部细节及全局场景特征,突破图像单一特征表达能力有限的瓶颈;同时,为改善文本生成准确性欠佳的问题,利用跨模态注意力机制结合LSTM解码器,整合图像多源特征与文本特征,打破模态壁垒,引导生成的文本紧密贴合图像内容。此外,借助跨模态对齐损失函数优化图像与文本特征在语义空间的匹配度,有效缓解模态间语义鸿沟,提升图文语义一致性。

1.2 图像特征提取模块

多尺度特征融合模块是视觉特征处理的重点。该模块利用空间注意力与通道注意力机制深度解析主干网络ResNet50输出的不同层级特征。首先,我们提取ResNet50的最后三层卷积层的输出特征$\boldsymbol{X}1 \in \mathbb{R}^{C/2\times 2H\times 2W}$、$\boldsymbol{X}2 \in \mathbb{R}^{C\times H\times W}$、$\boldsymbol{X}3 \in \mathbb{R}^{2C\times H/2\times W/2}$,以$\boldsymbol{X}2$特征为主特征,经过简单卷积进一步提取关键信息。

$$\boldsymbol{X}2' = \mathrm{Conv}(\boldsymbol{X}2) \tag{1}$$

其中,Conv()表示普通3×3卷积,并利用分组卷积的方式适当减少参数量,$\boldsymbol{X}2'$保持与$\boldsymbol{X}2$相同的尺度,并挖掘上下文信息。

其次利用浅层特征$\boldsymbol{X}1$优化主特征$\boldsymbol{X}2$的细节特征,我们使用深度可分离卷积提取$\boldsymbol{X}1$各个特征层的细节信息,有效减少参数量和计算量,空间注意力机制通过对特征图进行全局池化操作深化特征层间关系,并生成空间注意力权重图,聚焦图像中的纹理细节;

$$\boldsymbol{X}1' = \mathrm{DSConv}(\boldsymbol{X}1) \tag{2}$$

$$\boldsymbol{M}_S = \sigma(\mu(\mathrm{GAP}(\mathrm{Conv}_{1\times 1}(\boldsymbol{X}1'))) \otimes \mathrm{Conv}_{1\times 1}(\boldsymbol{X}1')) \tag{3}$$

其中,DSConv()为深度可分离卷积,利用DSConv()将$\boldsymbol{X}1' \in \mathbb{R}^{C\times H\times W}$与$\boldsymbol{X}2$对齐尺度。$\mathrm{Conv}_{1\times 1}()$为逐通道卷积,用于深化特征通道间的交互。$\sigma$为Sigmoid函数,$\mu$为softmax函数,GAP()表示空间维度的平均池化。$\boldsymbol{M}_S \in \mathbb{R}^{1\times H\times W}$最终的空间注意力图。

同时利用深层特征$\boldsymbol{X}3$优化主特征$\boldsymbol{X}2$的整体特征,我们使用膨胀卷积提取$\boldsymbol{X}3$的全局特征,通道注意力机制则根据各通道特征分配通道注意力权重,学习通道间依赖关系,强化整体特征传递。

$$\boldsymbol{X}3' = \mathrm{DilConv}(\boldsymbol{X}3) \tag{4}$$

$$\boldsymbol{M}_C = \sigma(\mu(\mathrm{Conv}_{1\times 1}(\boldsymbol{X}3')) \otimes \mathrm{Conv}_{1\times 1}(\boldsymbol{X}3')) \tag{5}$$

$$\boldsymbol{F}_{ms} = \boldsymbol{X}2' + \boldsymbol{M}_S \odot \boldsymbol{X}2' + \boldsymbol{M}_C \odot \boldsymbol{X}2' \tag{6}$$

其中,DilConv()为膨胀卷积,膨胀率分别设置为1、2、5,利用DSConv()将$\boldsymbol{X}3' \in \mathbb{R}^{C\times H\times W}$与$\boldsymbol{X}2$对齐尺度。$\boldsymbol{M}_C \in \mathbb{R}^{C\times 1\times 1}$为最终的通道注意力图。$\boldsymbol{F}_{ms} \in \mathbb{R}^{C\times H\times W}$为残差连接后的多尺度融合输出。

对多尺度特征完成加权融合后,能整合从细粒

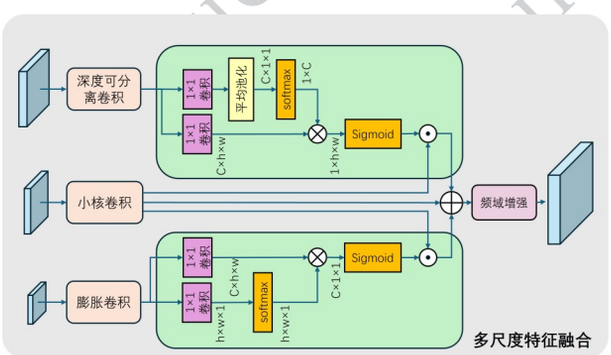

图2 多尺度特征融合模块

Fig. 2 Multi-scale feature fusion module

度纹理到粗粒度结构的多维度视觉信息，让特征层次更丰富。此外，我们对处理后的特征进行频域增强，使模型自主学习resnet50图像特征的频谱信息，通过可学习参数控制频段选择，通过通道注意力在频域进一步优化语义信息选择，从而增强纹理细节信息和结构信息。

$$\boldsymbol{F}_{fred}=\mathrm{FFT}(\boldsymbol{F}_{ms}) \tag{7}$$

$$\boldsymbol{F}_{fred}^{C}=\sigma(\mathrm{GAP}(\mathrm{Conv}(\boldsymbol{F}_{fred})))\odot\boldsymbol{F}_{fred} \tag{8}$$

$$\boldsymbol{F}_{fred}^{S}=\boldsymbol{F}_{fred}\odot\alpha \tag{9}$$

$$\boldsymbol{Q}2=\mathrm{IFFT}(\boldsymbol{F}_{fred}^{C}+\boldsymbol{F}_{fred}^{S}) \tag{10}$$

其中，FFT()和IFFT()分别表示傅里叶变换和傅里叶逆变换，α表示可学习参数矩阵，$\boldsymbol{F}_{fred}^{C}$和$\boldsymbol{F}_{fred}^{S}$分别表示通道和空间维度增强后的特征矩阵，$\boldsymbol{Q}2$为最终输出矩阵。

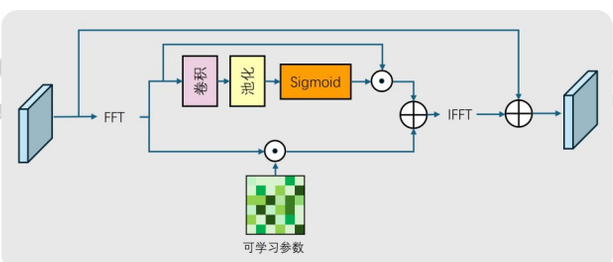

图3　频域增强模块

Fig. 3　Frequency domain enhancement module

多语义特征融合模块输入图像首先进入ResNet基础骨干网络，利用多尺度融合模块对图像进行分层特征提取，得到纹理特征$\boldsymbol{Q}1\in\mathbb{R}^{C1\times L1}$，为后续深度特征挖掘筑牢根基。在基础纹理特征之上，Faster RCNN目标检测网络挖掘物体类别、位置等区域特征，聚焦局部关键内容，得到局部特征$\boldsymbol{Q}2\in\mathbb{R}^{C2\times L2}$。CLIP图像编码器从全局视角介入，凭借大规模图文数据上预训练的优势，提取图像整体语义特征$\boldsymbol{Q}3\in\mathbb{R}^{L3}$，与局部特征形成互补，把握场景的全局语义，全方位编织图像的视觉信息网络。

我们首先将CLIP提取的全局特征经过全连接层，使其特征维度与局部特征和纹理特征对齐，再进行复制，增强全局特征的影响力，得到$\boldsymbol{Q}3'\in\mathbb{R}^{C3\times L3}$，接着将纹理特征、局部特征和全局特征拼接，最后经过空间-通道注意力进行融合，此处空间-通道注意力抛弃了可学习参数，利用自注意力优化最终特征。首先，将拼接后的特征进行转置变形，与原特征矩阵相乘，从而分别在空间和通道两个维度进行相似度计算，根据特征相似度对不同维度进行softmax处理生成注意力权重。

$$\boldsymbol{Q}=\mathrm{Concat}(\boldsymbol{Q}1,\boldsymbol{Q}2,\boldsymbol{Q}3') \tag{11}$$

$$\boldsymbol{Q}_{C}=\mu(\boldsymbol{Q}^{T}\otimes\boldsymbol{Q})\otimes\boldsymbol{Q} \tag{12}$$

$$\boldsymbol{Q}_{S}=\mu(\boldsymbol{Q}\otimes\boldsymbol{Q}^{T})\otimes\boldsymbol{Q} \tag{13}$$

$$\boldsymbol{Q}_{out}=\boldsymbol{Q}+\boldsymbol{Q}_{C}+\boldsymbol{Q}_{S} \tag{14}$$

其中，$\boldsymbol{Q}_{C}$和$\boldsymbol{Q}_{S}$分别表示经过自注意力得到的通道和空间优化的混合语义特征矩阵。

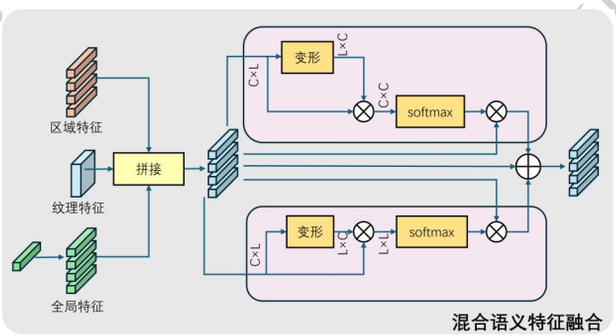

图4　混合语义融合模块

Fig. 4　Hybrid semantic integration module

1.3　跨模态对齐模块

跨模态对齐模块由跨模态注意力和跨模态对齐损失函数组成。在字幕生成阶段，我们构建文本-图像跨模态注意力模块，将文本语义信息高效映射至图像特征空间。该机制依托注意力权重矩阵精准定位图像中与文本描述强关联的视觉区域，基于注意力分布对图像特征进行精细化筛选与重构，以此强化关键视觉特征的表征强度，同时抑制冗余无关特征的干扰。为提升图像特征选择的针对性，我们通过融合上下文特征$\boldsymbol{h}2$与单词嵌入向量$\boldsymbol{w}$构建引导信号。首先，利用LSTM网络输出的隐状态（蕴含丰富时序上下文信息）与单词嵌入向量进行维度拼接，构建更完备的文本语义表征。将上述文本语义表征与图像特征进行点积运算，量化两者间的特征相似度；再将运算结果输入全连接层，生成针对图像特征空间维度的关注度权重。

$$\boldsymbol{X}=\mu(\mathrm{Linear}(\boldsymbol{Q}_{out}\odot\mathrm{Concat}(\boldsymbol{h}_{2},\boldsymbol{w})))\odot\boldsymbol{Q}_{out} \tag{15}$$

其中，Concat()为按空间维度进行拼接。Linear()表示全连接层。

以文本语义表征为查询，以通道增强后的图像特征为键与值，通过多头注意力机制完成对图像特征的精准注意力聚焦，确保图像特征与文本语义的

深度对齐。

$$\mathrm{A}(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v})=\mathrm{softmax}(\frac{\boldsymbol{q}\boldsymbol{k}^{T}}{\sqrt{d_k}})\cdot\boldsymbol{v} \tag{16}$$

$$\mathrm{MH}(\boldsymbol{X})=\mathrm{C}(\mathrm{A}(\boldsymbol{q}_1,\boldsymbol{k}_1,\boldsymbol{v}_1),...,\mathrm{A}(\boldsymbol{q}_h,\boldsymbol{k}_h,\boldsymbol{v}_h))\cdot\boldsymbol{W}^{O} \tag{17}$$

其中，$\mathrm{A}(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v})$表示注意力函数，MH()表示多头注意力。C()表示拼接操作，$\boldsymbol{W}^{O}$为输出变换矩阵。

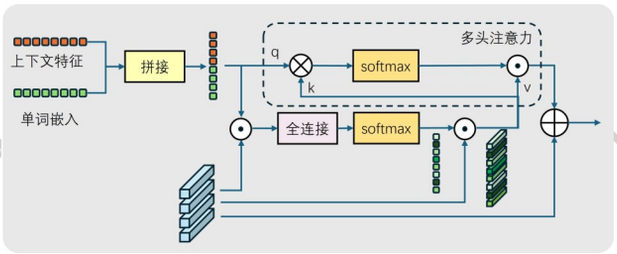

图5　跨模态注意力模块

Fig. 5　Cross-modal attention module

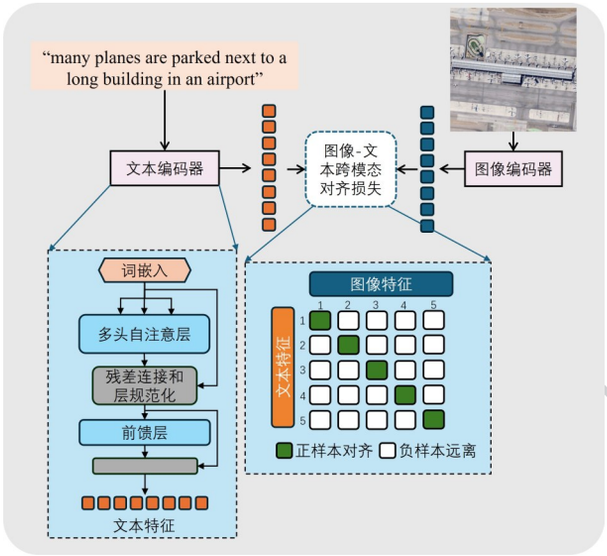

图6　跨模态对齐损失

Fig. 6　Cross-modal alignment loss

跨模态对齐损失目的是让图像的特征和对应的文本描述的特征在特征空间中尽可能接近。这样可以使模型更好地理解图像内容与文本之间的对应关系，从而能够根据图像准确地生成相应的字幕。温度缩放的交叉熵损失利用编码器提取的图像特征和额外的Transformer编码器提取的全局文本特征，计算一个批次中所有图像-文本对的余弦相似度矩阵，然后对相似度矩阵分别沿图像和文本维度计算交叉熵损失，以优化双向对齐，同时，引入了温度参数τ来调整概率分布，控制预测的"平滑度"。

$$\mathcal{L}_{it}=-\frac{1}{B}\log(\frac{\exp(\frac{z_i\cdot\boldsymbol{q}_i^{T}}{\tau})}{\sum_{j=1}^{n}\exp(\frac{z_j\cdot\boldsymbol{q}_j^{T}}{\tau})}) \tag{18}$$

$$\mathcal{L}_{ti}=-\frac{1}{B}\log(\frac{\exp(\frac{\boldsymbol{q}_i\cdot z_i^{T}}{\tau})}{\sum_{j=1}^{n}\exp(\frac{\boldsymbol{q}_j\cdot z_j^{T}}{\tau})}) \tag{19}$$

$$\mathcal{L}_{c}=(\mathcal{L}_{it}+\mathcal{L}_{ti})/2 \tag{20}$$

其中，$\mathcal{L}_{it}$和$\mathcal{L}_{ti}$分别表示图像-文本对齐损失和文本-图像对齐损失，B表示批量大小，z_i表示图像特征，$\boldsymbol{q}_i$表示文本特征，$\mathcal{L}_{c}$表示跨模态对齐损失。

2　实　验

2.1　数据集

UCM-Captions数据集(Qu等，2016)是基于UCM-Merced大学土地利用数据集构建的。图像来自美国地质调查局的国家地图城市区域。UCM-Captions数据集包含21个类别，包括飞机、海滩、高架桥和体育场等，总共有2100张遥感图像。每张遥感图像的分辨率为256×256像素，并配备有5个不同的标题标签。整个数据集使用368个不同的词汇生成了10,500个描述图像的标题标签。

RSICD数据集(Lu等，2017)是一个专门为遥感图像captioning(图像描述)任务设计的大规模数据集。在该数据集中，从谷歌地球、百度地图、MapABC和天地图收集了超10,000张遥感图像。数据集被划分为30个场景类别，像机场、裸地、棒球场、海滩等等。图像在保持原始分辨率的同时，被调整为224×224像素大小。遥感图像总数为10,921张，每张图像配有5条描述性语句。

SAR-Captions数据集是我们人工制作的SAR图像字幕数据集。整体制作围绕"数据筛选-精准标注-结构化整合"展开数据层面，图像筛选自公开SAR船舶目标检测数据集(Wang等，2019；Xu等，2022)，包含多种成像模式、多种极化方式、多种分辨率。重点覆盖海洋、海岸、海岛等典型船舶活动场景，确保场景代表性；标注层面，采用"一图五句"的人工标注模式，按以下标注要求进行标注。首先，舰船数量标注依据原始目标检测标签确定舰船数量，当数量大于5艘时，统一用"many"描述，避免计数冗余；其次，仅描述舰船的相对大小(如"部分舰船尺寸相近")与相对朝向(如"多数舰船朝向一致")因图像

存在多分辨率差异，标注中不涉及任何绝对大小信息；人工补充海岛、海岸等场景背景信息，将场景要素与舰船目标信息深度结合，让标注更具场景关联性（如“海岸附近停泊多艘舰船”）。最终，该数据集形成结构化资源包，包含4000张SAR船舶图像及对应的20000条标注语句，可直接用于SAR图像字幕生成任务。

2.2 评价指标

在图像字幕生成任务中，模型性能的核心评估维度包括生成文本与图像视觉内容的语义对齐度，以及生成语句的语法规范性与语言流畅度。当前学术界广泛采用的评价指标主要有四类：BLEU（Papineni 等，2002）、METEOR（Banerjee 等，2005）、ROUGE（Lin等，2004）与CIDEr（Vedantam等，2015），各指标基于不同设计逻辑实现差异化评估，共同构成模型性能的综合衡量体系。

BLEU最初面向机器翻译任务设计，后经适配应用于图像字幕生成评估。其核心原理是通过计算生成字幕与参考字幕间的N-gram重叠度来量化精确率，同时引入短句惩罚机制以抑制模型生成投机性短句（通过缩短输出长度提升表面重合度），并采用平滑策略缓解低N-gram重叠场景下的分数失真问题。该指标的优势在于计算效率高、结果可解释性强，能够快速反映文本表层信息的匹配程度，但本质局限于字面级匹配，无法有效捕捉语义层面的关联性与一致性，难以区分字面差异但语义等价的生成结果。

METEOR为弥补BLEU在语义捕捉能力上的不足，METEOR通过词干提取、同义词映射等语义匹配策略突破字面限制，实现语义层面的跨表述匹配。在评估逻辑上，其融合精确率与召回率构建F1评分函数，平衡“生成内容的准确性”与“参考信息的覆盖完整性”，显著提升了语义匹配的灵活性。然而，该指标的性能高度依赖外部词典的领域适配性，若词典中缺失特定领域（如遥感、医疗等）的专业术语，将直接导致语义匹配精度下降。

ROUGE以文本连贯性与信息完整性评估为核心，核心聚焦召回率维度的性能衡量。ROUGE-L通过最长公共子序列（LCS）分析，捕捉文本中词汇的顺序关联特征。该指标尤其适用于图像字幕生成等多参考文本场景，能够有效降低单一参考文本带来的评估偶然性，提升结果稳定性。

CIDEr是唯一专为图像字幕生成任务设计的评价指标，其核心创新在于引入TF-IDF加权机制对N-gram进行差异化赋权。通过TF-IDF算法，模型对图像独特语义表征的N-gram赋予更高权重，对通用高频词汇赋予较低权重，从而精准识别生成字幕是否捕捉到图像的核心独特信息。该设计既保证了评估的精准性，又具备良好的场景适配性，能够有效区分不同图像的语义差异。

在实际模型评估中，四类指标形成功能互补：BLEU侧重文本表层信息的准确性验证，METEOR与CIDEr聚焦语义层面的贴合度衡量，ROUGE则关注信息传递的完整性与连贯性。通过多指标联合评估，可有效规避单一指标的固有偏差，全面、客观地反映模型在语义对齐、语言流畅性与信息完整性等维度的综合性能。

2.3 训练设置

词嵌入的维度设置为512。训练阶段，用于构建词汇表的计数单词数量设置为1，生成句子的最大长度设置为50。学习率设置为2×10^{-4}。网络的批量大小固定为16。此外，我们定义耐心度（patience）来监控模型在验证集上的性能，并决定是否调整学习率或者提前终止训练。在推理过程中，我们选择束搜索值为3。我们使用Adam优化器优化模型。所有实验均在NVIDIA RTX 4090D上使用PyTorch 2.0实现。

2.4 对比试验

为了验证本文拟议模型的有效性，我们选择了几种现有的方法，分别为Soft-Attention、Hard-Attention（Lu 等，2017）、FC-Att、SM-Att（Zhang 等，2019）、AoANet（Huang等，2019）、MLA（Li等，2020）、MLAT（Liu 等，2022）、HC-Net（Yang 等，2024）、MG-Transformer（Meng等，2024），这些方法覆盖了遥感图像字幕生成领域不同发展阶段、不同技术路线的典型代表，Soft-Attention、Hard-Attention作为基础注意力方法代表，FC-Att、SM-Att属于基于属性、空间特征优化的注意力变体，可验证所提算法在基础注意力机制上的改进效果。AoANet、MLA以及MLAT模型代表了RSIC在特征融合与文本生成阶段的主流优化方向，对比可验证所提算法在多尺度特征整合、跨模态交互上的优势。HC-Net、MG-Transformer等RSIC领域最新研究模型，二者均聚焦于特征提取优化与跨模态对齐，与所提算法的核心优化方向契

合，可直接验证所提算法在模块设计上的创新性与性能优越性能。在三个数据集上与提出的方法进行对比，同时验证所提出的SAR-Captions图像字幕数据集的有效性。

如表1所示，本文拟议的模型在UCM-Captions遥感图像字幕生成任务中表现最优，其在BLEU1-BLEU4、METEOR、ROUGE、CIDEr所有评价指标上均取得最高值，显著优于其他对比模型，充分验证了其在遥感图像语义理解与字幕生成上的有效性。BLEU1-BLEU4分别衡量生成文本与参考文本的1-gram至4-gram重叠度，对比MLAT模型（BLEU1-BLEU4衰减18.7%），拟议模型从BLEU1到BLEU4的数值衰减幅度最小（仅衰减17.9%），说明拟议模型不仅能精准匹配单个关键词，还能生成更连贯的长句。METEOR指标通过考量同义词、词干匹配，更侧重评估生成文本的语义一致性，拟议模型的METEOR值达0.4604，显著高于第二名AoANet模型，证明拟议模型能更准确理解SAR图像中的语义信息，拟议模型的ROUGE0.8196是所有模型中最高，说明其生成的字幕能更全面覆盖参考描述中的核心信息；CIDEr衡量生成文本的信息量的一致性，拟议模型远超其他对比模型，证明其生成的字幕与人工标注的语义一致性更强。拟议模型具有这种表现有两个主要原因，多尺度特征提取模块相较于其他模型（如HC-Net使用简单的CA注意力）能提取更全面的目标信息，多语义融合模块则兼顾图像语义信息；跨模态注意机制能够更好地对齐不同模态特征。

表1　在UCM-Caption数据集上的对比实验定量结果

Table 1　Quantitative results of comparative experiments on the UCM-Caption dataset

方法	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr
Soft-Attention	0.8103	0.7428	0.6856	0.6341	0.4123	0.7524	3.0057
Hard-Attention	0.8008	0.7453	0.6931	0.6410	0.4066	0.7459	2.8809
FC-Att	0.8174	0.7537	0.7027	0.6571	0.4123	0.7626	2.9118
SM-Att	0.8152	0.7552	0.7000	0.6472	0.4181	0.7522	2.9431
AoANet	0.8237	0.7564	0.7028	0.6571	0.4428	0.7841	2.9483
MLA	0.8320	0.7702	0.7146	0.6625	0.4338	0.7889	3.0450
MLAT	0.8578	0.8015	0.7494	0.6974	0.4377	0.7942	3.1580
HC-Net	0.8449	0.7849	0.7362	0.6946	0.4379	0.7916	3.1264
ours	**0.8691**	**0.8133**	**0.7611**	**0.7130**	**0.4604**	**0.8196**	**3.3543**

注：加粗字体表示各列最优结果。

如表2所示，本文拟议的模型在RSICD遥感图像字幕生成任务中表现最优，其在所有评价指标上均取得最高值，优于其他对比模型，Soft-Attention、FC-Att等模型因较简单的注意力机制，并不能很好的提取遥感图像的多尺度特征和有效对齐跨模态特征，因此指标较低；MLAT则通过卷积拼接的操作融合多尺度特征，使用LSTM虽一定程度上融合图像-文本跨模态特征，但大大提高了模型复杂度。HC-Net则没有拟议模型的多语义融合的优势，对图像语义信息的挖掘不够全面。

如表3所示，本文拟议模型在自建的SAR-Captions图像字幕数据集上表现同样最优，其在所有评价指标上均取得最高值，优于其他对比模型，同时，体现了我们所提的SAR图像字幕数据集的有效性，在一定程度上证明了光学图像字幕方法同样适用于SAR图像。

2.5　消融实验

为了验证本文拟议模块的有效性，我们在RSICD数据集和自建的SAR-Captions图像字幕数据集上进行了消融实验，验证了多尺度和多语义特征提取以及跨模态对齐模块的优势。

如表4所示，基础模型（Base）的BLEU4达0.7426，CIDEr达3.3030，整体性能偏高，数据具有样本分布均匀、语义复杂度低等特点。跨模态对齐模块和多尺度、多语义模块均对模型具有正向作用，其中，对比基线模型，加了跨模态对齐模块后，

表2　在RSICD数据集上的对比实验定量结果

Table 2　Quantitative results of comparative experiments on the RSICD dataset

方法	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr
Soft-Attention	0.6210	0.4576	0.3589	0.2927	0.2521	0.4704	0.7999
Hard-Attention	0.6061	0.4499	0.3536	0.2877	0.2409	0.4677	0.7495
FC-Att	0.5898	0.4404	0.3462	0.2829	0.2357	0.4636	0.7460
SM-Att	0.6015	0.4445	0.3486	0.2826	0.2394	0.4643	0.7547
AoANet	0.6467	0.4820	0.3797	0.3082	0.2610	0.4845	0.8528
MLA	0.6253	0.4676	0.3715	0.3068	0.2537	0.4809	0.8134
MLAT	0.6386	0.4649	0.3611	0.2933	0.2573	0.4752	0.8277
HC-Net	0.6326	0.4792	0.3815	0.3140	0.2586	0.4862	0.8605
Ours	**0.6645**	**0.5028**	**0.3988**	**0.3277**	**0.2724**	**0.5081**	**0.9223**

注：加粗字体表示各列最优结果。

表3　在SAR-Captions数据集上的对比实验定量结果

Table 3　Quantitative results of comparative experiments on the SAR-Captions dataset

方法	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr
Soft-Attention	0.8734	0.7993	0.7420	0.6965	0.4768	0.8157	3.3563
Hard-Attention	0.8741	0.8032	0.7475	0.7036	0.4807	0.8197	3.3699
FC-Att	0.8672	0.7989	0.7445	0.7015	0.4879	0.8200	3.1273
SM-Att	0.8873	0.8188	0.7637	0.7206	0.4872	0.8285	3.4733
AoANet	0.8847	0.8241	0.7749	0.7346	0.5072	0.8474	3.4778
MLA	0.9007	0.8472	0.8018	0.7653	0.5230	0.8617	3.5877
MLAT	0.9337	0.8960	0.8626	0.8361	0.5636	0.9007	4.1293
HC-Net	0.9263	0.8888	0.8548	0.8276	0.5626	0.8953	4.1250
MG-Transformer	0.9000	0.8447	0.7997	0.7641	0.5081	0.8595	3.9302
ours	**0.9299**	**0.9014**	**0.8766**	**0.8570**	**0.5936**	**0.9128**	**4.2606**

注：加粗字体表示各列最优结果。

表4　在SAR-Captions数据集上的消融实验定量分析

Table 4　Quantitative analysis of ablation experiments on the SAR-Captions dataset

方法	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr
Base	0.8814	0.8227	0.7778	0.7426	0.5079	0.8366	3.3030
CMAM	0.8878	0.8355	0.7931	0.7612	0.5188	0.8598	3.7133
MSFFM+HSIM	0.8991	0.8596	0.8301	0.8072	0.5570	0.8831	4.0180
CMAM+HSIM	0.9152	0.8726	0.8364	0.8078	0.5579	0.8832	4.0569
CMAM+MSFFM	0.9210	0.8779	0.8446	0.8185	0.5653	0.8946	4.1374
CMAM+MSFFM(无频域增强)	0.9083	0.8600	0.8193	0.7848	0.5391	0.8759	3.6897
ours	**0.9299**	**0.9014**	**0.8766**	**0.8570**	**0.5936**	**0.9128**	**4.2606**

注：加粗字体表示各列最优结果。

表5　在RSICD数据集上的消融实验定量分析

Table 5　Quantitative analysis of ablation experiments on the RSICD dataset

方法	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr
Base	0.6072	0.4443	0.3485	0.2841	0.2418	0.4634	0.7581
CMAM	0.6314	0.4686	0.3672	0.2986	0.2572	0.4796	0.8260
MSFFM+HSIM	0.6561	0.4919	0.3864	0.3137	0.2667	0.4979	0.8936
CMAM+HSIM	0.6530	0.4859	0.3804	0.3082	0.2647	0.4912	0.8607
CMAM+MSFFM	0.6359	0.4819	0.3843	0.3170	0.2597	0.4914	0.8554
CMAM+MSFFM(无频域增强)	0.6417	0.4766	0.3755	0.3080	0.2599	0.4813	0.8644
ours	**0.6645**	**0.5028**	**0.3988**	**0.3277**	**0.2724**	**0.5081**	**0.9223**

注:加粗字体表示各列最优结果。

BLEU1指标并未有较大提升,而BLEU4以及CIDEr指标提升较多,证明该模块具有强化语义匹配的作用,多尺度和多语义模块联合提取完善的图像特征,对比基线模型所有指标均有较大提升。在跨模态对齐的基础上,多尺度模块和多语义模块也大大提升了模型的生成效果。同时,我们对比了有无频域增强模块的MSFFM模块,试验结果表明频域增强大大提升了模型生成完整字幕的能力,在所有指标上均有提高。

如表5所示,基础模型(Base)的BLEU4为0.2841,CIDEr为0.7581,相对而言数据样本分布离散、语义复杂度高。对比基线模型,跨模态对齐模块提升了模型的整体性能,促进了解码器中视觉特征和文本特征的有效交互。而多尺度模块和多语义模块的联合作用在提升通过多模态信息弥补文本语义的不足,效果更突出。通过对比有无频域增强模块的MSFFM,可以看出频域增强模块在BLEU2-BLEU4、ROUGE等反映长程依赖与语义完整性的指标上取得明显提升,表明其有助于让模型更关注全局结构、纹理与语义信息,生成更连贯、更贴合图像内容的描述,而不是局部单词硬匹配。

2.6　可视化结果

为了直观地体现我们模型地优越性,我们分别在RSICD数据集和SAR-Caption数据集上展示了部分预测结果。从提供的SAR-Caption数据来看,本文拟议模型生成的船舶描述与GT(Ground Truth,真值)较为一致,显著优于Base(基础模型),如图7(a)、7(b)、7(e)和7(f)所示,基础模型虽然准确的表达了船舶数量,我们的方法则很好的表达了船舶的数量信息和位置信息,如图7(c)所示,我们的方法较基础模型区分了集群的舰船和单独的舰船,总体而言,基础模型的表述集中于数量,且较为简单,我们的方法则有效地集成了多尺度和多语义图像特征,增强了特征表示的能力,很好的匹配真实标注。

在RSICD数据集上,本文拟议的模型也展现了较好的预测结果,如图8(a)所示,拟议的模型识别出了一个棒球场,而基础模型仅提到操场,对比标注真值一个棒球场更符合图像信息,如图8(b)、8(c)和8(e)所示,拟议模型较基础模型都能预测更准确、完整的图像信息,图8(d)和8(f)表示拟议模型挖掘到了图像中的道路信息和桥梁信息,而这些并没有在标注真值中出现,说明我们的模型在生成字幕时具有良好的灵活性。实验结果表明,拟议的方法能够达到较高的预测准确性和完整性。

3　结　论

本文针对遥感图像字幕生成(RSIC)中特征提取不全面、跨模态信息对齐不足的核心问题,以及SAR图像缺乏专用字幕数据集的现状,开展了系统性研究。

主要工作包括三方面:一是设计多尺度特征融合模块与混合语义融合模块,分别通过注意力机制、频域增强整合多尺度纹理与结构信息,结合局部对象特征、全局语义特征丰富图像表达,缓解数据稀缺问题;二是构建跨模态对齐模块,融合LSTM隐状态与单词嵌入向量引导视觉特征选择,并设计跨模态特征对齐损失,缩小图文语义鸿沟;三是人工标注构

图7　SAR-Caption数据集中部分图像的可视化结果(Base表示基线模型,GT表示标注语句。标红表示拟议模型优于基线模型)

Fig. 7　Visualization results of some images in the SAR-Caption dataset (Base denotes the baseline model and GT stands for ground truth sentences; content highlighted in red indicates that the proposed model outperforms the baseline model)

建SAR-Captions数据集,通过"数据筛选-精准标注-结构化整合"流程构建数据集:图像筛选自公开SAR船舶目标检测数据集,聚焦海洋、海岸、海岛等典型场景,包含4000张SAR船舶图像及20000条标注语句。

实验结果表明,所提方法在UCM-Captions、RSICD及SAR-Captions三个数据集上,在BLEU1-BLEU4、METEOR等所有评价指标上均优于主流对比模型,其中SAR-Captions数据集上BLEU4达0.8570、CIDEr达4.2606,消融实验验证了各核心模块的关键作用,可视化结果显示生成字幕更贴合图像真实内容,能兼顾细节与全局信息,体现出良好的语义理解与灵活表达能力。

本文通过模块创新与数据集构建,有效解决了RSIC的核心痛点:MSFFM与HSIM模块实现了多尺度纹理、局部对象、全局语义的全面特征提取;CMAM模块缩小了视觉-文本语义鸿沟。研究不足在于,模型对复杂场景下多类地物的语义关联捕捉仍有提升空间,且数据集虽覆盖典型船舶场景,但地物类型多样性有待拓展。未来可进一步优化多模态特征融合策略,增强对复杂场景的适配能力,并扩充数据集的地物类别与场景覆盖范围,提升模型的泛化性能。

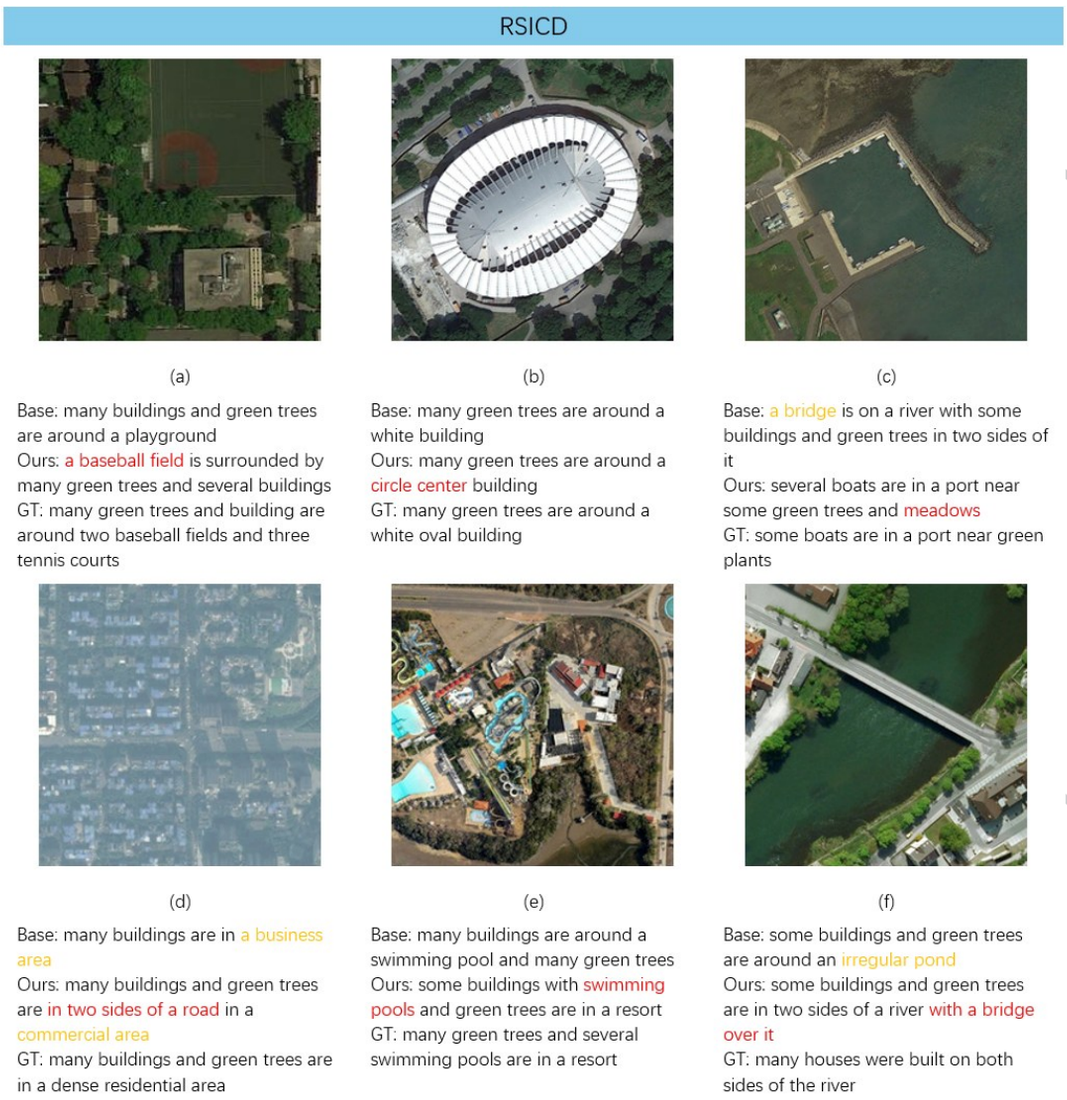

图8　RSICD数据集中部分图像的可视化结果(Base表示基线模型,GT表示标注语句。标红表示拟议模型优于基线模型或标注,标黄表示与标注或图像内容不符)

Fig. 8　Visualization results of some images in the RSICD dataset (Base denotes the baseline model, GT stands for ground truth sentences; content highlighted in red indicates that the proposed model outperforms the baseline model or the ground truth, and content highlighted in yellow indicates inconsistency with the ground truth or the image content).

参考文献(References)

Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S and Zhang L. 2018. Bottom-up and top-down attention for image captioning and visual question answering//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, Utah. IEEE: 6077-6086.[DOI: 10.1109/CVPR. 2018.00636]

Banerjee S and Lavie A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments//Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, Ann Arbor, Michigan. Association for Computational Linguistics: 65-72.

Chen K, Chen B, Liu C, Li W, Zou Z and Shi Z. 2024. Rsmamba: Remote sensing image classification with state space model. IEEE Geoscience and Remote Sensing Letters, 21: 1-5.[DOI: 10.1109/LGRS.2024.3407111]

Cheng Q, Huang H, Xu Y, Zhou Y, Li H and Wang Z. 2022. NWPU-captions dataset and MLCA-net for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-19.[DOI:10.1109/TGRS.2022.3201474]

Dhinesh A and Sumathy P. 2023. Remote Sensing Image Captioning (RSIC): A Technical Review//International Conference on Data Engineering and Machine Intelligence. Singapore: Springer Nature Singapore. 309-320.[DOI: 10.1007/978-981-97-7616- 0_22]

Du H and Liu X.2020. Image description generation method based on inhibitor learning. Journal of image and graphics, 25(2): 333-342.(杜海骏, 刘学亮.2020. 融合约束学习的图像字幕生成方法.中国图象图形学报, 25(2):333-342).[DOI:10.11834/jig.190222.]

Gao F, Jin X, Zhou X, Dong J and Du Q. 2025 MSFMamba: Multiscale feature fusion state space model for multi-source remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 63:1-16.[DOI: 10.1109/ TGRS.2025.3535622]

Huang L, Wang W, Chen J and Wei X. 2019. Attention on attention for image captioning//Proceedings of the IEEE/CVF international conference on computer vision. Seoul, Korea. IEEE: 4634-4643.[DOI: 10.1109/ICCV.2019.00473.]

Hou Z, Qu M, Li J, Ma S, Wang Y and Yang X. 2025. Lightweight CNN-Transformer combined network for real-time semantic segmentation. Journal of Image and Graphics, 30(7):2437-2450 (侯志强, 屈敏杰, 李俊歌, 马素刚, 王昀琛, 杨小宝. 2025. 轻量级CNN-Transformer相结合的实时语义分割网络. 中国图象图形学报, 30(7):2437-2450)[DOI:10.11834/jig.240527]

Lin J, Gao F, Shi X, Dong J and Du Q. 2023. SS-MAE: Spatial - spectral masked autoencoder for multisource remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 61 :1-14.[DOI: 10.1109/TGRS.2023. 3331717]

Lu X, Wang B, Zheng X and Li X. 2017. Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing, 56(4): 2183-2195.[DOI: 10.1109/TGRS.2017.2776321]

Li Y, Zhang X, Gu J, Li C, Wang X and Tang X. 2021. Recurrent attention and semantic gate for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-16.[DOI: 10.1109/TGRS.2021.3102590]

Liu J, Yao R, Gao N, Liang R and Chen P. 2025. VSRI: Visual Semantic Relational Interactor for Image Caption. Computer Science, 52(08):222-231.(刘健,姚任远,高楠,梁荣华,陈朋.2025. VSRI:基于视觉语义关系交互的图像字幕生成方法.计算机科学,52(08):222-231).

Liu C, Zhao R and Shi Z. 2022. Remote-sensing image captioning based on multilayer aggregated transformer. IEEE Geoscience and Remote Sensing Letters, 19: 1-5.[DOI: 10.1109/ LGRS. 2022. 3150957]

Li J and Liu J. 2025. Improved Ship Target Detection in SAR Images Based on YOLOv7//2025 6th International Conference on Computer Vision, Image and Deep Learning (CVIDL). Ningbo,China.IEEE: 18-21.[DOI:10.1109/CVIDL65390.2025.11085819]

Lin C Y. 2004. Rouge: A package for automatic evaluation of summaries//Text summarization branches out. Barcelona, Spain. Association for Computational Linguistics (ACL): 74-81.

Li Y, Fang S, Jiao L, Liu R and Shang R. 2020. A multi-level attention model for remote sensing image captions. Remote Sensing, 12(6): 939.[DOI: 10.3390/rs12060939]

Meena T, Vijaya J and Harsha B. 2025. Swin Transformers for Remote Sensing SAR Image Classification//2025 IEEE International Conference on Emerging Technologies and Applications (MPSec ICETA). Gwalior, India. IEEE: 1-6.[DOI: 10.1109/ MPSecICETA64837. 2025.11118726.]

Meng L, Wang J, Meng R, Yang Y and Xiao L. 2024. A multiscale grouping transformer with clip latents for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing, 62: 1-15.[DOI: 10.1109/TGRS.2024.3385500]

Papineni K, Roukos S, Ward T and Zhu W. 2002. Bleu: a method for automatic evaluation of machine translation//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA. Association for Computational Linguistics (ACL):311-318.

Qu B, Li X, Tao D and Lu X. 2016. Deep semantic understanding of high resolution remote sensing image//2016 International conference on computer, information and telecommunication systems (Cits). Kunming, China IEEE: 1-5. [DOI: 10.1109/CITS. 2016. 7546397.]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision[EB/OL].[2021-02-26]. https://arxiv.org/pdf/2103.00020.pdf

Ren Z, Gou S, Guo Z, Mao S and Li R. 2022. A mask-guided transformer network with topic token for remote sensing image captioning. Remote Sensing, 14(12): 2939.[DOI:10.3390/ rs14122939]

Ren S, He K, Girshick R and Sun J. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 39(6): 1137-1149.[DOI: 10.1109/TPAMI.2016.2577031]

Stefanini M, Cornia M, Baraldi L, Cascianelli S, Fiameni G and Cucchiara R. 2022. From show to tell: A survey on deep learning-based image captioning. IEEE transactions on pattern analysis and machine intelligence, 45(1): 539-559.[DOI: 10.1109/TPAMI. 2022.3148210]

Shi Z, Li C, Zhou L, Zhang Z, Wu C, You Z and Ren W. 2023. Survey on Transformer for image classification. Journal of Image and Graphics, 28(09):2661-2692 (石争浩,李成建,周亮,张治军,仵晨伟,尤珍臻,任文琦. 2023. Transformer驱动的图像分类研究进展. 中国图象图形学报, 28(09):2661-2692)[DOI:10.11834/jig.220799]

Tan Y, Tang P, Zhang L and Luo Y. 2021. From image to language: image captioning and description. Journal of image and graphics, 26(4): 727-750 (谭云兰, 汤鹏杰, 张丽, 罗玉盘. 2021. 从图像到语言:图像标题生成与描述. 中国图象图形学报, 26(4):727-750)[DOI:10.11834/jig.200177.]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A Kaiser L and Polosukhin I. 2017. Attention is all you need[EB/OL].

[2017-06-12].
https://arxiv.org/pdf/1706.03762.pdf

Vedantam R, Lawrence Zitnick C and Parikh D. 2015. Cider: Consensus-based image description evaluation[EB/OL]. [2015-06-03].
https://arxiv.org/pdf/1411.5726.pdf

Wang Q, Yang Z, Ni W, Wu J and Li Q. 2024. Semantic-spatial collaborative perception network for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing, 62: 1-12. [DOI: 10.1109/TGRS.2024.3502805]

Wang R, Wang L, Li C, Huo C and Chen J. 2023. IIQ-CNN-based cross-domain change detection of SAR images. Journal of Image and Graphics, 28(07): 2208-2220 (王蓉芳, 王良, 李畅, 霍春雷, 陈佳伟. 2023. 整型推理量化CNN的SAR图像跨域变化检测. 中国图象图形学报, 28(07): 2208-2220)[DOI: 10.11834/jig.211159]

Wang Y, Wang C, Zhang H, Dong Y and Wei S.2019. A SAR dataset of ship detection for deep learning under complex backgrounds. remote sensing, 11(7): 765.[DOI: 10.3390/rs11070765]

Xiao T, Liu Y, Huang Y, Li M and Yang G.2023. Enhancing multiscale representations with transformer for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 61: 1-16.[DOI: 10.1109/ TGRS.2023.3256064]

XU C, SU H, LI J, Liu Y, Yao L, Gao L, Yan W and Wang T.2022. RSDD-SAR: Rotated ship detection dataset in SAR images. Journal of Radars, 11(4): 581 - 599. (徐从安, 苏航, 李健伟, 刘瑜, 姚力波, 高龙, 闫文君, 汪韬阳. 2022. RSDD-SAR: SAR舰船斜框检测数据集. 雷达学报, 11(4): 581 - 599).[DOI: 10.12000/JR22007]

Yang Z, Li Q, Yuan Y and Wang Q. 2024. HCNet: Hierarchical feature aggregation and cross-modal feature alignment for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing, 62: 1-11.[DOI: 10.1109/ TGRS.2024.3401576]

Yao Y, Cheng G, Lang C, Yuan X, Xie X and Han J. 2024. Hierarchical mask prompting and robust integrated regression for oriented object detection. IEEE Transactions on Circuits and Systems for Video Technology, 34(12): 13071-13084.[DOI: 10.1109/TCSVT.2024.3444795]

Yu L, Hao J and Zuo L. 2025. Supervised attention-based oriented object detection in remote sensing images. Journal of Image and Graphics, 30(03): 0696-0709 (余凌霄, 郝洁, 左量. 2025. 基于监督注意力的遥感图像定向目标检测. 中国图象图形学报, 30(03): 0696-0709)[DOI: 10.11834/jig.240247]

Zhang C, Su J, Ju Y, Lam K and Wang Q. 2023. Efficient inductive vision transformer for oriented object detection in remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 61: 1-20.[DOI: 10.1109/TGRS.2023.3292418]

Zhang X, Li X, An J, Gao L Hou B and Li C. 2017. Natural language description of remote sensing images based on deep learning//2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Fort Worth, Texas, USA. IEEE: 4798-4801.[DOI: 10.1109 /IGARSS.2017.8128075.]

Zhang X, Li Y, Wang X, et al. 2023. Multi-source interactive stair attention for remote sensing image captioning. Remote Sensing, 15(3): 579.[DOI: 10.3390/rs15030579]

Zhang X, Wang X, Tang X, Zhou H and Li C. 2019. Description generation for remote sensing images using attribute attention mechanism. Remote Sensing, 11(6): 612.[DOI: 10.3390/ rs11060612]

Zhang Z, Zhang W, Yan M, Gao X, Fu K and Sun X. 2021. Global visual feature and linguistic state guided attention for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-16. [DOI: 10.1109/TGRS.2021. 3132095]

Zhao R, Shi Z and Zou Z. 2021. High-resolution remote sensing image captioning based on structured attention. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-14. [DOI: 10.1109/TGRS.2021.3070383]

Zhao K and Xiong W. 2024. Cooperative connection transformer for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing, 62: 1-14. [DOI: 10.1109/ TGRS. 2024.3360089]

Zhou J., Shi S and Wang H. 2025. A Survey of Image Caption Generation Based on Deep Learning. Software Guide, 24(01): 211-220. (周峻宇, 施水才, 王洪俊. 2025. 基于深度学习的图像字幕生成综述. 软件导刊, 24(01): 211-220).

作者简介

周凯立, 2002年生, 男, 硕士研究生, 主要研究方向为深度学习遥感图像处理。E-mail: 1783829175@qq.com。

王鹏, 通信作者, 男, 副教授, 博士生导师, 主要研究方向为深度学习遥感图像处理。E-mail: Pengwang_B614080003@nuaa.edu.cn。

程剑, 男, 教授, 博士生导师, 主要研究方向为天基频谱遥感载荷及感存算网一体化。E-mail: chengjian_nuaa@nuaa.edu.cn。