

用于遥感图像变化描述的多尺度特征补偿与变化增强方法

李 浩, 谢明鸿

(昆明理工大学 信息工程与自动化学院, 昆明 650500)

摘要: 遥感图像变化描述 (RSICC) 旨在利用多时相图像生成反映地表变化的自然语言文本, 在国土监测、灾害评估等领域具有重要价值。然而, 在实际应用中, 双时相图像常受光照波动、季节演替等非语义因素干扰, 导致模型易产生“语义幻觉”, 现有方法的普遍采用的简单差分机制难以抑制干扰因素, 且对微弱变化信号的感知力不足, 致使生成的描述文本不够精准。针对上述挑战, 本文提出一种基于多尺度特征补偿与变化增强的遥感图像变化描述框架。首先, 通过构建特征补偿网络 (FCN), 在特征层面强制非变化区域实现双向交互对齐, 显式抑制阴影、光照等伪变化干扰; 其次, 设计变化感知引导模块 (DAGM) 及多尺度融合策略, 在空间与通道维度同步强化判别性特征, 提升模型对复杂地物微小演变的响应灵敏度。在 LEVIR-CC 和 Dubai-CC 基准数据集上的实验结果表明, 本文方法在多项核心评价指标上均优于主流先进算法。

关键词: 遥感图像变化描述; 特征补偿; 多尺度变化增强; 伪变化抑制; 深度学习

中图分类号: TP391.4

文献标志码: A

引 言

随着对地观测技术的飞速发展, 对高分辨率、多时相遥感图像的获取与分析需求日益增长。在海量数据中精准捕捉地表演变, 不仅是遥感领域的核心科学问题, 更是国土监测、灾害评估及城市规划等场景的基础。传统的遥感图像变化检测 (RSICD) 旨在识别地物差异并生成二值掩码, 尽管其检测性能与识别精度已得到显著优化, 如通过多尺度 Transformer 提升判别性、融合 CNN 与 Transformer 增强全局建模、或设计多向交互结构实现特征协同, 但此类任务仅停留在像素级分类, 难以满足用户对变化细节进行深层语义解释的需求。

由此, 遥感图像变化描述 (RSICC) 应运而生, 其核心任务是将视觉感知跨越至自然语言生成。研究者们通过引入空间感知、门控融合及双分支并行结构, 初步实现了从差异建模到高层语义表达的过渡。为应对复杂场景带来的挑战, 诸如渐进式尺度感知、空间-通道联合建模及跨域对比学习等技术被相继引入。近期, RSICC 正迈向大模型驱动与多任务统一化, 通过引入 GPT-2 等语言先验、利用像素级任务提供约束、以及探索扩散模型、指令微调与 Mamba 线性架构, 显著增强了描述的逻辑性与效率。同时, 交互式系统与时序建模的出现, 标志着 RSICC 已由单向静态解译转向智能化、泛化性的主动分析。

尽管上述方法已取得显著进展, 但在实际复杂场景下, 由于缺乏对非语义伪变化的显式抑制机制以及对微弱变化特征的增强手段, 模型易受到光照等噪声干扰而产生语义幻觉, 并因特征感知力不足导致生成的描述语句缺失细粒度地物细节。

基金项目: 国家自然科学基金(62161015)。

收稿日期: 2026-04-24; 修订日期: 2026-05-29

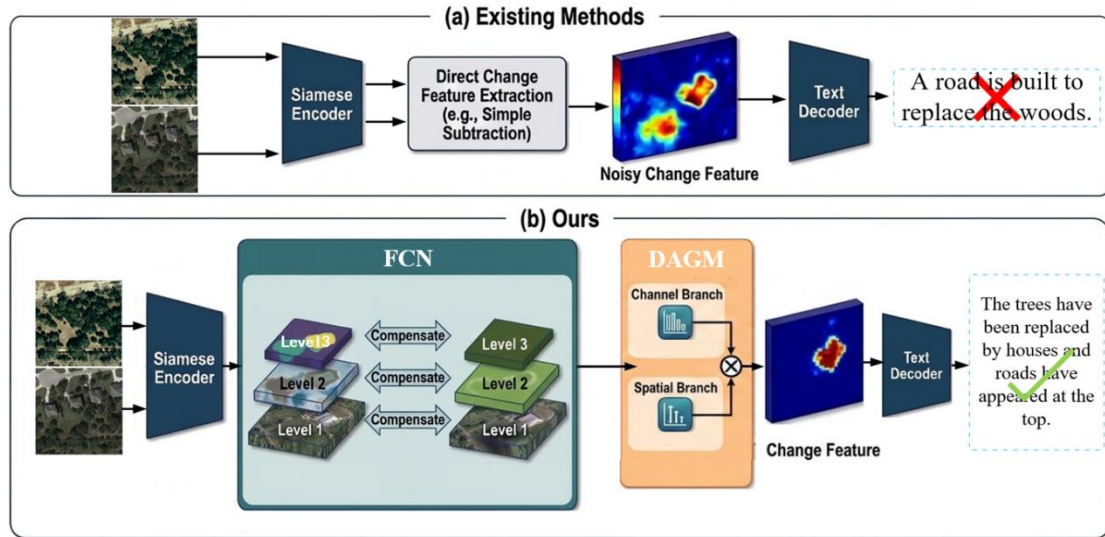


图1 现有 RSICC 方法与本文提出方法的对比示意图。(a) 现有方法, (b) 本文提出的方法

Fig.1 Schematic illustration of the comparison between existing RSICC methods and the proposed method. (a) Existing methods. (b) Our proposed method

如图 1 所示, 现有方法主要采用差分机制等方法直接提取变化信息, 由于缺乏对干扰的抑制能力, 生成的特征图中通畅包含大量伪变化干扰, 导致描述语句偏离事实。

为解决上述问题, 本文的贡献主要包括:

(1) 提出特征补偿网络 (FCN) 通过对多尺度双时相特征进行双向交互调制, 在特征层面使得背景区域保持一致, 实现了对光照、阴影伪变化干扰的显式抑制。

(2) 引入变化感知引导模块 (DAGM) 与多尺度融合策略通过在特征中同步挖掘空间与通道层面的判别信息, 显著强化了模型对真实变化区域的响应强度, 从而解决了模型对变化特征不敏感及描述细节缺失的问题。

(3) 本文在 LEVIR-CC 和 Dubai-CC 数据集上进行了实验, 结果表明该方法在 BLEU、ROUGE-L、METEOR 和 CIDEr 指标上均获得了较优的结果。

1 方法

1.1 总体框架概述

本文提出了一种用于遥感图像变化描述的特征补偿与多尺度变化增强方法。如图 2 所示, 该方法主要由多尺度编码器、双时相特征补偿与差异表示构建、变化感知引导模块以及文本解码器组成。其中, 多尺度编码器用于提取变化前后图像的多层级特征, 双时相特征补偿与差异表示构建通过特征补偿网络 (FCN) 进行双向条件补偿, 从而减轻抑制阴影和光照变化等伪变化干扰, 而后利用变化聚合块 (DAB) 从原始差异和补偿差异两个层面构建并增强变化特征; 变化感知引导模块 (DAGM) 根据变化特征生成注意力图, 以突出真实变化区域; 将处理后的各个特征通过判别性表征构建后, 最终由文本解码器生成对应的变化描述文本。

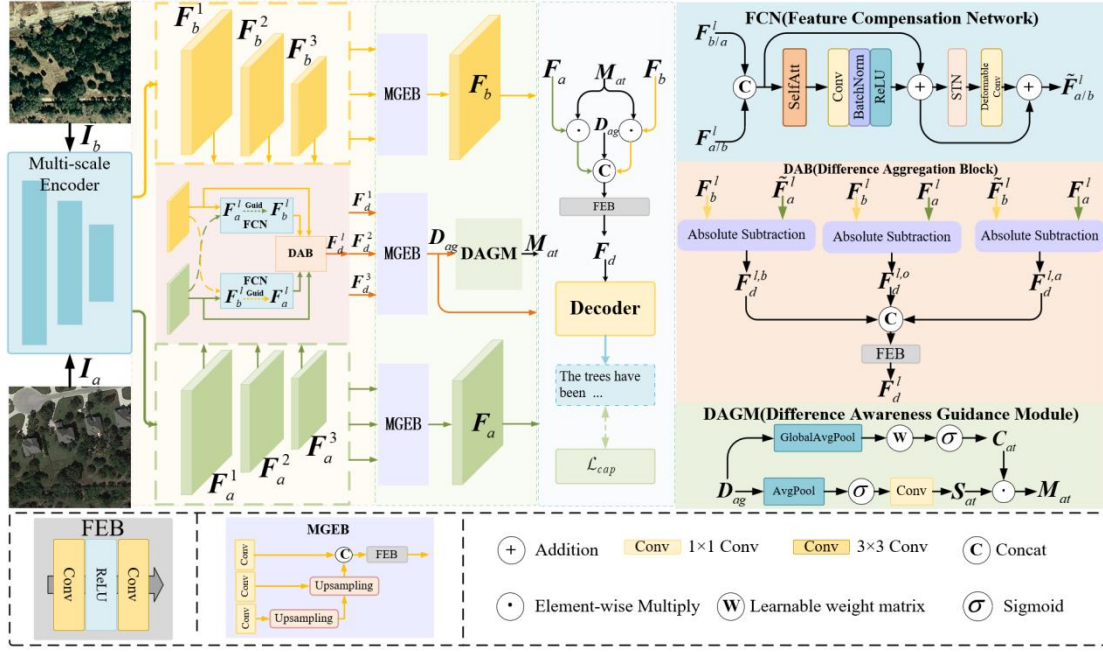


图2 网络总框架

Fig.2 Overall Network Framework

1.2 多尺度编码器

考虑到双时相遥感图像变化描述任务通常对局部纹理、边界结构以及多尺度信息具有较高要求，同时为了避免主干网络过于复杂，本文选用 ResNet-18 作为多尺度编码器（Multi-scale Encoder）。ResNet-18 具有结构简洁、训练稳定、参数量适中以及层级特征表达清晰等优点，能够为后续特征补偿与变化特征建模提供可靠的多尺度特征。

给定变化前图像和变化后图像，记为 $I_b, I_a \in \mathbb{R}^{B \times C \times H \times W}$ ， b 表示变化前图片， a 表示变化后图片， B 表示批大小（batch size）， C 表示输入图像的通道数， H 和 W 分别表示输入图像的高度和宽度。为了保证双时相图像特征处于统一的表示空间，本文采用共享参数的孪生编码方式，即变化前和变化后图像通过同一个编码器 $E(\cdot)$ 进行特征提取，其过程可表示为：

$$F_b^1, F_b^2, F_b^3 = E(I_b) \quad (1)$$

$$F_a^1, F_a^2, F_a^3 = E(I_a) \quad (2)$$

其中， F_b^l 和 F_a^l 分别表示变化前图像和变化后图像在第 l 个尺度下的图像特征， $l \in \{1, 2, 3\}$ 表示三个不同尺度。

具体地，本文从 ResNet-18 的后三个残差阶段提取多尺度特征，即将 Layer2、Layer3 和 Layer4 的输出分别作为三个尺度的编码特征。对于任意时相 $t \in \{b, a\}$ ，其编码过程可进一步写为：

$$X_t^0 = \text{MaxPool}(\phi(\text{BN}(\text{Conv}_{7 \times 7}(I_t)))) \quad (3)$$

$$X_t^1 = \text{Layer1}(X_t^0) \quad (4)$$

$$F_t^1 = \text{Layer2}(X_t^1) \quad (5)$$

$$F_t^2 = \text{Layer3}(F_t^1) \quad (6)$$

$$F_t^3 = \text{Layer4}(F_t^2) \quad (7)$$

其中， $\text{Conv}_{7 \times 7}(\times)$ 表示卷积核大小为 7×7 的卷积操作， $\text{BN}(\cdot)$ 表示批归一化， $\phi(\cdot)$ 表示 ReLU 激活函数， $\text{MaxPool}(\cdot)$ 表示最大池化操作，Layer1 至 Layer4 分别表示 ResNet-18 的四个残差阶段。因此，三个尺度

下的特征张量维度分别为 $\mathbf{F}_b^1, \mathbf{F}_a^1 \in \mathbb{R}^{B \times 128 \times \frac{H}{8} \times \frac{W}{8}}$, $\mathbf{F}_b^2, \mathbf{F}_a^2 \in \mathbb{R}^{B \times 256 \times \frac{H}{16} \times \frac{W}{16}}$, $\mathbf{F}_b^3, \mathbf{F}_a^3 \in \mathbb{R}^{B \times 512 \times \frac{H}{32} \times \frac{W}{32}}$ 。

上述三尺度图像特征具有层次互补性：浅层特征 $\mathbf{F}_t^1$ 保留了较丰富的边缘、纹理和局部细节信息；中层特征 $\mathbf{F}_t^2$ 能够进一步聚合局部结构和上下文关系，用于建模区域级变化特征；深层特征 $\mathbf{F}_t^3$ 则包含更强的高层语义信息。基于多尺度图像特征表示，后续特征补偿网络能够在不同尺度上有针对性地消除伪变化干扰，从而提升双时相变化特征建模的精确性。

1.3 双时相特征补偿与差异表示构建

为减弱光影等伪变化对后续变化特征建模的影响，本文提出特征补偿网络（Feature Compensation Network, FCN），用于在每一个尺度上对双时相图像特征进行双向条件补偿。其核心思想是：利用另一时相图像中包含的伪变化条件对当前时相特征进行调制，使得补偿后的双时相特征在非真实变化因素上尽可能保持一致，从而突出真实变化区域的特征差异。

设多尺度编码器在第 l 个尺度上输出的变化前和变化后图像特征分别为 $\mathbf{F}_b^l, \mathbf{F}_a^l \in \mathbb{R}^{B \times C_l \times H_l \times W_l}$, $l \in \{1, 2, 3\}$ 表示三个不同尺度，考虑到不同尺度特征所承载的信息存在差异，本文在三个尺度上分别采用独立的特征补偿网络，各尺度之间不共享参数。对于任意尺度 l ，使用双向补偿方式分别对变化前和变化后图像特征进行调制。记 $\mathbf{C}^l(\cdot, \cdot)$ 为第 l 个尺度上的特征补偿网络，则其双向补偿过程可表示为：

$$\tilde{\mathbf{F}}_b^l = \mathbf{C}^l(\mathbf{F}_a^l, \mathbf{F}_b^l) \quad (8)$$

$$\tilde{\mathbf{F}}_a^l = \mathbf{C}^l(\mathbf{F}_b^l, \mathbf{F}_a^l) \quad (9)$$

其中， $\tilde{\mathbf{F}}_b^l$ 表示在 $\mathbf{F}_a^l$ 的条件引导下，对 $\mathbf{F}_b^l$ 进行补偿后得到的结果； $\tilde{\mathbf{F}}_a^l$ 表示在 $\mathbf{F}_b^l$ 的条件引导下，对 $\mathbf{F}_a^l$ 进行补偿后得到的结果。

具体地，对于待补偿特征 $\mathbf{F}_x^l$ 及其条件引导特征 $\mathbf{F}_y^l$ ，其中 $x, y \in \{b, a\}$ 且 $x \neq y$ ，首先将二者在通道维进行拼接，得到联合输入特征：

$$\mathbf{Z}_0^l = \mathbf{C}(\mathbf{F}_x^l, \mathbf{F}_y^l) \quad (10)$$

其中， $\mathbf{C}(\cdot)$ 表示通道拼接操作。随后，拼接特征首先进行自注意力操作，以增强双时相特征之间的依赖建模能力，在此基础上，再经过卷积、批归一化和 ReLU 激活进行局部特征提炼与非线性映射。该过程可表示为：

$$\mathbf{Z}_1^l = \mathbf{Z}_0^l + \phi\left(\text{BN}\left(\text{Conv}\left(\text{SelfAtt}\left(\mathbf{Z}_0^l\right)\right)\right)\right) \quad (11)$$

其中， $\text{SelfAtt}(\cdot)$ 表示自注意力操作，用于建模联合特征的全局依赖关系； $\text{Conv}(\cdot)$ 表示 1×1 的卷积操作，用于进一步聚合局部邻域信息并完成通道映射； $\text{BN}(\cdot)$ 表示批归一化操作； $\phi(\cdot)$ 表示 ReLU 激活函数。残差连接用于保留原始双时相图像特征，防止在补偿过程中出现有效结构信息丢失，同时有助于稳定网络训练。

为了进一步减弱由局部错位、形变和阴影形状差异带来的伪变化干扰，本文在上述特征调制之后引入空间变换网络（Spatial Transformer Network, STN）和可变形卷积（Deformable Convolution, DConv）进行逐步补偿。具体过程可写为：

$$\mathbf{Z}_2^l = \text{STN}(\mathbf{Z}_1^l) \mathbf{Z}_1^l = \mathbf{Z}_1^l + \text{DConv}(\mathbf{Z}_2^l; \Delta \mathbf{P}^l) \quad (12)$$

其中， $\text{STN}(\cdot)$ 表示空间变换网络，用于对中间特征进行空间校正； $\text{DConv}(\cdot; \Delta \mathbf{P}^l)$ 表示带有可学习采样偏移的可变形卷积操作， $\Delta \mathbf{P}^l$ 为第 l 个尺度下学习得到的偏移量集合，用于对卷积采样位置进行动态调整，从而更灵活地拟合不规则边界、局部形变以及复杂阴影区域带来的扰动。第二次残差连接用于在空间校正后保留稳定的语义结构信息。因此，第 l 个尺度上的特征补偿网络输出可记为 $\{\tilde{\mathbf{F}}_b^l, \tilde{\mathbf{F}}_a^l\}$ 。

在获得该尺度下补偿后的双时相图像特征后，为了充分提取强化该尺度双时相图像之间的变化信息，本文进一步设计变化聚合块(Difference Aggregation Block, DAB)，在每个尺度内进行差异表示构建。需要说

明的是，该模块是针对同一尺度下的双时相图像特征，从原始变化和双向补偿变化两个层面挖掘更加充分的变化表示。具体而言，对于第 l 个尺度，首先构建原始双时相图像特征之间的变化特征，即：

$$\mathbf{F}_d^{l,o} = \left| \mathbf{F}_b^l - \mathbf{F}_a^l \right| \quad (13)$$

其中， $\mathbf{F}_d^{l,o}$ 表示由原始变化前图像特征 $\mathbf{F}_b^l$ 和原始变化后图像特征 $\mathbf{F}_a^l$ 进行绝对值减法构造得到的变化特征。随后，以同样的方式，利用特征补偿网络得到的双向补偿结果，分别从两个方向进一步构建变化特征。可以得到在变化前图像为条件引导下构造的变化特征 $\mathbf{F}_d^{l,b}$ ，以及在变化后图像为条件引导下构造的变化特征 $\mathbf{F}_d^{l,a}$ 。

在通过三种方式得到的变化特征后，将它们在通道维上进行拼接，得到同一尺度下的联合变化表示：

$$\mathbf{F}_c^l = C(\mathbf{F}_d^{l,b}, \mathbf{F}_d^{l,o}, \mathbf{F}_d^{l,a}) \quad (14)$$

其中， $C(\cdot)$ 表示通道拼接操作。随后，将拼接后的联合变化特征送入特征强化块 (Feature Enhancement Block, FEB) 进行增强，得到该尺度最终的变化特征，本文所采用的 FEB 由两个 3×3 卷积层和一个 ReLU 激活函数构成，其具体形式为：

$$\mathbf{F}_d^l = \text{Conv}_{3 \times 3}^{(2)} \left(\phi \left(\text{Conv}_{3 \times 3}^{(1)} (\mathbf{F}_c^l) \right) \right) \quad (15)$$

其中， $\mathbf{F}_d^l$ 表示第 l 个尺度下经过变化提取与融合后的最终变化特征， $\text{Conv}_{3 \times 3}^{(1)}(\cdot)$ 和 $\text{Conv}_{3 \times 3}^{(2)}(\cdot)$ 分别表示第一个和第二个卷积操作， $\phi(\cdot)$ 表示 ReLU 激活函数。

综上，变化聚合块在每个尺度内分别从较完整保留原始地物边界和轮廓的原始双时相变化、以及消除由伪变化带来干扰的补偿变化三种方式构建变化特征，并通过特征强化块对三者进行融合增强，从而得到单尺度变化特征。

1.4 变化感知引导模块

为增强变化前图像特征和变化后图像特征中的关键变化，本文进一步设计变化感知引导模块 (Difference Awareness Guidance Module, DAGM)，用于从变化特征中挖掘对变化区域更敏感的信息，并生成对应的注意力图。具体而言，对于来自每个尺度 DAB 输出的变化特征，首先通过多尺度聚合增强块 (Multiscale Aggregation Enhancement Block, MGEB) 将多尺度特征融合增强为单个统一的特征表示。MGEB 对来自 DAB 输出的特征处理过程可以表示为：

$$\mathbf{D}'_{ag} = C \left(U \left(U \left(\text{Conv}_{1 \times 1} (\mathbf{F}_d^3) \right) \right), U \left(\text{Conv}_{1 \times 1} (\mathbf{F}_d^2) \right), \mathbf{F}_d^1 \right) \quad (16)$$

其中， $\text{Conv}_{1 \times 1}(\cdot)$ 表示 1×1 卷积操作， $U(\cdot)$ 表示上采样操作。随后，将联合特征 $\mathbf{D}'_{ag}$ 经过 FEB 进行增强得到最终的变化特征表示 $\mathbf{D}_{ag}$ 。DAGM 以聚合后的变化特征 $\mathbf{D}_{ag}$ 作为输入，同时从通道和空间两个角度对其进行建模，得到变化注意力图 $\mathbf{M}_{at}$ 。

对于通道分支，首先对 $\mathbf{D}_{ag}$ 进行全局平均池化，以提取其全局通道统计信息，随后通过线性映射和 Sigmoid 激活函数生成通道注意表示，其过程可表示为：

$$\mathbf{C}_{at} = \sigma \left(\mathbf{W} \left(\text{GAP}(\mathbf{D}_{ag}) \right) \right) \quad (17)$$

其中， $\text{GAP}(\cdot)$ 表示全局平均池化操作， $\mathbf{W}(\cdot)$ 表示可学习的权重矩阵， $\sigma(\cdot)$ 表示 Sigmoid 激活函数， $\mathbf{C}_{at}$ 表示由变化特征生成的通道注意表示。

对于空间分支，先对 $\mathbf{D}_{ag}$ 进行平均池化，以压缩冗余信息并保留显著的空间响应，再经过 Sigmoid 激活和卷积操作生成空间注意表示，即：

$$\mathbf{S}_{at} = \text{Conv} \left(\sigma \left(\text{AP}(\mathbf{D}_{ag}) \right) \right) \quad (18)$$

其中， $\text{AP}(\cdot)$ 表示平均池化操作， $\text{Conv}(\cdot)$ 表示 1×1 卷积操作， $\mathbf{S}_{at}$ 为空间注意表示。在得到通道注意表示和空间注意表示后，通过逐元素乘法将二者进行融合，得到最终的变化注意力图：

$$\mathbf{M}_{at} = \mathbf{C}_{at} \square \mathbf{S}_{at} \quad (19)$$

其中, $\square$ 表示逐元素乘法, $\mathbf{M}_{at}$ 表示最终生成的变化注意力图。

通过上述过程, 变化感知引导模块能够从聚合后的变化特征 $\mathbf{D}_{ag}$ 中提取变化相关的通道与空间响应信息, 并生成对应的注意力图 $\mathbf{M}_{at}$, 为后续特征增强提供引导。

1.5 判别性表征构建与文本生成

为融合多个特征中的变化部分, 在送入解码器之前, 本文对由 MEGB 处理得到的统一尺度的双时相图像特征 $\mathbf{F}_b$ 与 $\mathbf{F}_a$ 进行调制, 并与聚合变化特征 $\mathbf{D}_{ag}$ 共同构成最终的变化特征表示。具体来说, 首先利用注意力图对变化前图像特征 $\mathbf{F}_b$ 和变化后图像特征 $\mathbf{F}_a$ 进行逐元素加权, 以突出变化区域的响应, 可表示为:

$$\hat{\mathbf{F}}_b = \mathbf{F}_b \square \mathbf{M}_{at} \quad (20)$$

$$\hat{\mathbf{F}}_a = \mathbf{F}_a \square \mathbf{M}_{at} \quad (21)$$

其中, $\square$ 表示逐元素乘法, $\hat{\mathbf{F}}_b$ 和 $\hat{\mathbf{F}}_a$ 分别表示经过注意力调制后的变化前图像特征和变化后图像特征。随后, 将调制后的双时相图像特征与聚合变化特征 $\mathbf{D}_{ag}$ 在通道维进行拼接, 并通过 FEB 进一步融合, 得到最终的变化输出特征:

$$\mathbf{F}_d = \text{FEB}\left(C\left(\hat{\mathbf{F}}_b, \mathbf{D}_{ag}, \hat{\mathbf{F}}_a\right)\right) \quad (22)$$

其中, $C(\cdot)$ 表示通道拼接操作, $\mathbf{F}_d$ 表示将要输入到文本解码器的最终变化特征。

在变化描述生成阶段, 最终的变化特征 $\mathbf{F}_d$ 送入文本解码器后, 解码器以自回归方式生成描述变化内容的自然语言。为约束生成描述与真实文本标签之间的一致性, 本文采用交叉熵损失对整个生成过程进行监督, 定义为:

$$\ell_{cap} = -\sum_{t=1}^L \sum_{v=1}^V \tilde{y}_t^{(v)} \log p_t^{(v)} \quad (23)$$

其中, $p_t = [p_t^{(1)}, \dots, p_t^{(V)}]$ 表示模型在第 t 个时间步对词汇表中各词的预测概率向量, $\tilde{y}_t$ 表示第 t 个时间步参考词对应的独热向量, V 表示词汇表大小, L 表示生成句子的长度。

这一过程使模型不断优化, 其生成的文本在语义上逐步逼近真实标注。

2 实验

通过在公开数据集上的对比实验、定性分析、消融实验以及复杂度分析, 对本文方法的有效性进行全面验证。

2.1 数据集

LEVIR-CC 错误!不能识别的开关参数。 是遥感变化描述任务中常用的基准数据集之一。该数据集由美国德克萨斯州 20 个区域的双时相遥感图像构成, 共包含 10,077 组图像对, 其中变化样本 5,038 组, 无变化样本 5,039 组。变化类型涵盖建筑物、道路、停车场、植被和水体等五类典型地物。每组图像配备 5 条人工标注的文本描述, 因此共包含 50,385 条句子标注。数据集中句子的最大长度为 39 个单词, 平均长度为 7.99 个单词。按照标准划分方式, 训练集、验证集和测试集分别包含 6,815 组、1,333 组和 1,929 组样本。

Dubai-CC 错误!不能识别的开关参数。 面向城市区域长期演化场景下的变化描述任务构建, 数据来源于迪拜地区 2000 年至 2010 年间的双时相遥感图像。该数据集总计包含 500 组图像对, 标注类别涉及道路、居民区、建筑物、绿地、湖泊和岛屿六类变化。每组图像同样对应 5 条变化描述, 共计 2,500 条标注文本。句子最大长度为 23 个单词, 平均长度为 7.35 个单词, 词汇表规模为 246。训练集、验证集与测试集分别包含 300 组、50 组和 150 组样本。

2.2 评价指标

为便于与已有方法公平比较，本文采用 BLEU-4 错误:不能识别的开关参数。、ROUGE-L 错误:不能识别的开关参数。、METEOR 错误:不能识别的开关参数。和 CIDEr-D 错误:不能识别的开关参数。四项指标评价生成文本的质量。其中，BLEU-4 从 n-gram 匹配角度衡量预测语句与参考语句之间的一致程度；ROUGE-L 基于最长公共子序列考察生成结果对关键信息的覆盖能力；METEOR 在匹配机制中进一步考虑词形变化与惩罚项，因此更能反映文本的流畅性与可读性；CIDEr-D 则针对图像描述任务设计，通过 TF-IDF 加权方式评价生成语句与参考文本在语义层面的相似性。上述指标均满足“数值越大，生成效果越好”的评价准则。

2.3 实现细节

本文方法基于 PyTorch 实现，所有实验均在单张 NVIDIA GeForce RTX 4090 GPU 上完成。训练过程中采用 Adam 优化器，初始学习率设置为 1×10^{-4} ，并结合余弦退火策略动态更新学习率。批大小设为 16，训练轮数设为 50。为提高模型泛化能力，在优化过程中引入权重衰减，系数设置为 1×10^{-2} ；同时采用早停策略，当验证集性能连续 10 轮不再提升时终止训练。在网络结构方面，本文采用 ResNet-18 作为共享参数的多尺度编码器，从双时相图像中提取三个层级的特征表示。针对不同尺度，分别使用独立权重的特征补偿网络，以实现双向条件补偿并削弱伪变化干扰。所有输入图像尺寸均为 256×256 ，文本最大生成长度设为 40。

2.4 定量分析

为验证本文方法的优势，本文将其与多种代表性变化描述方法进行比较，包括采用了动态扬声器来替代传统的 LSTM 进行语言解码的方法 DUDA 错误:不能识别的开关参数。、通过自注意力同时建模图像内和图像间关联和利用协同注意力机制来捕捉双时相图像之间的复杂交互的方法 MCCFormer 错误:不能识别的开关参数。、基于概率扩散的模型，通过跨模态融合与堆叠自注意来强化跨模态对齐并降低像素干扰的方法 Diffusion-RSCC 错误:不能识别的开关参数。、利用双分支 Transformer 编码器架构和注意力机制的方法 RSICCformer 错误:不能识别的开关参数。、基于大模型指令微调的方法 PromptCC 错误:不能识别的开关参数。、使用单流编码器架构和双时相对比预训练策略的方法 SEN 错误:不能识别的开关参数。以及从视频建模视角重构变化检测与描述任务方法 Change3D 错误:不能识别的开关参数。，RsCaMa 错误:不能识别的开关参数。，实验结果分别列于表 1 和表 2。其中，最优结果采用加粗标记。

表 1 不同方法在 LEVIR-CC 数据集上的性能对比。加粗表示最优结果

Tab.1 Performance comparison of different methods on the LEVIR-CC dataset. Bold values indicate the best results

Method	BLEU-4	METEOR	ROUGE-L	CIDEr-D
DUDA	57.79	37.15	71.04	124.32
MCCFormer-S	56.36	39.60	69.46	120.39
MCCFormer-D	56.38	39.91	70.44	124.44
Diffusion-RSCC	60.90	37.80	71.50	125.60
RSICCformer	62.77	39.61	74.12	134.12
PromptCC	63.54	38.82	73.72	136.44
SEN	64.09	39.59	74.57	136.02
Change3D	64.38	40.03	75.12	138.29
Ours	65.72	40.20	76.91	140.13

由表 1 可见，本文方法在 LEVIR-CC 数据集的四项评价指标上均取得最佳结果。相较于性能较强的 Change3D，本文方法在 BLEU-4、METEOR、ROUGE-L 和 CIDEr-D 上分别提升了 1.34、0.17、1.79 和 1.84 个百分点。这表明，所提方法不仅能够更加准确地捕捉场景中的关键变化，还能够生成内容更完整、表达更自然的描述文本。尤其是在 ROUGE-L 与 CIDEr-D 两项指标上的提升更为明显，说明本文提出的模型对真实变化内容的建模更加充分，生成语句与参考描述之间的语义一致性也更高。

表 2 不同方法在 Dubai-CC 数据集上的性能对比。加粗表示最优结果

Tab.2 Performance comparison of different methods on the Dubai-CC dataset. Bold values indicate the best **results**

Method	BLEU-4	METEOR	ROUGE-L	CIDEr-D
DUDA	25.39	22.05	48.34	62.78
MCCFormer-S	22.57	18.64	43.29	53.81
MCCFormer-D	29.48	25.09	51.27	66.51
RSICCFomer	31.28	25.41	51.96	66.54
PromptCC	40.32	26.48	55.82	85.44
SEN	31.01	23.67	48.19	65.15
Change3D	36.80	27.06	56.04	86.19
Ours	42.34	28.63	57.69	88.83

在 Dubai-CC 数据集上的实验结果同样验证了本文方法的有效性。如表 2 所示，本文方法在所有指标上均优于对比方法。与当前较优的 PromptCC 和 Change3D 相比，本文方法在 BLEU-4、METEOR、ROUGE-L 和 CIDEr-D 上均保持稳定领先。由于 Dubai-CC 数据量更小、场景构成更复杂，因此这组结果进一步说明，本文方法在数据分布变化较大时仍具有较好的鲁棒性和泛化能力。其主要原因在于：特征补偿网络能够在特征层面缓解阴影和亮度等干扰因素带来的伪变化影响，而后续变化感知引导增强机制则有助于进一步聚焦真实变化区域，从而提升描述质量。

2.5 定性分析

为了更直观地观察本文方法的描述能力，本文分别在 LEVIR-CC 和 Dubai-CC 数据集上选取若干具有代表性的样本进行对比。



GT: A small track appears.
 RSICCFomer: The scene is the same as before.
 MCCFormer-S: There is no difference.
 MCCFormer-D: The scene is the same as before.
 Our: A road has been built on the bare land.



GT: Some trees have disappeared and some roads are built.
 RSICCFomer: A road is built and some houses are built along the road.
 MCCFormer-S: A road is built on the left and a building is built at the bottom.
 MCCFormer-D: A road appears in the bare land.
 Our: Some trees have disappeared and some roads are built.



GT: Part of the woodland becomes the road and the houses on both sides of the road.
 RSICCFomer: The road with houses built along appears in the desert.
 MCCFormer-S: A road with many houses on both sides of the road.
 MCCFormer-D: a road with many houses around appears in the desert.
 Our: A road with houses on both sides replaces some trees.



GT: The road at the top has been rebuilt and four houses appear replacing some trees.
 RSICCFomer: Houses and road have been built.
 MCCFormer-S: A road is built to replace the woods.
 MCCFormer-D: Trees are removed and houses appear.
 Our: Four houses appear replacing some trees and the road at the top has been rebuilt.

图 3 不同方法在 LEVIR-CC 数据集上生成的描述文本对比。其中绿色文本标注了与变化相关的关键短语

Fig.3 Comparison of generated captioning results by different methods on the LEVIR-CC dataset. Key phrases related to changes are highlighted in green

图 3 展示了 LEVIR-CC 数据集上的部分结果。可以发现，当面临如左上角样本所示的裸地新增道路场景时，对比方法由于受地表纹理与光照等伪变化因素干扰，普遍出现了未检测到变化的误报现象，而本文方法能够对干扰进行显式抑制，从而精准地识别出了新增地物的类别。在处理包含多类地物同时演变的复杂场景时，本文提出的方法能够捕捉到更精细的变化特征，不仅准确描述了房屋的出现与树木的消失，还进一步刻画了地物间的替代关系及道路重建等细节，生成的描述在逻辑连贯性与语义准确度上均最接近标签。



图 4 不同方法在 Dubai-CC 数据集上生成的描述文本对比。其中绿色文本标注了与变化相关的关键短语

Fig.4 Comparison of generated captioning results by different methods on the Dubai-CC dataset. Key phrases related to changes are highlighted in green

图 4 展示了 Dubai-CC 数据集上的定性结果。数据集中包含大量的沙漠背景，地表纹理的细微波动极易对模型造成伪变化干扰，如在处理新增绿地或垂直道路等细长目标时，本文提出的特征补偿与多尺度变化增强方法展现了更强的描述准确性，特征补偿网络通过对双时相特征的补偿调制，有效消除了由光照等引起的伪变化噪声。在描述复杂空间布局时，本文方法生成的文本不仅地物类别识别准确，且能够清晰表达出地物间的位置关系，整体逻辑连贯性最接近标签。

为验证 FCN 补偿机制的有效性，本文在 LEVIR-CC 数据集和 Dubai-CC 数据集进行了可视化分析，为了更直观清晰地展示特征补偿效果，我们可视化了包含更多细节的浅层特征对比图，如图 5 所示。

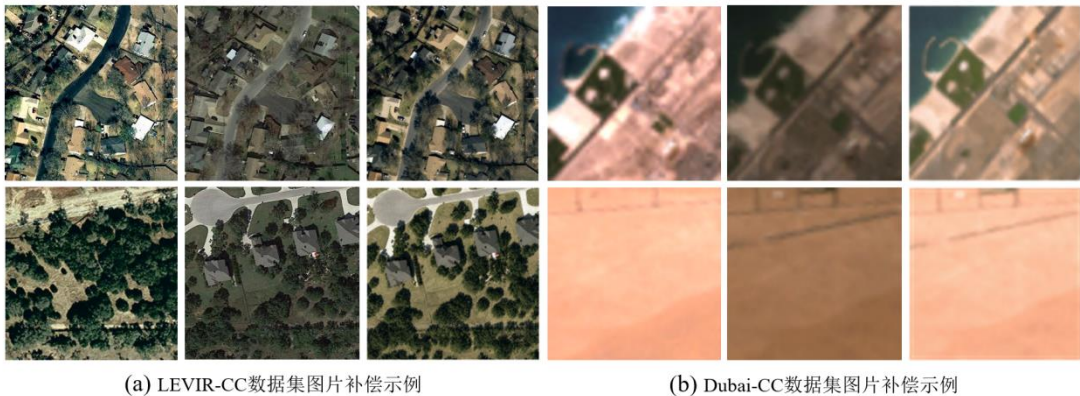


图5 FCN补偿可视化结果图

Fig5. Visualization results of FCN compensation

其中, (a) 为 LEVIR-CC 数据集图片补偿前后示例, 左侧为引导图片, 中间为待补偿图片, 右侧为补偿结果, 可以观察到, 待补偿图片补偿后的结果与引导图片在光照等伪变化因素上趋于一致, (b)为 Dubai-CC 数据集图片的补偿前后示例, 可以观察到与 LEVIR-CC 数据集补偿示例相似的结果。这直观且有效地证明了 FCN 模块在光照等伪变化因素的补偿能力。

2.6 消融实验

为进一步验证本文设计的有效性, 本文在 LEVIR-CC 和 Dubai-CC 两个数据集上开展消融实验。考虑到 FCN 和 DAGM 分别对应本文方法中两个最核心的环节, 相比之下, 其余模块主要承担差异构建、跨尺度聚合与输出整合等功能, 并与整体信息流耦合较为紧密。若对这些模块进行单独拆分, 往往会同时改变网络的特征传递路径与输入输出形式, 从而削弱消融对比的公平性。因此, 本文选择对 FCN 和 DAGM 进行独立模块消融实验, 以更清晰地评估两者对整体性能的贡献。

表 3 LEVIR-CC 数据集上的消融实验结果

Tab.3 Ablation study results on the LEVIR-CC dataset

FCN	DAGM	BLEU-4	METEOR	ROUGE-L	CIDEr-D
×	×	64.06	38.15	74.24	135.02
√	×	65.04	39.57	75.91	139.67
×	√	65.31	39.44	75.84	139.71
√	√	65.72	40.20	76.91	140.13

表 3 展示了 LEVIR-CC 数据集上的消融结果。当 FCN 和 DAGM 均被移除时, 模型仍能完成基本的变化描述任务, 但整体性能仍存在进一步提升空间。加入 FCN 后, 四项指标均有不同程度改善, 说明双向条件补偿能够有效减弱阴影、光照变化及局部错位等伪变化干扰, 从而为后续变化建模提供更加稳定的特征基础。仅引入 DAGM 时, 模型性能同样有所提升, 表明基于聚合变化特征生成的注意力图可以有效突出与真实变化相关的关键区域, 增强最终特征表达的针对性。进一步地, 当 FCN 与 DAGM 同时引入时, 模型在各项指标上均达到最优, 说明前端伪变化抑制与后端差异区域强化之间具有良好的互补关系, 二者协同能够进一步提升变化描述性能。

表 4 Dubai-CC 数据集上的消融实验结果

Table 4. Ablation study results on the Dubai-CC dataset

FCN	DAGM	BLEU-4	METEOR	ROUGE-L	CIDEr-D
×	×	38.93	26.14	55.62	86.90
√	×	40.68	26.86	57.20	87.86
×	√	41.51	27.66	57.11	87.65
√	√	42.34	28.63	57.69	88.83

表 4 给出了 Dubai-CC 数据集上的消融结果, 可以观察到与 LEVIR-CC 上相似的变化趋势。无论是单独加入 FCN 还是单独加入 DAGM, 模型性能均较基线有所提升, 说明这两个模块在复杂场景下同样能够有效发挥作用。其中, FCN 的增益更加明显, 这表明在样本规模较小且场景干扰更复杂的 Dubai-CC 数据集上, 特征层面的伪变化抑制对于提升变化描述质量尤为重要。与此同时, DAGM 也带来了稳定增益, 说明在聚合特征基础上进一步强化变化敏感区域, 有助于模型更准确地捕获关键变化语义。最终, 当两者同时引入时, 模型取得最佳性能, 进一步说明本文方法从抑制伪变化到突出真实变化的整体设计是有效且合理的。

综合两个数据集的实验结果可以看出, FCN 和 DAGM 分别从不同方面提升了模型对真实变化信息的建模能力。前者着重缓解双时相图像之间由成像条件差异带来的干扰, 后者则进一步强化变化区域在最终表征中的响应强度。二者相互配合, 共同促进了变化描述性能的提升。

为了探究 DAB 内部三种提取变化特征的方式对模型整体的贡献，在其余模块都兼备的情况下，考虑到大型数据集具备更丰富的样本多样性与更复杂的遥感场景，能够提供更具说服力的评测基准，本文选择在相对更大规模的 LEVIR-CC 数据集上对 DAB 的三种提取变化特征的方式进行消融实验分析，由原始变化前图像特征和原始变化后图像特征进行绝对值减法构造得到的变化特征我们记为“原始变化”，由原始变化前图像特征和原始变化前图像特征引导补偿得到的变化后图像特征进行绝对值减法构造得到的变化特征我们记为“变化前引导”，由原始变化后图像特征和原始变化后图像特征引导补偿得到的变化前图像特征进行绝对值减法构造得到的变化特征我们记为“变化后引导”，鉴于 DAB 对上述三种方式进行了差异提取，所以我们设计了 4 组对照实验，如表 5 所示。

表 5 DAB 模块三种方式变化提取在 LEVIR-CC 数据集的消融实验结果对比。加粗表示最优结果

Table 5. Comparison of ablation study results for the three change extraction approaches of the DAB module on the LEVIR-CC dataset.

Bold indicates the best results

原始变化	变化前引导	变化后引导	BLEU-4	METEOR	ROUGE-L	CIDEr-D
✓			65.23	39.70	76.03	139.69
✓	✓		65.61	39.95	76.79	140.05
✓		✓	65.55	39.88	76.88	139.97
✓	✓	✓	65.72	40.20	76.91	140.13

表 5 展示了 DAB 内部三种方式进行变化提取的消融实验结果，仅保留原始变化分支时，仅通过原始双时相特征图显式做差构造变化特征，易受到光照阴影等伪变化的干扰，因而性能受限。而在引入变化前引导或变化后引导后，特征补偿网络的条件补偿机制利用另一时相图像中包含的伪变化条件对当前时相特征进行调制使该方向上补偿，抑制了伪变化干扰，模型的各项核心指标均有提升，最终，当三种方式构造的变化特征共同参与时，通过协同融合的方式，不仅保留了原始的结构信息，更最大化地强化了双时相特征之间的依赖建模能力，获得更具判别力和上下文信息的视觉表征，模型取得了最优结果，这也体现出，由于 FCN 双向补偿分支并存所建立的双向交叉校验与冗余容错机制，使双时相特征能够互为约束、协同纠错，一定程度地规避了单向调制带来的潜在失效风险，提升了模型的整体鲁棒性，所以当完全引入双向补偿机制后，模型性能达到最优结果。

2.7 模型复杂度分析

除了精度指标外，本文还对模型的参数量、计算量以及推理速度进行了分析，并将其与已有方法进行比较。

表 5 不同变化描述方法的复杂度对比。输入图像尺寸统一为 256×256，推理时间单位为毫秒/样本

Tab.5 Complexity comparison of different change captioning methods. The input image size is unified to 256×256, and the inference time is measured in milliseconds per sample (ms/sample)

Methods	Params(M)	FLOPs(G)	Inference(ms/sample)
DUDA	52.78	20.28	14.0
MCCFormer-S	135.01	25.09	15.0
MCCFormer-D	135.01	25.09	15.0
RSICCformer	145.26	27.10	10.0
PromptCC	196.28	19.88	13.0
SEN	39.90	10.09	8.0
Change3D	5.05	2.39	7.0
Ours	71.84	16.72	11.0

由表 5 可知，本文方法在取得较优性能的同时，整体复杂度仍处于较为合理的范围内。得益于以

ResNet-18 为基础的编码结构以及轻量化的差异建模模块, 本文方法的参数量和 FLOPs 均低于多种基于 Transformer 骨干的对比模型, 并且推理速度保持在较高水平。虽然与极轻量模型相比, 本文方法在复杂度上并非最低, 但结合前述实验结果可以看出, 其在性能与效率之间实现了较好的平衡, 具备实际应用潜力。

3 结束语

本文聚焦于遥感图像变化描述任务, 旨在从双时相图像中生成描述变化的准确、自然的语言描述。为克服现有方法在复杂成像条件下易受光照波动、阴影等伪变化干扰, 以及对真实细微变化特征感知能力不足的问题, 我们提出了一种基于多尺度特征补偿与变化增强的遥感图像变化描述方法。该方法的核心在于构建了一个层次化的特征处理框架, 通过特征补偿网络 (FCN) 实现非变化区域的双向交互对齐与显式干扰抑制, 而后利用补偿特征构建多维差异表征, 并利用变化感知引导模块 (DAGM) 深度挖掘并强化了真实变化区域的细节。在 LEVIR-CC 和 Dubai-CC 两个公开数据集上的系统实验表明, 本文所提方法在多个关键指标上均达到了优于主流先进算法的性能, 验证了其有效性。

参 考 文 献:

- [1] TOTH C and JOZKOW G, Remote sensing platforms and sensors: A survey[J].ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 115:22-36.
- [2] ZHANG Z and ZHU L, A review on unmanned aerial vehicle remote sensing: Platforms, sensors, data processing methods, and applications[J]. Drones, 2023, 7(6):398.
- [3] ZHU L, SUOMALAINEN J, LIU J, et al. A review: Remote sensing sensors[M]//RUSTAMOV J B, SABITOV S E. Multi-purposeful Application of Geospatial Data. London, UK: IntechOpen, 2018,19:19-42.
- [4] CHEN B, LIU L,LIU C, et al. Spectral-cascaded diffusion model for remote sensing image spectral super-resolution[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62:1-14.
- [5] 史振威,张浩田,郭涵,等.基于深度学习的双时相光学遥感图像变化检测方法综述[J].数据采集与处理,2026,41(2):566-591.
SHI Zhenwei,ZHANG Haotian,GUO Han,et al.A Review of Deep Learning-Based Change Detection Methods for Bi-temporal Optical Remote Sensing Images[J].Journal of Data Acquisition and Processing,2026,41(2):566-591.
- [6] 朱永进,李熹,李永胜,等. MS-SwinCE:融合多尺度Swin Transformer与对比特征增强的遥感图像变化检测[J]. 计算机系统应用, 2025, 34 (12): 67-74.
ZHU Yongjin, LI Xi, LI Yongsheng, et al. MS-SwinCE: Remote sensing image change detection integrating multi-scale Swin Transformer and contrastive feature enhancement [J]. Computer Systems Applications, 2025, 34 (12): 67-74.
- [7] 武一, 贡世林. C-T Net:融合CNN和Transformer的遥感图像变化检测模型[J]. 华东师范大学学报(自然科学版), 2025, (4): 49-60.
WU Yi, YUN Shilin. C-T Net: Remote sensing image change detection model integrating CNN and Transformer [J]. Journal of East China Normal University (Natural Science), 2025, (4): 49-60.
- [8] 王雷全,童寿梁,耿辰东. 用于遥感图像变化检测的交互式多编码器和多解码器网络设计[J]. 实验技术与管理, 2025, 42 (4): 48-58.
WANG LeiQuan, TONG Shouliang, GENG Chendong. Design of interactive multi-encoder and multi-decoder network for remote sensing image change detection [J]. Experimental Technology and Management, 2025, 42 (4): 48-58.
- [9] QIN Z, ZHANG L. Dsqnet: Enhancing change detection network via deep semantics query for remote sensing images[J]. IEEE Geoscience and Remote Sensing Letters, 2024, 21:1-5.
- [10] WANG H, LIU J, ZHANG W, et al. Cross-temporal attention and adaptive ensemble kernel learning for change detection in remote sensing images[C]//IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2025:8124-8128.
- [11] 卓力,于婉婷,贾童瑶,等. 基于Transformer的遥感图像变化检测研究进展[J]. 北京工业大学学报, 2025, 51 (7): 851-866.
ZHUO Li, YU Wanting, JIA Tongyao, et al. Research progress of remote sensing image change detection based on Transformer [J]. Journal of Beijing University of Technology, 2025, 51 (7): 851-866.
- [12] 李新宇. 基于深度学习的高分遥感图像建筑物变化检测研究[D]. 大连交通大学, 2025.
LI Xinyu. Research on building change detection of high-resolution remote sensing images based on deep learning [D]. Dalian Jiaotong University, 2025.
- [13] HOXHA G, CHOUAF S, MELGANI F, et al. Change captioning: A new paradigm for multitemporal remote sensing image analysis[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60:1-14.

- [14] PARK D H, DARRELL T, ROHRBACH A. Robust change captioning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:4624-4633.
- [15] QIU Y, YAMAMOTO S, NAKASHIMA K, et al. Describing and localizing multiple changes with transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:1971-1980.
- [16] LIU C, ZHAO R, CHEN H, et al. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60:1-20.
- [17] CHANG S, GHAMISI P. Changes to captions: An attentive network for remote sensing change captioning[J]. IEEE Transactions on Image Processing, 2023, 32:6047-6060.
- [18] CHEN C, WANG Y, YAP K H. Multi-scale attentive fusion network for remote sensing image change captioning[C]//2024 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2024:1-5.
- [19] LIU C, YANG J, QI Z, et al. Progressive scale-aware network for remote sensing image change captioning[C]//IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2023:6668-6671.
- [20] XU K, HAN Y, YANG R, et al. Mfnet: A new multi-scale feature refining method for remote sensing image change captioning[C]//IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2024:7119-7123.
- [21] XIAN T, ZHOU Z, ZENG D, et al. Dynamic and asymmetric enhancement for remote sensing image change captioning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63:1-15.
- [22] SUN D, WANG Y, YAO J, et al. Scnet: Lightweight spatial-channel attention network for remote sensing change captioning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2026, 64:1-13.
- [23] MENG X, ZHAO Y, CAO K, et al. Dualfocus-capnet: A dual-stream network for real change and interscale relationship-aware change captioning in remote sensing images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2025, 19:2045-2059.
- [24] ZHOU Q, GAO J, YUAN Y, et al. Single-stream extractor network with contrastive pre-training for remote-sensing change captioning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62:1-14.
- [25] KARACA A C, OZELBAS E, BERBER S, et al. Robust change captioning in remote sensing: Second-cc dataset and mmodalcc framework[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2025, 18:21494-21513.
- [26] LIU C, ZHAO R, CHEN J, et al. A decoupling paradigm with prompt learning for remote sensing image change captioning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61:1-18.
- [27] ZHU Y, LI L, CHEN K, et al. Semantic-cc: Boosting remote sensing image change captioning via foundational knowledge and semantic guidance[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62:1-16.
- [28] LIU C, CHEN K, QI Z, et al. Pixel-level change detection pseudo-label learning for remote sensing change captioning[C]//IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2024:8405-8408.
- [29] YANG C, LI Z, JIAO H, et al. Enhancing perception of key changes in remote sensing image change captioning[J]. IEEE Transactions on Image Processing, 2025, 34:7378-7390.
- [30] YVSHNAV D, GUTHA L, MANINDRA A P V, et al. Intelli-change remote sensing-a novel transformer approach[C]//2024 Second International Conference on Data Science and Information System (ICDSIS). IEEE, 2024:1-7.
- [31] KUMAR S, REDDY P L, VAJPAYEE K, et al. H-ldm: A hierarchical llm-driven framework for remote sensing change detection and domain assistance[C]//2025 IEEE 2nd International Conference for Women in Computing (InCoWoCo). IEEE, 2025:1-6.
- [32] YU X, MA J, QIAO L. Kd-rsc: A karras diffusion framework for efficient remote sensing change captioning[J]. IEEE Geoscience and Remote Sensing Letters, 2025, 22:1-5.
- [33] YU X, LI Y, MA J, et al. Diffusion-rsc: Diffusion probabilistic model for change captioning in remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63:1-13.
- [34] SUN D, YAO J, XUE W, et al. Mask approximation net: A novel diffusion model approach for remote sensing change captioning[J]. IEEE transactions on geoscience and remote sensing, 2025, 63:1-11.
- [35] CAI M, CHEN H, LI C, et al. Image editing based on diffusion model for remote sensing image change captioning[C]//2024 IEEE International Conference on Signal, Information and Data Processing (ICSIDP). IEEE, 2024:1-5.
- [36] LIU C, CHEN K, CHEN B, et al. Rscama: Remote sensing image change captioning with state space model[J]. IEEE Geoscience and Remote Sensing Letters, 2024, 21:1-5.
- [37] CAP X, DONG W, QU J, et al. Rmnet: Dual-dimensional difference recalibration-guided cnn vmamba synergistic network for remote sensing image change captioning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2026, 64:1-13.
- [38] SUN D, BAO Y, LIU J, et al. A lightweight sparse focus transformer for remote sensing image change captioning[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 17:18727-18738.
- [39] DENG P, ZHOU W, WU H. Changechat: An interactive model for remote sensing change analysis via multimodal instruction

- tuning[C]//ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025:1-5.
- [40] ZHU D, HUANG X, HUANG H, et al. Change3d: Revisiting change detection and captioning from a video modeling perspective[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025:24011-24022.
- [41] DUMAN E, SERTTAS O, OZELBAS E, et al. Blip-cc: Adapting the blip for change captioning task in remote sensing[C]//2025 33rd Signal Processing and Communications Applications Conference (SIU). IEEE, 2025:1-4.
- [42] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002:311-318.
- [43] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]//Proceedings of the ACL-04 Workshop on Text Summarization Branches Out. 2004:74-81.
- [44] BANERJEE S, LAVIE A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments[C]//Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005:65-72.
- [45] VEDANTAM R, LAWRENCE ZITNICK C, PARIKH D. Cider: Consensus-based image description evaluation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015:4566-4575.

作者简介:



李浩 (2002-), 男, 硕士, 研究方向: 图像处理, 遥感图像变化描述,
E-mail: lh18537392583@163.com。



谢明鸿 (1976-), **通信作者**, 男, 博士, 高级工程师, 研究方向: 图像处理, 计算机视觉, E-mail: minghongxie@163.com。

Multi-scale Feature Compensation and Change Enhancement for Remote Sensing Image Change Captioning

LI Hao, XIE Minghong

(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming Yunnan 650500, China)

Abstract: Remote Sensing Image Change Captioning (RSICC) aims to generate natural language descriptions that reflect land-surface alterations using multi-temporal images, holding substantial value in domains such as land-resource monitoring and disaster assessment. However, in practical deployment, bi-temporal images are frequently vulnerable to complex non-semantic interference, such as seasonal successions and sudden illumination fluctuations, which easily induces "semantic hallucinations" in deep networks. Moreover, the ubiquitous simple differencing mechanisms adopted by existing methods struggle to effectively suppress these confounding factors and lack sufficient sensitivity to highly localized change signals, thereby yielding imprecise descriptive captions with poor contextual alignment. To tackle these multi-faceted challenges, this paper proposes a robust remote sensing image change captioning framework grounded on multi-scale feature compensation and change enhancement. To effectively alleviate the severe interference stemming from pseudo-change factors like lighting variations and environmental dynamics, a novel Feature Compensation Network (FCN) is designed. Leveraging a bidirectional guided compensation mechanism via cross-temporal attention, the FCN adaptively modulates bi-temporal features to enforce strict semantic and stylistic consistency across non-change background regions, thereby achieving the explicit suppression of illumination and shadow pseudo-changes. Concurrently, to overcome the perceptual deficiency regarding subtle change signals, a Discriminative Change-Aware Guided Module (DAGM) and a multi-scale fusion strategy are introduced. By synchronously reinforcing highly discriminative representations across both spatial and channel dimensions, this module substantially elevates the framework's operational sensitivity to the minor and intricate evolution of complex ground objects. Comprehensive quantitative comparison experiments conducted on the challenging LEVIR-CC and Dubai-CC benchmark datasets demonstrate that the proposed method consistently and significantly outperforms mainstream state-of-the-art algorithms across multiple core evaluation metrics, such as BLEU-4 and CIDEr, validating its distinct superiority in generating accurate and grammatically coherent text descriptions. Furthermore, qualitative analysis and comprehensive visualization results provide intuitive, empirical evidence supporting the robust cross-temporal interpretability of the model. Meanwhile, rigorous ablation studies systematically verify the theoretical necessity and practical effectiveness of each key proposed component. Ultimately, both experimental and theoretical investigations indicate that the proposed Feature Compensation Network (FCN) successfully achieves bidirectional interactive alignment and explicit interference suppression within non-change regions. Meanwhile, the seamlessly integrated Change-Aware Guided Module (DAGM) and multi-scale feature fusion strategy can effectively mine, preserve, and reinforce fine-grained details within genuine change regions, offering a highly promising and generalizable solution for high-precision earth observation and semantic change interpretation tasks.

Highlights:

1. A novel Feature Compensation Network (FCN) is proposed to adaptively modulate multi-scale bi-temporal features, enforcing background consistency and achieving explicit suppression of illumination and shadow pseudo-changes.
2. A Difference Awareness Guidance Module (DAGM) and a multi-scale fusion strategy are introduced to synchronously mine spatial and channel discriminative information, significantly enhancing sensitivity to genuine change regions and addressing the issues of feature insensitivity and missing textual details.
3. Extensive experiments conducted on the LEVIR-CC and Dubai-CC datasets demonstrate that the proposed

Foundation items: National Natural Science Foundation of China (Nos 62161015).

Received: 2026-04-24; **Revised:** 2026-05-29

***Corresponding author, E-mail:** minghongxie@163.com.

framework achieves superior performance across all core evaluation metrics, including BLEU, ROUGE-L, METEOR, and CIDEr.

Key words: remote sensing image change captioning; feature compensation; multi-scale difference enhancement; pseudo-change suppression; deep learning