中图法分类号:TP18;TP75　　**文献标识码**:A　　**文章编号**:1006-8961(2026)09-3293-15

论文引用格式:Yu J X, Fu Y J and Liu J. 2026. Bi-temporal feature enhancement-guided multi-relational difference coupling for remote sensing change captioning. Journal of Image and Graphics, 31(9):3293-3307(于洁潇, 付雨杰, 刘婧. 2026. 双时相特征增强引导的多关系差异耦合遥感变化字幕生成. 中国图象图形学报, 31(9):3293-3307)[DOI:10.11834/jig.260171]

双时相特征增强引导的多关系差异耦合遥感变化字幕生成

于洁潇,付雨杰,刘婧*

天津大学电气自动化与信息工程学院,天津300072

摘　要:**目的**　针对遥感变化字幕(remote sensing change captioning,RSCC)生成任务中关键变化区域表征不足、双时相差异关系建模不充分以及模型复杂度较高等问题,提出一种双时相特征增强引导的多关系差异耦合遥感变化字幕生成方法。**方法**　采用基于ResNet-50(residual network)的RemoteCLIP(remote sensing constrastive language-image pre-training),即RemoteCLIP-RN50,提取双时相遥感影像深层语义特征,在此基础上,通过双时相特征增强策略对关键变化区域和重要语义通道进行自适应强化;随后构建多关系差异耦合框架,对双时相原始特征、绝对差异、乘积交互和相似性信息进行关系建模,以增强变化语义表达能力,最后利用文本解码器实现变化描述语句生成。**结果**　在LEVIR-CC数据集的实验结果表明,所提方法的双语评估替补指标BLEU-1为83.62%,BLEU-4为60.22%,基于最长公共子序列的面向召回摘要评估指标ROUGE-L(recall-oriented understudy for gisting evaluation-longest common subsequence)为64.94%,基于共识的图像描述评估指标CIDEr(consensus-based image description evaluation)为128.58%。其中BLEU-1和CIDEr优于对比方法;同时,本文方法参数量为41.50 M,低于多种经典及近年代表方法,体现出较好的性能—复杂度平衡。在DUBAI-CC数据集的补充实验中,本文方法取得的BLEU-1为63.75%、BLEU-4为34.14%、ROUGE-L为56.62%、CIDEr为90.09%,其中BLEU-4、ROUGE-L和CIDEr均取得最优结果,说明所提方法具有一定跨数据集适用性。消融实验表明,双时相特征增强策略和多关系差异耦合单元均能有效提升变化描述性能;进一步的关系项协同实验表明,在双时相特征增强后,完整多关系耦合在BLEU-1、ROUGE-L和CIDEr上取得最佳效果,说明特征增强有助于提升不同差异关系信息之间的互补表达能力。**结论**　本文方法围绕双时相变化表征过程开展针对性设计,在保持模型结构相对简洁的前提下有效提升了遥感变化字幕生成性能,并在不同数据集上表现出较好的适用性。

关键词:遥感变化字幕(RSCC);双时相遥感影像;特征增强;多关系差异耦合;变化描述

Bi-temporal feature enhancement-guided multi-relational difference coupling for remote sensing change captioning

Yu Jiexiao, Fu Yujie, Liu Jing*

School of Electrical and Information Engineering, Tianjin University, Tianjin　300072, China

Abstract: **Objective** Remote sensing change captioning (RSCC) aims to automatically generate natural language descrip-

收稿日期:2026-04-13;**修回日期**:2026-06-02;**预印本日期**:2026-06-09

*通信作者:刘婧　jliu_tju@tju.edu.cn

基金项目:国家自然科学基金项目(62371333);自动目标识别全国重点实验室基础研究基金项目(WDZC20265290201)

Supported by: National Natural Science Foundation of China (62371333); Fundamental Research Foundation of the National Key Laboratory of Automatic Target Recognition (WDZC20265290201)

tions for changes that occur between remote sensing images acquired over the same geographic region. Unlike traditional change detection, which mainly focuses on determining whether changes occur and where they are located, RSCC further requires the model to understand the semantic category, spatial position, and transformation relationship of changed objects. As a result, RSCC becomes a challenging task that integrates visual change perception, bitemporal feature interaction, and language generation. Thus far, existing methods have made considerable progress by introducing attention mechanisms, Transformer-based structures, and generative modeling strategies. However, several problems remain. First, the deep features extracted by visual backbones usually contain a large amount of unchanged background information, while the truly changed regions occupy only a small portion of the image. As a result, key changed regions and discriminative semantic channels may not be sufficiently highlighted before difference modeling. Second, the relationship between bitemporal features is not limited to simple subtraction. The original pre- and post-change features, change magnitudes, multiplicative interactions, and semantic similarities may all contribute to change understanding. Directly relying on a single difference representation or simple feature concatenation may be insufficient to describe complex changes. Third, some recent methods have been shown to achieve better performance by increasing model complexity, but the balance between captioning performance and model size remains important for practical remote sensing applications. To address the abovementioned problems, this paper proposes an RSCC framework guided by bitemporal feature enhancement and multirelation difference coupling. **Method** The proposed framework follows an encoder-decoder structure. RemoteCLIP-RN50, which is a RemoteCLIP with a residual network-50 backbone, is adopted as the shared visual feature extractor to encode pre- and post-change remote sensing images. The two temporal images are processed by the same backbone to ensure that their visual features are represented in a unified semantic space. After feature extraction, a bitemporal spatial-channel enhancement (BSCE) strategy is introduced before difference modeling. The BSCE strategy then performs channel enhancement and spatial enhancement on the features of both temporal images. Channel enhancement uses global contextual information to recalibrate semantic channels, thereby allowing channels related to changed objects to obtain stronger responses. Spatial enhancement further emphasizes local regions with change potential and suppresses irrelevant background responses. In this way, the enhanced bitemporal features provide clearer and more discriminative inputs for subsequent difference modeling. On this basis, a multirelation difference coupling (MRDC) unit is constructed. Instead of using only simple subtraction, MRDC jointly models the original bitemporal features, absolute difference, multiplicative interaction, and cosine similarity. Here, absolute difference represents local change magnitude between the two temporal features. Multiplicative interaction captures coresponse and feature interaction between the two temporal images. Cosine similarity describes semantic consistency between corresponding spatial positions. These complementary relationships are concatenated and compressed through feature fusion layers to obtain a compact visual change representation. The fused feature map is then flattened into a visual memory sequence. Finally, a Transformer-based text decoder is employed to generate change captions autoregressively. During training, the model is optimized using cross-entropy loss under the teacher-forcing strategy. During inference, greedy search is adopted to generate the final change description. **Result** Experiments on the LEVIR-CC dataset reveal that the proposed method achieves 83.62% in BLEU-1, 60.22% in BLEU-4, 64.94% in ROUGE-L, and 128.58% in CIDEr, with BLEU-1 and CIDEr outperforming the comparison methods. Meanwhile, the proposed method contains 41.50 M parameters—a value that is lower than several classic and recent representative methods, demonstrating a favorable balance between performance and model complexity. Additional experiments on the DUBAI-CC dataset show that the proposed method achieves 63.75% in BLEU-1, 34.14% in BLEU-4, 56.62% in ROUGE-L, and 90.09% in CIDEr. The best performance is obtained in BLEU-4, ROUGE-L, and CIDEr, thus indicating its applicability across different datasets. Ablation studies demonstrate that both the bitemporal feature enhancement strategy and the MRDC unit improve change captioning performance. Further relation synergy experiments reveal that, after bitemporal feature enhancement, complete multirelation coupling achieves the best performance in BLEU-1, ROUGE-L, and CIDEr, thus indicating that feature enhancement helps improve the complementary representation of different difference relations. **Conclusion** The proposed method designs a targeted bitemporal change representation process for RSCC. By introducing BSCE before difference modeling and constructing MRDC to jointly represent multiple change relations, the method improves the discriminability of visual change features and provides more effective visual cues for language generation. Furthermore, experimen-

tal results on LEVIR-CC and DUBAI-CC demonstrate that the proposed method not only achieves a favorable balance between captioning performance and model complexity but also shows good applicability across different datasets. The ablation, relation decomposition, synergy, and visualization analyses further confirm that the performance gain is achieved from the cooperation between bitemporal feature enhancement and MRDC. Although the proposed method still has room for improvement in fine-grained word-level matching and complex scene description, it provides a practical and relatively simple framework for RSCC. Future works will explore stronger language decoding strategies, cross-dataset semantic alignment, multidataset joint training, and lightweight deployment to further improve the robustness and generalization ability of RSCC models.

Key words: remote sensing change captioning(RSCC); bi-temporal remote sensing image; feature enhancement; multi-relation difference coupling; change description

0 引　言

现代对地观测系统能够持续获取大范围、多时相遥感影像，为地表变化监测、灾害评估、城市扩张分析和土地利用动态更新等应用提供重要支撑。长期以来，变化检测一直是遥感时序分析中的核心任务之一，其目标主要集中于判别“是否发生变化”以及“变化发生在何处”。然而，随着人机交互需求和遥感智能应用场景的不断拓展，仅输出变化区域或变化掩膜已难以满足实际应用需要。相较于传统变化检测结果，自然语言形式的变化描述能够进一步回答“发生了什么变化”、“从什么变成了什么”以及“变化位于何处”等更高层次问题，因此遥感变化字幕(remote sensing change captioning，RSCC)生成逐渐成为多时相遥感影像理解的重要研究方向。

遥感变化字幕生成任务通常以同一区域前后时相遥感影像对为输入，以描述地表变化的自然语言句子为输出，兼具变化检测与图像描述两类任务特点。从更广义的研究脉络来看，变化描述任务最早可追溯到相似图像对差异文本生成研究。Jhamtani和Berg-Kirkpatrick(2018)系统讨论了相似图像对差异描述问题，为后续变化字幕生成奠定了基础；随后，Park等人(2019)提出鲁棒的变化字幕生成任务与双重动态注意力(dual dynamic attention，DUDA)模型，推动了通用变化字幕生成研究的发展。此后，Hosseinzadeh和Wang(2021)从辅助任务学习角度改进图像变化描述，Qiu等人(2021)进一步提出多变化字幕Transformer(multi-change captioning Transformer，MCCFormer)，将Transformer引入多变化区域描述与定位，Yao等人(2022)和Guo等人(2022)又分别从预训练、对比学习以及对比语言—图像预训练(contrastive language-image pre-training，CLIP)迁移等角度推进了图像差异描述研究。

在遥感领域，变化字幕生成的发展建立在遥感图像描述研究的长期积累之上。Shi和Zou(2017)较早探讨了遥感图像自然语言描述生成问题，Lu等人(2018)系统分析了遥感图像描述模型与数据集，Wang等人(2021)和Li等人(2022)则分别从词句层级建模和循环注意力机制角度推动了遥感图像描述研究的发展。在此基础上，Chouaf等人(2021)首先将变化描述思想引入双时相遥感影像分析，Hoxha等人(2022)进一步将变化字幕生成明确为多时相遥感影像分析的新范式。随后，Liu等人(2022)提出LEVIR-CC数据集与RSICCformer(remote sensing image change captioning)，Chang和Ghamisi(2023)提出Chg2Cap，Liu等人(2023)提出渐进式尺度感知网络(progressive scale-aware network，PSNet)，Ferrod等人(2024)进一步探索了兼顾字幕生成与检索的多模态变化描述框架，表明遥感变化字幕生成正由单一描述任务向更广义的双时相视觉—语言联合建模方向拓展。国内相关研究也开始关注遥感图像字幕生成中的多尺度语义融合与跨模态表达问题，例如周凯立等人(2026)提出多尺度多语义融合协同的遥感图像字幕生成方法，为遥感图像语义描述研究提供了有益参考。

近年来，生成式建模思想也开始被引入遥感变化字幕生成任务。Yu等人(2025)提出遥感变化字幕生成扩散概率模型(diffusion probabilistic model for change captioning in remote sensing images，Diffusion-RSCC)，将扩散概率模型用于遥感变化字幕生成，通过条件噪声预测和跨模态特征融合建模真实描述分布，以缓解长时间跨度遥感影像中像素级差异对变化语义定位和描述生成的影响。该类方法表明，遥

感变化字幕生成正在从传统特征差异建模进一步发展到生成式语义建模阶段，但复杂生成模型通常也会带来更高的建模和推理开销。因此，如何在保持模型结构相对简洁的前提下获得有效的双时相变化表示，仍是该任务中值得进一步研究的问题。

随着视觉语言预训练模型的发展，面向遥感场景的基础模型也开始被引入到变化描述任务。Liu等人（2024）提出的遥感对比语言—图像预训练模型（remote sensing contrastive language-image pre-training，RemoteCLIP）表明，基于遥感领域大规模图文对预训练得到的视觉骨干能够为下游任务提供更具领域适应性的语义表征，这为遥感变化字幕生成中的双时相特征提取提供了新的技术条件。现有方法虽然通过多尺度差异感知、多注意力交互以及扩散式建模等机制不断提升变化描述能力，但部分方法对模型规模和特征交互复杂度的依赖也随之增加。相比之下，围绕双时相特征本身进行有效增强，并对变化关系进行有针对性的耦合建模，是一种更简洁且具有实际意义的研究思路。

尽管现有研究已在遥感变化字幕生成方面取得了较大进展，但在双时相变化建模方面仍存在两方面不足：一方面，遥感影像中目标尺度变化明显，小目标和局部变化区域容易受到背景复杂性与特征表示不足的影响，已有遥感影像小目标检测研究也指出，特征表示瓶颈和前背景混淆是影响遥感智能解译的重要因素（袁翔 等，2023）。因此，骨干网络提取的原始特征往往包含较多背景冗余响应，关键变化区域和高判别性语义通道难以被充分突出，导致后续描述生成受到干扰；另一方面，双时相特征之间并非只存在简单差分关系，变化前信息、变化后信息、变化幅度、交互响应和语义相似性等因素都会影响变化语义表达。若仅依赖单一差分或简单特征拼接，难以充分刻画复杂变化关系；而不同关系信息之间也可能存在冗余或互补关系，需要结合特征增强过程进行合理耦合。因此，如何在保持模型结构相对简洁的前提下，增强关键双时相特征并有效融合多类差异关系，是遥感变化字幕生成任务中值得关注的问题。

对于遥感变化字幕生成任务而言，仅感知“哪里不同”并不足以支撑准确描述，还需要进一步理解“变化前后是什么关系”以及“哪些关系信息真正有助于语言生成”。因此，围绕双时相特征增强、差异关系建模和关系项协同作用开展针对性设计，具有明确的研究意义。

基于上述认识，本文提出一种双时相特征增强引导的多关系差异耦合遥感变化字幕生成方法。首先，采用RemoteCLIP-RN50作为共享视觉骨干提取前后时相深层特征，并通过双时相特征增强策略强化关键变化区域与重要语义通道响应；随后，构建多关系差异耦合框架，在保留双时相原始特征的基础上，联合建模绝对差异、乘积交互和余弦相似性信息，以获得更具判别性的变化表示；最后，通过文本解码器实现变化描述语句生成。进一步地，本文通过协同作用实验分析不同关系项在双时相特征增强后的实际贡献。实验结果表明，在引入双时相特征增强后，完整多关系耦合在多项指标上取得最佳效果，说明特征增强能够提高多类关系信息之间的互补表达能力。LEVIR-CC数据集和DUBAI-CC数据集实验表明，本文方法表现优异，验证了所提方法在不同遥感变化场景下的适用性。

1 研究方法

遥感变化描述任务旨在根据同一区域前后时相遥感影像，自动生成能够表征地表变化语义的自然语言句子。与传统变化检测主要关注“变化位置”和“变化范围”不同，变化描述不仅需要识别变化区域，还需要刻画变化前后的语义关系，并将其转化为可读的文本表达。针对现有方法在关键变化区域表征不足、双时相差异关系建模不充分等问题，构建了一种由共享视觉骨干、双时相空间通道增强单元（bi-temporal spatial-channel enhancement，BSCE）、多关系差异耦合单元（multi-relation difference coupling，MRDC）和文本解码器组成的遥感变化描述框架。该框架首先利用共享参数的视觉骨干提取双时相深层特征；然后通过空间与通道两个维度对双时相特征进行自适应增强，进一步结合原始特征、绝对差异、乘积交互和相似性信息进行差异耦合；最后将得到的变化表示输入文本解码器，逐词生成变化描述语句。模型整体结构可概括为“双时相特征提取—特征增强—差异耦合—文本生成”4个阶段。

1.1 整体框架

设输入的前时相遥感影像和后时相遥感影像分别为 $\boldsymbol{I}^A$ 和 $\boldsymbol{I}^B$，对应输出变化描述序列为 $\boldsymbol{Y}=$

$\{y_1, y_2, \cdots, y_T\}$。针对遥感变化描述任务中“变化区域小、背景冗余多、前后时相语义关系复杂”的特点，构建了一种端到端的编码—解码框架，如图 1 所示。首先，采用共享参数的基于残差网络 50（residual network-50，ResNet-50）的 RemoteCLIP（remote sensing constrastive language-image pre-training），即 RemoteCLIP-RN50 视觉骨干分别提取前后时相深层语义特征，并通过线性投影将骨干输出统一映射到同一隐空间维度，从而为后续跨时相建模提供一致的特征基础。随后，为突出变化相关区域并抑制背景噪声，在双时相特征进入差异建模前，引入双时相空间通道增强单元，从空间位置和语义通道两个维度对前后时相特征进行自适应重标定，使网络能够更加聚焦于潜在变化区域及其判别性语义响应。进一步地，在差异建模阶段，将增强后的双时相特征送入多关系差异耦合单元，在保留前后时相原始特征的同时，构建多关系差异耦合框架，从绝对差异、乘积交互和相似性等多个角度刻画双时相特征之间的关系，以增强变化语义表达能力。此外，为进一步分析不同关系项在双时相特征增强后的作用，本文在实验部分对绝对差异、乘积交互和余弦相似性进行协同作用分析，以验证完整多关系耦合与双时相特征增强之间的互补性。最后，将耦合后的融合特征展平为视觉记忆序列，并输入文本解码器；解码器先利用带因果掩码的自注意力建模句内依赖，再通过跨模态注意力从视觉记忆中提取与当前词生成相关的变化信息，最终逐词输出变化描述语句。整个模型能够在兼顾变化区域定位的同时，更充分地表达双时相语义变化关系。模型整体可表示为

$$\boldsymbol{Y} = D\Big(C\Big(S\big(\varepsilon(\boldsymbol{I}^A), \varepsilon(\boldsymbol{I}^B)\big)\Big)\Big) \tag{1}$$

式中，$\varepsilon(\cdot)$表示共享视觉编码器特征提取过程；$S(\cdot)$表示双时相空间通道增强；$C(\cdot)$表示多关系差异耦合；$D(\cdot)$表示文本解码过程；$\boldsymbol{Y}$表示最终生成的变化描述序列。

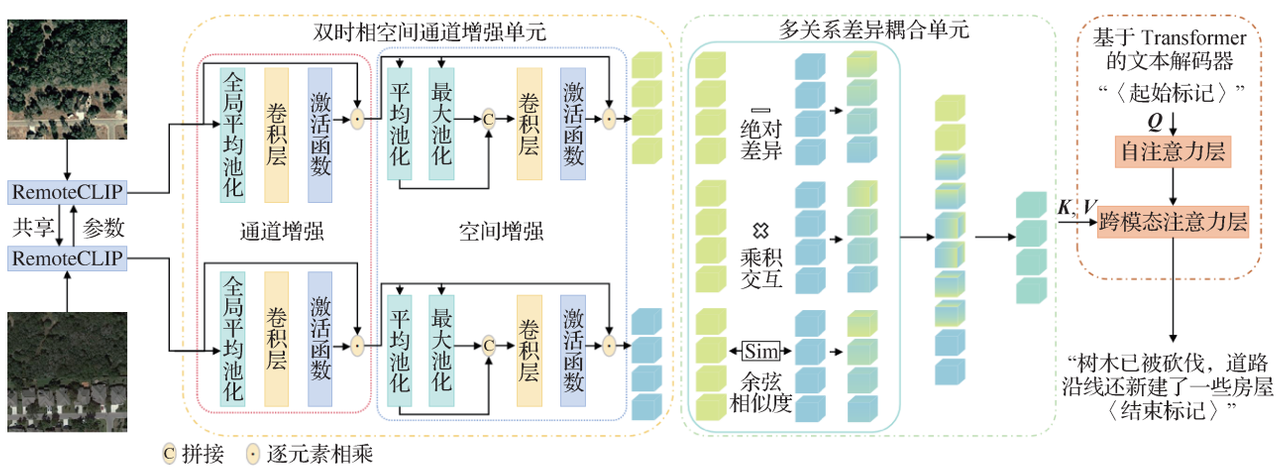


图 1　整体框架图

Fig. 1　Overall framework diagram

1.2　双时相视觉特征提取

为充分挖掘双时相遥感影像中的高层语义信息，采用共享参数的双分支视觉编码方式分别对前后时相影像进行特征提取。当前模型选用 RemoteCLIP-RN50 作为视觉骨干，并利用其遥感领域预训练权重增强对遥感场景语义的表征能力。设共享视觉编码器为$\varepsilon(\cdot)$，则前后时相特征分别表示为

$$\begin{cases} \boldsymbol{F}^A = \varepsilon(\boldsymbol{I}^A) \\ \boldsymbol{F}^B = \varepsilon(\boldsymbol{I}^B) \end{cases} \tag{2}$$

式中，$\boldsymbol{F}^A, \boldsymbol{F}^B \in \mathbf{R}^{C\times H\times W}$分别表示前时相和后时相的深层特征图；$C$、$H$、$W$分别表示通道数和空间尺寸。

共享骨干提取方式具有两方面优势：一方面，前后时相特征由同一组参数生成，可保证双时相特征处于统一语义嵌入空间，减少由独立编码器带来的表示偏差；另一方面，参数共享能够避免模型规模过大，有利于提高训练稳定性和泛化能力。

1.3　双时相空间通道增强单元

遥感影像中的变化通常具有明显的局部性和语义稀疏性，若直接对骨干输出特征进行差异建模，容

易受到大面积不变背景和弱判别通道的干扰。为此，在双时相特征进入差异耦合前，引入双时相空间通道增强单元，对前后时相特征分别进行自适应重标定，以强化关键变化区域和重要语义通道的响应。当前实现中，该单元由通道增强和空间增强两部分串联构成，并分别作用于前后时相特征。需要说明的是，该单元借鉴了空间注意力与通道注意力的基本思想，其作用并不在于构建新的注意力形式，而是在双时相差异建模之前对前后时相特征进行一致性增强。与直接对单幅图像特征进行注意力重标定不同，本文将其嵌入双时相变化描述框架中，用于为后续多关系差异耦合提供更清晰的变化相关输入特征。

1.3.1 通道增强

给定任一时相特征 $\boldsymbol{F} \in \mathbf{R}^{C \times H \times W}$，首先通过全局平均池化压缩空间信息，获得通道级语义描述向量；随后通过两层 1×1 卷积和非线性激活生成通道权重 $\boldsymbol{M}_{\mathrm{C}}$，其计算过程可表示为

$$\boldsymbol{M}_{\mathrm{C}} = \sigma\left(\boldsymbol{W}_2\delta\left(\boldsymbol{W}_1\left(GAP(\boldsymbol{F})\right)\right)\right) \tag{3}$$

式中，$GAP(\cdot)$ 表示全局平均池化；$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 表示卷积映射；$\delta(\cdot)$ 表示整流线性单元（rectified linear unit，ReLU）激活函数；$\sigma(\cdot)$ 表示 sigmoid 函数。之后，利用通道权重对原始特征进行逐通道重标定，具体为

$$\boldsymbol{F}_{\mathrm{C}} = \boldsymbol{M}_{\mathrm{C}} \odot \boldsymbol{F} \tag{4}$$

式中，$\odot$ 表示逐元素乘法。

通道增强的目的是提升对变化语义更敏感通道的响应，抑制冗余背景通道和低贡献通道带来的干扰。

1.3.2 空间增强

在通道增强基础上，进一步通过空间注意力突出具有变化潜力的局部区域。具体地，对通道增强后的特征 $\boldsymbol{F}_{\mathrm{C}}$ 分别在通道维执行平均池化（Avg）和最大池化（Max），得到两幅单通道空间图；将其在通道维拼接后，利用 7×7 卷积生成空间权重图 $\boldsymbol{M}_{\mathrm{S}}$，具体为

$$\boldsymbol{M}_{\mathrm{S}} = \sigma\left(f^{7\times7}\left(\left[Avg\left(\boldsymbol{F}_{\mathrm{C}}\right); Max\left(\boldsymbol{F}_{\mathrm{C}}\right)\right]\right)\right) \tag{5}$$

式中，$[\cdot]$ 表示通道拼接；$f^{7\times7}$ 表示卷积核大小为 7×7 的卷积操作。最终增强特征可表示为

$$\boldsymbol{F}' = \boldsymbol{M}_{\mathrm{S}} \odot \boldsymbol{F}_{\mathrm{C}} \tag{6}$$

对前后时相特征分别执行上述操作，可获得增强后的双时相特征 $\boldsymbol{F}^{A'}$ 和 $\boldsymbol{F}^{B'}$。

与直接使用骨干输出相比，该过程能够在进入差异耦合前优先提升变化区域及其相关语义的显著性，为后续跨时相变化建模提供更具判别性的输入特征。

1.4 多关系差异耦合单元

仅依赖简单差分往往难以充分表达双时相之间复杂的语义变化关系。对于变化描述任务而言，模型不仅需要感知“哪里发生了变化”，还需要理解变化前后地物状态、变化幅度以及双时相语义关系。为此，本文构建多关系差异耦合框架，在保留双时相原始特征的基础上，从绝对差异、乘积交互和余弦相似性等角度刻画双时相特征关系，以增强变化语义表达能力。需要说明的是，不同关系项对变化描述性能的贡献并不完全一致。在实验部分进一步分析了不同关系项与双时相特征增强策略之间的协同作用，结果表明，在经过特征增强后，完整多关系耦合能够更充分发挥不同关系信息之间的互补性。因此，最终采用由原始双时相特征、绝对差异、乘积交互和余弦相似性共同组成的完整多关系差异耦合结构。

1.4.1 多关系差异构建

给定增强后的双时相特征 $\boldsymbol{F}^{A'}$ 和 $\boldsymbol{F}^{B'}$，首先构建绝对差异特征，具体为

$$\boldsymbol{F}_{\mathrm{d}} = \left|\boldsymbol{F}^{B'} - \boldsymbol{F}^{A'}\right| \tag{7}$$

该特征反映前后时相在同一空间位置上的显著变化幅度。进一步地，构建乘积交互特征，具体为

$$\boldsymbol{F}_{\mathrm{P}} = \boldsymbol{F}^{A'} \odot \boldsymbol{F}^{B'} \tag{8}$$

其能够刻画双时相之间的共现关系与一致性响应，有助于区分真实变化与轻微纹理扰动。与此同时，沿通道维计算双时相特征的余弦相似性图，具体为

$$\boldsymbol{M}_{\cos}(u,v) = \frac{\boldsymbol{F}^{A'}(u,v) \times \boldsymbol{F}^{B'}(u,v)}{\left\|\boldsymbol{F}^{A'}(u,v)\right\| \left\|\boldsymbol{F}^{B'}(u,v)\right\|} \tag{9}$$

式中，(u,v) 表示空间位置。余弦相似性越低，通常意味着对应位置发生语义变化的可能性越高。

1.4.2 差异耦合与特征压缩

为综合利用双时相原始增强特征与多类差异关系信息，首先将它们在通道维进行拼接，构造综合变化表示，具体为

$$\boldsymbol{F}_{\mathrm{cat}} = \left[\boldsymbol{F}^{A'}; \boldsymbol{F}^{B'}; \boldsymbol{F}_{\mathrm{d}}; \boldsymbol{F}_{\mathrm{P}}; \boldsymbol{M}_{\cos}\right] \tag{10}$$

式中，$[\cdot]$ 表示通道拼接。由于拼接后特征维数较高，采用 1×1 卷积对其进行维度压缩，并结合批归

一化、ReLU 激活和残差块进一步提炼融合表示，得到最终变化特征，具体为

$$\boldsymbol{F}_{\mathrm{f}} = R\left(\delta\left(BN\left(Conv_{1\times1}\left(\boldsymbol{F}_{\mathrm{cat}}\right)\right)\right)\right) \tag{11}$$

式中，$Conv_{1\times1}$表示 1 × 1 卷积；$BN(\cdot)$表示批归一化（batch normalization）；$R(\cdot)$表示残差映射。

与传统简单差分相比，该设计同时保留了变化前信息、变化后信息、变化幅度、交互响应和语义一致性等多维关系，有利于后续生成具有方向性和语义完整性的变化描述。后续协同作用实验进一步表明，虽然不同关系项的单独贡献存在差异，但在双时相特征增强后，完整多关系耦合能够取得更优效果，说明特征增强有助于提升不同关系信息之间的互补表达能力。

1.4.3　视觉记忆序列构建

多关系差异耦合后得到的融合特征$\boldsymbol{F}_{\mathrm{f}}$仍为二维特征图，难以直接输入序列式文本解码器。为此，将其按空间维展开为视觉 token 序列，具体为

$$\boldsymbol{V} = Flatten\left(\boldsymbol{F}_{\mathrm{f}}\right) \in \mathbf{R}^{HW\times C} \tag{12}$$

式中，*Flatten* 为空间展平操作；HW 个 token 对应融合特征图中的 $H \times W$ 个空间位置，每个 token 包含一个空间位置的融合变化语义。

1.5　文本解码器与描述生成

在获得视觉记忆序列 $\boldsymbol{V}$ 后，采用基于 Transformer 的文本解码器生成变化描述。解码器由词嵌入层、位置编码、自注意力解码层和跨模态注意力解码层组成。对于输入词序列$\boldsymbol{Y}_{<t}$，首先通过词嵌入层映射到连续语义空间，并叠加位置编码以保留序列顺序信息，具体为

$$\boldsymbol{X}_0 = PE\left(Emb\left(\boldsymbol{Y}_{<t}\right)\right) \tag{13}$$

式中，$Emb(\cdot)$表示词嵌入操作；$PE(\cdot)$表示位置编码。为保证文本生成满足自回归约束，解码器对输入序列施加上三角因果掩码，使当前时刻只能访问历史词而不能访问未来词。经过位置编码后的文本序列首先通过自注意力解码层，建模句内长距离依赖关系；随后通过跨模态注意力解码层，以文本隐状态为 query、以视觉记忆序列 $\boldsymbol{V}$ 为 key/value，与变化视觉表示进行交互，从而提取与当前词生成相关的视觉语义信息。该过程可表示为

$$\boldsymbol{H}_t^{(1)} = SelfAttn\left(\boldsymbol{X}_0\right) \tag{14}$$

$$\boldsymbol{H}_t^{(2)} = CrossAttn\left(\boldsymbol{H}_t^{(1)}, \boldsymbol{V}\right) \tag{15}$$

式中，$SelfAttn(\cdot)$表示文本自注意力；$CrossAttn(\cdot)$表示文本对视觉记忆的跨模态注意力。最后，经线性映射和 softmax 得到词表上的预测概率分布，具体为

$$P\left(y_t \middle| y_{<t}, \boldsymbol{V}\right) = softmax\left(\boldsymbol{W}_0\right)\boldsymbol{H}_t^{(2)} \tag{16}$$

实际上，解码器包含独立的文本自注意力层和跨模态注意力层，输出层与输入词嵌入层共享权重，以减少参数量并增强词语表示一致性。

训练阶段采用 Teacher forcing 策略，以真实前缀词序列预测下一个词；推理阶段以起始标记（START）作为初始输入，循环调用解码器，每一步取当前时刻最大概率词作为下一个输入，直到输出结束标记（END）或达到最大长度为止。当前测试脚本中，变化描述生成采用 Greedy search 完成，并利用最终生成句与参考句计算评价指标。

1.6　损失函数与训练策略

为实现端到端训练，采用基于序列移位的交叉熵损失对变化描述生成过程进行监督。设模型在时刻 t 的预测概率分布为 $P\left(y_t \middle| y_{<t}, \boldsymbol{V}\right)$，真实标签序列为 $\boldsymbol{Y}$，则描述生成损失可表示为

$$L_{\mathrm{cap}} = -\sum_{t=1}^{T}\log P\left(y_t \middle| y_{<t}, \boldsymbol{V}\right) \tag{17}$$

考虑到输入句子长度不一致，训练时对序列尾部进行填充，并且填充位置不参与损失计算。

综上，所构建的遥感变化描述方法通过“共享 RemoteCLIP-RN50 特征提取—双时相空间通道增强—多关系差异耦合—文本解码生成”的整体流程，将双时相遥感影像中的局部显著变化与高层语义差异统一编码为紧凑的视觉记忆序列，再借助自回归解码器实现自然语言变化描述生成。相较于仅依赖骨干特征或简单差分建模的方法，该框架更有利于同时兼顾变化区域感知和变化语义表达，为后续定量实验和可视化分析提供了明确的方法基础。

2　数据与实验设置

2.1　数据集介绍

本文在 LEVIR-CC（learning, vision, and remote sensing laboratory change captioning）和 DUBAI-CC（Dubai change captioning）两个公开遥感变化字幕生成数据集上开展实验，以验证所提方法的有效性与跨数据集适用性。

LEVIR-CC数据集是在LEVIR-CD双时相建筑变化检测数据基础上构建的大规模遥感变化描述数据集，面向双时相遥感图像与自然语言变化描述之间的语义建模任务。数据集中每个样本由一对不同时相获取的遥感图像和多条变化描述文本组成，文本内容主要围绕建筑物新增、消失和扩展等变化类型展开，从而能够为遥感变化描述任务提供图像—文本对齐监督。LEVIR-CC数据集共包含10 077对双时相遥感图像和50 385条变化描述文本，平均每对图像对应5条文本描述。数据集按照训练集、验证集和测试集进行划分，分别用于模型参数学习、模型选择和最终性能评估。

DUBAI-CC数据集用于进一步验证模型在不同遥感变化场景下的适用性。该数据集包含500组双时相遥感影像和2 500条变化描述文本，每组影像同样对应5条描述。与LEVIR-CC相比，DUBAI-CC在场景区域、影像尺度和描述分布方面存在差异，因此可作为跨数据集泛化实验的补充验证数据集。实验采用其公开划分方式，其中训练集、验证集和测试集分别包含300、50和150组样本。

在样本组织方式上，训练阶段采用多描述增强策略，即每幅双时相影像对应最多5条描述；当有效描述数量不足5条时，通过对已有描述进行重复采样补足，以提高监督信号密度。验证与测试阶段则保留同一样本的全部参考描述，用于多参考文本评价，从而更加客观地反映生成结果与人工标注之间的一致性。图像在输入模型前统一转换为红绿蓝（red-green-blue，RGB）格式，并完成张量化与归一化预处理。

2.2 评价指标

为全面评估模型生成变化描述的准确性与语言质量，实验采用双语评估替补指标（bilingual evaluation understudy，BLEU）、基于最长公共子序列的面向召回摘要评估指标（recall-oriented understudy for gisting evaluation-longest common subsequence，ROUGE-L）和基于共识的图像描述评估指标（consensus-based image description evaluation，CIDEr）作为定量评价指标。其中，BLEU-1和BLEU-4分别表示1元和4元n-gram匹配精度。上述指标分别从局部词汇匹配、长短语连续性、句子结构一致性以及整体语义相似性等多个角度衡量生成描述与参考描述之间的接近程度。

2.2.1 BLEU

BLEU指标通过比较生成句与参考句之间的n-gram重合程度衡量文本生成质量，其定义为

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N}\omega_n \log P_n\right) \tag{18}$$

式中，BP表示简短惩罚项；ω_n表示第n阶n-gram精度对应的权重；P_n表示第n阶修正n-gram精度；N表示采用的最大n-gram阶数。当$N=1$和$N=4$时，分别得到BLEU-1和BLEU-4。

2.2.2 ROUGE-L

ROUGE-L基于最长公共子序列（longest common subsequence，LCS）评价生成句与参考句之间的整体结构一致性。

ROUGE-L能够反映生成句与参考句在句子级结构上的一致程度，对描述生成任务中的整体表达质量具有较好的刻画能力。

2.2.3 CIDEr

CIDEr主要面向描述生成任务设计，通过比较生成句与多条参考句在n-gram层面的词频—逆文档频率（term frequency-inverse document frequency，TF-IDF）加权相似性衡量语义一致性。

CIDEr能够同时考虑多参考描述与词项区分度，对描述生成任务中的语义一致性评估更具代表性，因此通常可视为衡量变化描述质量的重要指标。

2.3 实验实现细节

实验在Linux操作系统下基于PyTorch深度学习框架实现，训练与测试均在NVIDIA GeForce RTX 4090 GPU上完成，显存为24 GB。所有实验均采用相同的硬件与软件平台，以保证不同模型和消融设置之间的可比性。

模型训练采用AdamW优化器，初始学习率设为1×10^{-4}，batch size设为64，训练轮数设为50，warmup比例设为0.025，并采用warmup后恒定学习率策略进行调度。为减小随机性带来的性能波动，实验固定随机种子为42。训练过程中，每个epoch后在验证集上计算平均损失；同时按固定间隔在验证集上执行变化描述推理，并计算BLEU-1、BLEU-4、ROUGE-L和CIDEr指标。综合考虑优化稳定性与生成质量，实验分别保存验证损失最优模型和CIDEr指标最优模型。

在训练监督方式上，采用自回归序列学习策略。输入描述序列经过移位后用于预测下一时刻词元，

填充位置不参与损失计算。测试阶段采用贪心搜索策略逐词生成变化描述，即以起始标记作为初始输入，在每个时刻选择当前概率最大的词作为下一时刻输入，直至生成结束标记或达到最大长度为止。最终将测试集生成结果与参考描述共同输入评价模块，得到各项定量指标，并同步输出可视化结果。

3 结果与分析

3.1 LEVIR-CC 数据集对比实验与复杂度分析

为验证所提方法在遥感变化字幕生成任务中的有效性，首先在 LEVIR-CC 数据集上开展对比实验。对比方法包括 Capt-Rep-Diff、Capt-Att、Capt-Dual-Att、DUDA、MCCFormer-S、MCCFormer-D 等经典变化字幕生成方法，同时加入 Diffusion-RSCC 和 RSICRC（remote sensing image change retrieval and captioning）等近年相关方法。为进一步分析模型性能与复杂度之间的关系，本文同时给出部分方法的参数量，并与本文方法进行比较。

表 1 给出了不同方法在 LEVIR-CC 数据集上的实验结果和参数量比较。可以看出，本文方法取得了 83.62% 的 BLEU-1、60.22% 的 BLEU-4、64.94% 的 ROUGE-L 和 128.58% 的 CIDEr。其中，BLEU-1 和 CIDEr 均取得最优结果，说明本文方法在变化关键词匹配和整体语义一致性方面具有较好的表现。与 RSICRC 相比，本文方法在 BLEU-1、BLEU-4、ROUGE-L 和 CIDEr 上分别提升了 4.07%、4.46%、2.79% 和 11.44%，表明双时相特征增强与多关系差异耦合能够有效提升变化描述质量。与 Diffusion-RSCC 相比，本文方法在 BLEU-1 和 CIDEr 上分别提升 0.04% 和 2.98%，在 BLEU-4 上略低 0.68%，这说明本文方法在整体语义一致性方面具有一定优势，但在长短语连续性和句子结构组织方面仍存在提升空间。特别是在 ROUGE-L 指标上，本文方法低于 DUDA、MCCFormer-D 和 Diffusion-RSCC 等方法，说明模型在句子级结构保持能力方面仍有进一步改进空间。从模型复杂度角度看，本文方法的参数量为 41.50 M，明显低于表 1 中已报告参数量的对比方法。例如，与 Capt-Rep-Diff、DUDA 和 MCCFormer-D 相比，本文方法参数量分别减少了 31.71 M、38.81 M 和 121.05 M。该结果表明，本文方法并非依赖大规模参数堆叠获得性能提升，而是在较小参数量条件下，通过双时相特征增强和多关系差异耦合实现了有效变化表征。

综合来看，本文方法在 LEVIR-CC 数据集上取得了较好的性能—复杂度平衡。虽然其在 BLEU-4 和 ROUGE-L 指标上并非最优，但在 BLEU-1、CIDEr 和参数量方面表现较优，说明所提出的双时相特征增强引导的多关系差异耦合框架能够在保持结构相对简洁的前提下获得具有竞争力的遥感变化描述性能。

3.2 DUBAI-CC 数据集对比实验

为进一步验证所提方法在不同遥感变化场景下的适用性，本文在 DUBAI-CC 数据集上补充开展泛化实验。与 LEVIR-CC 相比，DUBAI-CC 在场景区域、影

表 1 不同方法在 LEVIR-CC 数据集上的实验结果与参数量比较

Table 1 Experimental results and parameter comparison of different methods on the LEVIR-CC dataset

方法	BLEU-1/%	BLEU-4/%	ROUGE-L/%	CIDEr/%	参数量/M
Capt-Rep-Diff	72.90	47.41	65.64	110.57	73.21
Capt-Att	77.64	53.15	69.73	121.22	73.60
Capt-Dual-Att	79.51	57.46	70.69	124.42	75.58
DUDA	81.44	57.79	71.04	124.32	80.31
MCCFormer-S	79.90	56.68	69.46	120.39	162.55
MCCFormer-D	80.42	56.38	70.32	124.44	162.55
Diffusion-RSCC	83.58	**60.90**	**71.50**	125.60	–
RSICRC	79.55	55.76	62.15	117.14	–
本文	**83.62**	60.22	64.94	**128.58**	**41.50**

注：加粗字体表示各列最优结果。“–”表示未公开。

像尺度、变化类型和描述风格上均存在差异，因此能够用于检验模型在跨数据集条件下的变化描述能力。实验选取RSICCFormer、DUDA、MCCFormer-S、MCCFormer-D和Diffusion-RSCC等方法作为对比方法，结果如表2所示。

表2给出了不同方法在DUBAI-CC数据集上的实验结果。可以看出，本文方法取得了63.75%的BLEU-1、34.14%的BLEU-4、56.62%的ROUGE-L和90.09%的CIDEr。其中，BLEU-4、ROUGE-L和CIDEr均取得最优结果，说明本文方法在短语级生成质量、句子结构一致性和整体语义一致性方面具有较好的表现。与RSICCFormer相比，本文方法在BLEU-4、ROUGE-L和CIDEr上分别提升了2.86%、4.66%和23.55%；与MCCFormer-D相比，分别提升了4.66%、5.35%和23.58%。相较于Diffusion-RSCC，本文方法虽然在BLEU-1上低5.45%，但在BLEU-4、ROUGE-L和CIDEr上分别提升了0.84%、0.12%和1.39%。该结果表明，本文方法虽然在单词级匹配方面仍有提升空间，但在更能反映短语连续性和整体语义一致性的指标上表现更优。

表2 不同方法在DUBAI-CC数据集上的实验结果

Table 2 Experimental results of different methods on the DUBAI-CC dataset

/%

方法	BLEU-1	BLEU-4	ROUGE-L	CIDEr
RSICCFormer	67.92	31.28	51.96	66.54
DUDA	58.82	25.39	48.34	62.78
MCCFormer-S	52.97	22.57	43.29	53.81
MCCFormer-D	64.65	29.48	51.27	66.51
Diffusion-RSCC	**69.20**	33.30	56.50	88.70
本文	63.75	**34.14**	**56.62**	**90.09**

注：加粗字体表示各列最优结果。

综合来看，DUBAI-CC实验结果表明，本文方法在不同遥感变化场景下仍能够保持较好的变化描述能力，具有一定跨数据集适用性。这说明双时相特征增强能够有效突出关键变化区域和重要语义通道，多关系差异耦合能够增强双时相变化关系表达，从而提高模型在不同数据分布下的变化描述性能。同时，本文方法在BLEU-1指标上仍低于Diffusion-RSCC，说明模型在词汇层面的细粒度匹配能力仍有进一步提升空间，后续可通过更强的语言建模策略和跨数据集语义对齐进一步改善泛化性能。

3.3 模块级消融实验

为验证所提各模块在遥感变化字幕生成任务中的实际贡献，本文在LEVIR-CC数据集上开展模块级消融实验。实验以RSICRC框架作为Baseline，在此基础上分别加入双时相特征增强策略（BSCE）、多关系差异耦合单元（MRDC）以及二者组合，以分析不同模块对变化描述性能的影响。

表3给出了不同模块配置下的消融实验结果。可以看出，相比Baseline，单独加入BSCE后，BLEU-1、BLEU-4、ROUGE-L和CIDEr分别由79.55%、55.76%、62.15%和117.14%提升至82.07%、58.17%、63.38%和121.73%，说明双时相特征增强策略能够有效强化关键变化区域和重要语义通道，从而提升变化描述质量。单独加入MRDC后，模型性能同样得到提升，BLEU-1、BLEU-4、ROUGE-L和CIDEr分别达到81.07%、57.06%、63.03%和121.38%。该结果表明，相较于仅依赖基础变化描述分支，多关系差异耦合能够更充分地建模双时相特征之间的变化关系，增强模型对变化语义的表达能力。进一步地，同时引入BSCE和MRDC后，模型取得最优性能，BLEU-1、BLEU-4、ROUGE-L和CIDEr分别达到83.62%、60.22%、64.94%和128.58%。与Baseline相比，完整模型在4项指标上分别提升4.07%、4.46%、2.79%和11.44%；与单独加入BSCE或MRDC相比，也均取得进一步提升。该结果说明，BSCE和MRDC具有一定互补性：BSCE侧重于在差异建模前增强双时相特征质量，突出关键区域和判别性通道；MRDC则侧重于从关系层面对变化前后特征进行建模。二者结合后，能够在特征增强和关系建模两个层面共同提升变化描述性能。

综上，模块级消融实验验证了本文方法中双时相特征增强策略和多关系差异耦合单元的有效性。需要指出的是，BSCE主要作为特征增强策略，用于改善后续差异耦合的输入特征质量；模型性能的进一步提升主要来源于其与MRDC的协同作用，而非单一注意力结构本身。

3.4 多关系差异项拆解实验

为进一步分析多关系差异耦合单元（MRDC）中不同差异关系项的独立贡献，本文在不引入双时相特征增强策略（BSCE）的条件下，对绝对差异、乘积

表 3　不同模块配置下的消融实验结果

Table 3　Ablation results under different module configurations

/%

方法	BLEU-1	BLEU-4	ROUGE-L	CIDEr
Baseline	79.55	55.76	62.15	117.14
+BSCE	82.07	58.17	63.38	121.73
+MRDC	81.07	57.06	63.03	121.38
+MRDC + BSCE	**83.62**	**60.22**	**64.94**	**128.58**

注：加粗字体表示各列最优结果。

交互和余弦相似性 3 类关系进行拆解实验。实验首先以仅保留双时相原始特征 Pair-only 作为基础设置，然后分别加入单一关系项，并进一步分析去除某一关系项后的性能变化，以评估不同关系信息对变化描述性能的影响。

表 4 给出了不同关系项配置下的实验结果。可以看出，与 Pair-only 相比，单独引入绝对差异（Abs-only）和乘积交互（Prod-only）均能带来明显性能提升。其中，Abs-only 的 CIDEr 由 Pair-only 的 119. 57% 提升至 122. 94%，说明绝对差异能够有效反映双时相局部变化幅度，对变化区域表征具有积极作用。Prod-only 取得了 83. 31% 的 BLEU-1、60. 07% 的 BLEU-4、63. 93% 的 ROUGE-L 和 123. 88% 的 CIDEr，在单一关系项设置中表现最优，表明乘积交互关系能够较好地刻画双时相特征之间的协同响应和共现关系，对变化语义表达具有较强的贡献。

表 4　多关系差异项拆解实验结果

Table 4　Decomposition results of different relation terms in MRDC

/%

方法	BLEU-1	BLEU-4	ROUGE-L	CIDEr
Pair-only	81.59	57.87	62.60	119.57
Abs-only	82.31	58.85	63.41	122.94
Prod-only	**83.31**	**60.07**	**63.93**	**123.88**
Cos-only	81.02	55.81	62.27	119.61
w/o abs	82.36	58.66	63.38	123.21
w/o prod	82.35	58.60	63.45	123.18
w/o cos	82.31	58.89	63.36	123.29
Full MRDC	81.07	57.06	63.03	121.38

注：加粗字体表示各列最优结果。w/o 表示去除。

相比之下，仅保留余弦相似性（Cos-only）的 BLEU-1、BLEU-4、ROUGE-L 和 CIDEr 分别为 81. 02%、55. 81%、62. 27% 和 119. 61%，整体表现接近 Pair-only，但低于 Abs-only 和 Prod-only。这说明余弦相似性信息虽然能够从语义一致性角度刻画双时相特征关系，但单独使用时难以充分表达变化方向、变化幅度和目标类别等描述生成所需信息，因此对最终变化描述性能的提升有限。从去除单一关系项的实验结果来看，w/o abs、w/o prod 和 w/o cos 的 CIDEr 分别为 123. 21%、123. 18% 和 123. 29%，均高于 Full MRDC 的 121. 38%。这表明在未引入 BSCE 的情况下，直接叠加全部关系项并不一定取得最优结果，不同关系信息之间可能存在一定冗余或干扰。尤其是余弦相似性关系在原始特征基础上直接参与融合时，可能会引入与变化区域无关的语义一致性响应，从而影响多关系融合效果。

综合来看，MRDC 中不同关系项对变化描述性能的贡献并不完全一致。其中，乘积交互关系在单独关系建模中表现最优，是双时相差异关系建模中的重要组成部分；绝对差异关系能够提供稳定的变化幅度信息；余弦相似性关系单独使用时贡献相对有限。该实验说明，仅从 MRDC 本身来看，关系项的简单叠加并不能保证性能最优。因此，有必要进一步分析双时相特征增强策略是否能够改善多关系融合的输入质量，使不同关系项之间形成更有效的互补作用。后续 3. 5 节将进一步讨论 BSCE 与 MRDC 中不同关系项之间的协同关系。

3. 5　BSCE 与 MRDC 中多关系项的协同作用实验

3. 4 节结果表明，在未引入双时相特征增强策略（BSCE）的情况下，MRDC 中不同关系项对模型性能的贡献存在差异，且直接叠加全部关系项并不一定取得最优效果。为进一步分析 BSCE 是否能够改善多关系融合过程，本文在引入 BSCE 的基础上，对 MRDC 中不同关系项进行协同作用实验。具体而言，分别考察绝对差异、乘积交互和余弦相似性单独参与耦合时的性能，并进一步分析去除某一关系项后的模型表现，以验证不同关系项在增强特征基础上的贡献与互补性。

表 5 给出了 BSCE 与 MRDC 中不同关系项的协同作用实验结果。可以看出，在仅引入单一关系项的情况下，BSCE + Abs-only 取得了 82. 03% 的 BLEU-1、58. 61% 的 BLEU-4、63. 49% 的 ROUGE-L 和 124. 08%

的CIDEr，整体表现优于BSCE + Prod-only和BSCE + Cos-only。这说明在经过双时相特征增强后，绝对差异关系能够较直接地反映局部变化幅度，对变化区域表征和变化描述生成具有较稳定贡献。

相比之下，BSCE + Prod-only和BSCE + Cos-only的CIDEr分别为122.82%和121.08%，低于BSCE + Abs-only。结合3.4节结果可以发现，乘积交互关系在未引入BSCE时单独建模效果最好，而在BSCE增强特征基础上仅使用乘积交互并未取得最优结果。这表明，特征增强会改变不同关系项的作用方式：当关键区域和重要语义通道已被强化后，单一乘积交互关系虽然仍能提供协同响应信息，但难以完整表达变化幅度、变化方向和语义一致性等多维变化线索。

从去除单一关系项的实验结果来看，BSCE + w/o cos取得了83.07%的BLEU-1、60.51%的BLEU-4、64.44%的ROUGE-L和127.69%的CIDEr，整体性能明显高于各单一关系项设置，说明绝对差异和乘积交互之间具有较好的互补性。其中，BSCE + w/o cos在BLEU-4上略高于BSCE + Full，表明在部分短语级匹配方面，去除余弦相似性可能减少一定冗余干扰。然而，完整关系组合BSCE + Full在BLEU-1、ROUGE-L和CIDEr上均取得最优结果，分别达到83.62%、64.94%和128.58%，说明完整多关系耦合在变化关键词匹配、句子结构一致性和整体语义一致性方面更具优势。

综合来看，该实验进一步验证了BSCE与MRDC之间的互补关系。一方面，BSCE能够在差异耦合前突出关键变化区域和重要语义通道，为后续关系建模提供更高质量的输入特征；另一方面，MRDC能够从绝对差异、乘积交互和余弦相似性等不同角度刻画双时相变化关系。虽然不同关系项的贡献并不完全一致，且部分组合在个别指标上具有优势，但完整多关系耦合在CIDEr和ROUGE-L等更能反映整体语义一致性和句子结构质量的指标上表现最佳。因此，本文最终采用BSCE与完整MRDC相结合的结构作为最终模型。

表5　BSCE与MRDC中多关系项的协同作用实验结果

Table 5　Experimental results of the synergy between BSCE and different relation terms in MRDC

/%

方法	BLEU-1	BLEU-4	ROUGE-L	CIDEr
BSCE + Abs-only	82.03	58.61	63.49	124.08
BSCE + Prod-only	81.07	57.75	63.15	122.82
BSCE + Cos-only	80.97	56.73	62.43	121.08
BSCE + w/o abs	82.89	58.30	62.93	121.50
BSCE + w/o prod	82.77	58.55	64.44	125.74
BSCE + w/o cos	83.07	**60.51**	64.44	127.69
BSCE + Full	**83.62**	60.22	**64.94**	**128.58**

注：加粗字体表示各列最优结果。w/o表示去除。

3.6　可视化分析

为进一步直观分析所提方法的变化理解能力，本文选取典型样本对RSICRC与本文方法的生成结果和注意力响应进行可视化对比，具体如图2所示。图2(a)为人工参考描述中的一条，图2(b)为RSICRC生成的变化描述，图2(c)为本文方法生成的变化描述。为便于对比，图中采用下划线标注生成结果中与真实变化不一致或存在误检、漏检的描述内容，采用加粗标注本文方法中与真实变化区域及语义更加一致的描述内容。每组样本中，左上和右上分别为变化前、后图像，左下和右下分别为RSICRC和本文方法的注意力热图。

从第1组无变化样本可以看出，RSICRC错误生成了“a house has been built at the top”，将未发生显著变化的区域误判为新增房屋；而本文方法生成“the scene is the same as before”，与参考描述一致，说明本文方法能够更有效地抑制背景差异和成像条件变化带来的干扰，降低无变化场景中的误检风险。

第2组样本中，影像左上区域出现建筑物新增。RSICRC能够识别出建筑变化，但同时生成了“a house is built at the bottom”等不准确内容；本文方法生成“a house with a path to the forest appears in the top left”，不仅识别出新增房屋，还较准确地描述了其空间位置。对应注意力热图显示，本文方法在新增建筑附近形成较明显响应，说明其能够更好地聚焦于真实变化区域。

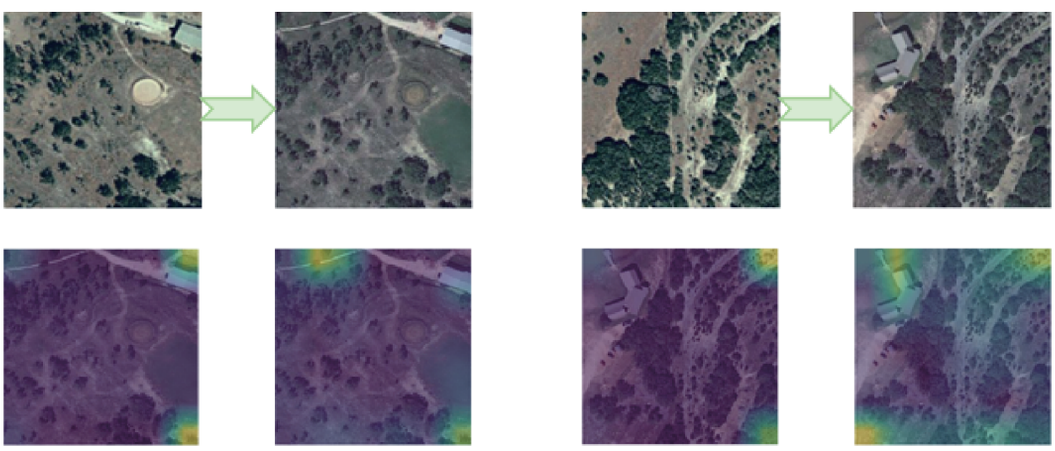
第3组样本中，道路附近出现新增房屋。RSICRC生成“two houses are built along the road”，虽然描述了道路附近建筑变化，但在数量和位置表达上不够准确；本文方法生成“a new house is built at the bottom right corner”，能够更准确地描述新增目标及其空间位置。该结果表明，本文方法在变化目标识别和空间位置表达方面具有更好的细粒度描述能力。

第4组样本中，道路结构发生变化。RSICRC生

成“the scene is the same as before”，未能识别道路变化；本文方法则生成“a road is built to replace the original one”，能够正确描述道路被替换或拓宽的变化现象。结合注意力热图可以看出，本文方法在道路交汇和变化区域附近具有更强响应，而RSICRC对变化区域关注不足。

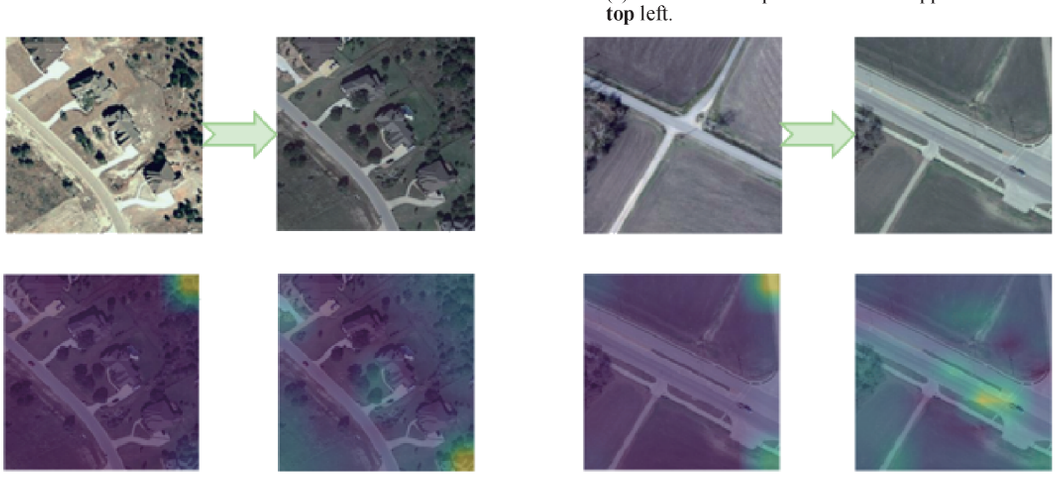
综合上述可视化结果可以看出，RSICRC在部分样本中容易出现无变化误判、变化目标遗漏以及空间位置描述不准确等问题。相比之下，本文方法能够更稳定地识别变化目标类别，并生成与真实变化区域更加一致的描述语句。这说明双时相特征增强策略有助于突出关键变化区域和重要语义通道，多关系差异耦合则能够进一步增强双时相变化关系表达，从而为文本解码器提供更具判别性的视觉线索。

同时也可以看到，注意力热图在部分样本中仍存在一定扩散现象，说明模型对复杂背景下细粒度变化边界的刻画仍有提升空间。总体而言，图2结果表明，本文方法不仅在定量指标上具有有效性，在定性可视化层面也表现出较好的变化区域关注能力和描述可解释性。



(a) the scene is the same as before.
(b) a house has been built at the top.
(c) **the scene is the same as before**.

(a) a building appears at the top and some trees disappear.
(b) a house is built at the top and a house is built at the bottom.
(c) a house with a path to the forest appears in **the top** left.



(a) a new house is built on the lower-right.
(b) two houses are built along the road.
(c) **a new house** is built at the bottom right corner.

(a) a road is replaced by a wider one.
(b) the scene is the same as before.
(c) **a road** is built to **replace** the original one.

图2　典型样本的变化描述与注意力可视化对比结果

Fig. 2　Comparison of change captions and attention visualizations on typical samples

4　结　论

针对遥感变化字幕生成任务中关键变化区域表征不足、双时相差异关系建模不充分以及模型复杂度较高等问题，本文提出了一种双时相特征增强引导的多关系差异耦合遥感变化字幕生成模型。该方法采用RemoteCLIP-RN50作为共享视觉骨干提取双

时相遥感影像深层语义特征，在差异建模前通过双时相特征增强策略强化关键变化区域与重要语义通道响应；随后，构建多关系差异耦合单元，联合建模双时相原始特征、绝对差异、乘积交互和余弦相似性信息，并通过Transformer解码器生成变化描述语句。

在LEVIR-CC数据集上的实验结果表明，本文方法取得了83.62%的BLEU-1、60.22%的BLEU-4、64.94%的ROUGE-L和128.58%的CIDEr，其中BLEU-1和CIDEr取得较优结果；同时，本文方法参数量为41.50 M，低于多种经典及近年代表方法，说明该方法在变化描述性能与模型复杂度之间取得了较好的平衡。在DUBAI-CC数据集上的补充实验中，本文方法在BLEU-4、ROUGE-L和CIDEr上取得最优结果，验证了所提方法在不同遥感变化场景下的适用性。

消融实验表明，双时相特征增强策略和多关系差异耦合单元均能有效提升变化描述性能，二者结合时取得最佳结果。进一步的关系项拆解与协同作用实验表明，不同差异关系项对模型性能的贡献存在差异；在引入双时相特征增强后，完整多关系耦合能够更充分发挥绝对差异、乘积交互和余弦相似性之间的互补作用。可视化结果进一步表明，本文方法能够更准确地关注真实变化区域，并生成与变化语义更加一致的描述文本。

总体而言，本文方法围绕双时相变化表征过程开展针对性设计，为结构相对简洁的遥感变化字幕生成模型设计提供了一种可行思路。但本文方法仍存在一定不足，例如在部分样本中注意力响应仍有扩散现象，且在词汇级匹配和复杂场景细粒度描述方面仍有提升空间。后续工作将进一步从跨数据集语义对齐、多数据集联合训练、更强语言解码策略以及轻量化部署等方面展开研究，以提升模型在复杂遥感变化场景下的泛化能力和实际应用价值。

参考文献(References)

Chang S Z and Ghamisi P. 2023. Changes to captions: an attentive network for remote sensing change captioning. IEEE Transactions on Image Processing, 32: 6047-6060 [DOI: 10.1109/TIP.2023.3328224]

Chouaf S, Hoxha G, Smara Y and Melgani F. 2021. Captioning changes in bi-temporal remote sensing images//2021 IEEE International Geoscience and Remote Sensing Symposium. Brussels, Belgium: IEEE: 2891-2894 [DOI: 10.1109/IGARSS47720.2021.9554419]

Ferrod R, Di Caro L and Ienco D. 2024. Towards a multimodal framework for remote sensing image change retrieval and captioning//Proceedings of the 27th International Conference on Discovery Science. Pisa, Italy: Springer: 231-245 [DOI: 10.1007/978-3-031-78980-9_15]

Guo Z X, Wang T J and Laaksonen J. 2022. CLIP4IDC: CLIP for image difference captioning//Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing. [s.l.]: Association for Computational Linguistics: 33-42 [DOI: 10.18653/v1/2022.aacl-short.5]

Hosseinzadeh M and Wang Y. 2021. Image change captioning by learning from an auxiliary task//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2724-2733 [DOI: 10.1109/CVPR46437.2021.00275]

Hoxha G, Chouaf S, Melgani F and Smara Y. 2022. Change captioning: a new paradigm for multitemporal remote sensing image analysis. IEEE Transactions on Geoscience and Remote Sensing, 60: #5627414 [DOI: 10.1109/TGRS.2022.3195692]

Jhamtani H and Berg-Kirkpatrick T. 2018. Learning to describe differences between pairs of similar images//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics: 4024-4034 [DOI: 10.18653/v1/D18-1436]

Li Y P, Zhang X R, Gu J, Li C, Wang X, Tang X, et al. 2022. Recurrent attention and semantic gate for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing, 60: #5608816 [DOI: 10.1109/TGRS.2021.3102590]

Liu C Y, Yang J J, Qi Z P, Zou Z X and Shi Z W. 2023. Progressive scale-aware network for remote sensing image change captioning//Proceedings of 2023 IEEE International Geoscience and Remote Sensing Symposium. Pasadena, USA: IEEE: 6668-6671 [DOI: 10.1109/IGARSS52108.2023.10283451]

Liu C Y, Zhao R, Chen H, Zou Z X and Shi Z W. 2022. Remote sensing image change captioning with dual-branch transformers: a new method and a large scale dataset. IEEE Transactions on Geoscience and Remote Sensing, 60: #5633520 [DOI: 10.1109/TGRS.2022.3218921]

Liu F, Chen D L, Guan Z Q Y, Zhou X C, Zhu J L, Ye Q L, et al. 2024. RemoteCLIP: a vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 62: #5622216 [DOI: 10.1109/TGRS.2024.3390838]

Lu X Q, Wang B Q, Zheng X T and Li X L. 2018. 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]

Wang Q, Huang W, Zhang X T and Li X L. 2021. Word-sentence framework for remote sensing image captioning. IEEE Transactions on

Geoscience and Remote Sensing, 59(12): 10532-10543 [DOI: 10.1109/TGRS.2020.3044054]

Park D H, Darrell T and Rohrbach A. 2019. Robust change captioning// Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea(South): IEEE: 4623-4632 [DOI: 10.1109/ICCV.2019.00472]

Qiu Y, Yamamoto S, Nakashima K, Suzuki R, Iwata K, Kataoka H, et al. 2021. Describing and localizing multiple changes with transformers//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 1951-1960 [DOI: 10.1109/ICCV48922.2021.00198]

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

Yao L L, Wang W Y and Jin Q. 2022. Image difference captioning with pre-training and contrastive learning//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 3108-3116 [DOI: 10.1609/aaai.v36i3.20218]

Yu X F, Li Y T, Ma J, Li C and Wu H L. 2025. Diffusion-RSCC: diffusion probabilistic model for change captioning in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 63: #5617013 [DOI: 10.1109/TGRS.2025.3554360]

Yuan X, Cheng G, Li G, Dai W, Yin W X, Feng Y C, et al. 2023. Progress in small object detection for remote sensing images. Journal of Image and Graphics, 28(6): 1662-1684 (袁翔, 程塨, 李戈, 戴威, 尹文昕, 冯瑛超, 等. 2023. 遥感影像小目标检测研究进展. 中国图象图形学报, 28(6): 1662-1684) [DOI: 10.11834/jig.221202]

Zhou K L, Wang P and Cheng J. 2026. Remote sensing image caption generation method based on multi-scale and multi-semantic fusion and collaboration [J/OL]. Journal of Image and Graphics, 1-15 (周凯立, 王鹏, 程剑. 2026. 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, 1-15 [2026-06-21]. https://www.cjig.cn/zh/article/doi/10.11834/jig.250591/ [DOI: 10.11834/jig.250591]

作者简介

于洁潇，女，副教授，主要研究方向为智能信号处理。
E-mail：yjx@tju.edu.cn
刘婧，通信作者，女，副教授，主要研究方向为多媒体信息处理。E-mail：jliu_tju@tju.edu.cn
付雨杰，男，硕士研究生，主要研究方向为遥感多模态学习。
E-mail：fyj_2023@tju.edu.cn