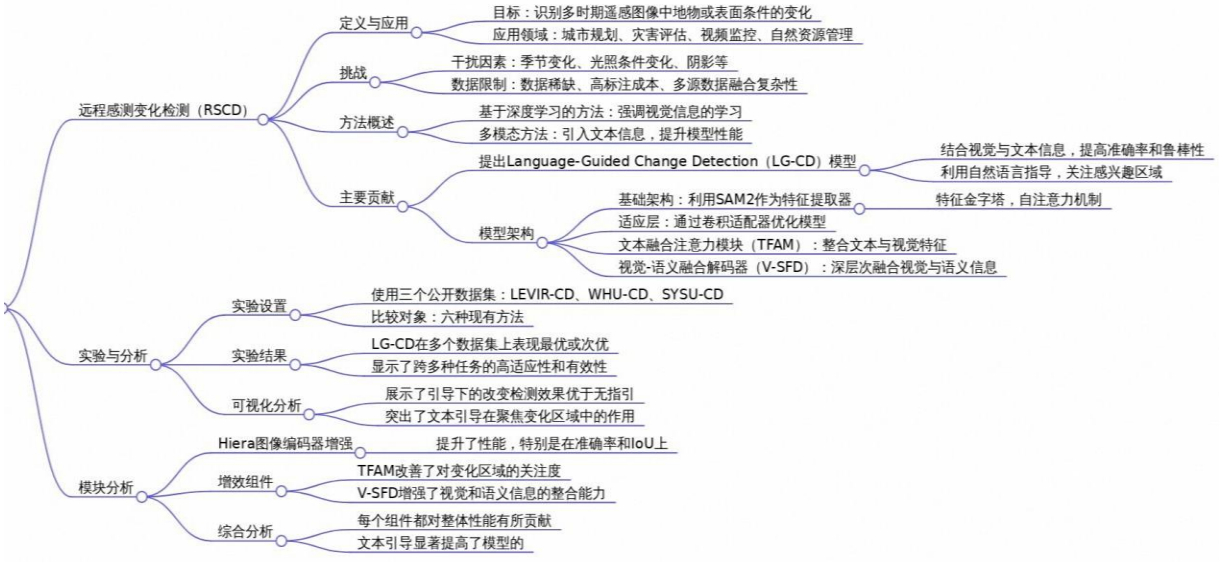
LG-CD:Enhancing Language-Guided ChangeDetection through SAM2 Adaptation

全文摘要

全文概述



本文提出了一种基于 SAM2 适配的语言引导变化检测模型 LG-CD，旨在解决传统遥感变化检测方法依赖单一视觉信息导致的泛化能力不足问题。该模型通过整合自然语言提示与视觉特征，显著提升了变化检测的准确性和鲁棒性。核心创新点包括：1）采用预训练的 SAM2 编码器作为视觉特征提取器，通过多层适配器进行微调，实现多尺度特征融合；2）设计文本融合注意力模块，将 CLIP 生成的文本语义嵌入与视觉特征对齐，引导模型聚焦关键变化区域；3）构建视觉-语义融合解码器，通过交叉注意力机制深度整合多模态信息，生成高精度变

化检测掩码。实验部分在 LEVIR-CD、WHU-CD 和 SYSU-CD 三个公开数据集上验证了模型性能，结果显示 LG-CD 在召回率指标上分别比次优方法高出 1.65%、2%和 2.79%，且在小目标检测和多类变化识别中表现出色。消融实验进一步证明了各模块的有效性，其中文本引导机制使模型在 LEVIR-CD 数据集上的 F1 分数提升了 3.2%。该研究为多模态信息融合在遥感变化检测中的应用提供了新思路，尤其在复杂场景和动态变化检测中展现出显著优势。

术语解释

1. **SAM2**: 基于视觉 Transformer 架构的预训练基础模型，通过分层编码器提取多尺度特征，具备强大的跨数据集泛化能力。在本文中作为 LG-CD 的视觉特征提取器，通过适配器微调实现变化检测任务的快速迁移。
2. **TFAM**: 文本融合注意力模块，采用多头交叉注意力机制将 CLIP 生成的文本语义嵌入与视觉特征对齐。该模块通过空间注意力机制增强模型对关键变化区域的聚焦能力，有效提升检测精度。
3. **V-SFD**: 视觉-语义融合解码器，通过多尺度特征融合与交叉注意力机制深度整合视觉和语义信息。该解码器采用类 FPN 结构，结合全局文本嵌入与局部视觉特征，生成高分辨率变化检测掩码。

论文速读

论文方法

方法描述

本篇论文提出了一种名为 LG-CD（Language Guided Change Detection）的多模态图像变化检测方法。该方法主要包含四个部分：SAM2 编码器及适配器、文本融合注意力模块、视觉语义融合解码器以及相似度计算与二值化处理。

首先，使用 SAM2 编码器提取两张不同时期的遥感影像数据中的多尺度特征，并通过多个轻量级适配器进行优化。接着，利用文本融合注意力模块将输入的文本提示信息与视觉特征相结合，进一步强化模型的任务导向性和关注焦点。最后，视觉语义融合解码器整合了视觉和语义信息，生成最终的变化检测掩膜。

方法改进

相较于传统的基于单一视觉或文本信息的图像变化检测方法，LG-CD 方法结合了视觉和语义信息，使得模型能够更好地理解任务目标并准确地定位变化区域。此外，文本融合注意力模块的设计可以有效地捕捉到任务相关的语义信息，从而提高模型的表现力。

解决的问题

LG-CD 方法旨在解决传统单模态图像变化检测方法存在的局限性，如无法充分考虑任务相关的信息、难以准确识别变化区域等问题。通过引入多模态信息和文本融合注意力机制，LG-CD 方法在图像变化检测任务中取得了更好的性能表现。

论文实验

本文主要介绍了作者提出的 LG-CD 方法在遥感影像变化检测任务中的性能表现，并与其他六种先进的遥感影像变化检测方法进行了比较。实验结果表明，LG-CD 在不同的数据集上都取得了最好的或第二好的性能，特别是在关键的召回指标上表现出色。此外，作者还进行了视觉分析和消融实验来验证 LG-CD 的有效性和各个模块的作用。具体来说：

首先，作者使用了三种广泛认可的遥感影像变化检测数据集（LEVIR-CD、WHU-CD 和 SYSU-CD）对 LG-CD 和其他六种先进的遥感影像变化检测方法进行了比较。实验结果表明，LG-CD 在不同的数据集上都取得了最好的或第二好的性能，特别是在关键的召回指标上表现出色。这些结果不仅证实了 LG-CD 在不同变化检测任务中的鲁棒性和优越性，而且证明了有效整合多模态信息可以进一步提高性能，突出了 LG-CD 的创新性和实用价值。

其次，作者进行了视觉分析来强调将语义信息融入到视觉特征提取过程中的重要性。实验结果表明，在没有语义指导的情况下，模型会同时关注多个目标，如建筑物、道路和桥梁等。但是，通过嵌入语义指导，模型的关注点集中在指定的建筑目标上。这一观察结果表明，将语义信息嵌入到视觉特征提取过程中可以有效地集中模型的注意力，使其更加专注于与任务相关的特征，从而提高了检测性能和任务相关性。

最后，作者进行了消融实验来验证 LG-CD 的有效性和各个模块的作用。实验结果表明，使用 Hiera 作为图像编码器显著优于 ResNet 基

关键图表解读

关键图表解读

图 1：LG-CD 模型整体架构展示了双时相遥感图像与文本提示的输入流程。SAM2 编码器提取多尺度特征，通过多层适配器进行任务微调，TFAM 模块融合文本与视觉特征，最终由 V-SFD 解码器生成高精度变化检测掩码。该架构突出了多模态信息的协同处理机制。

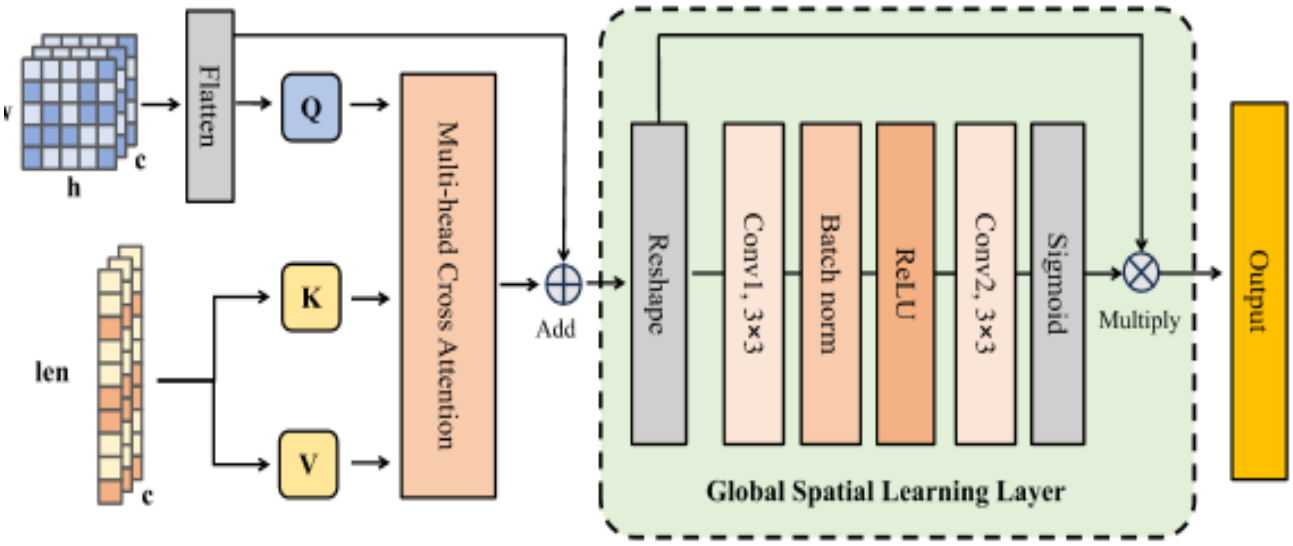


图 2：TFAM 模块结构揭示了文本特征与视觉特征的融合过程。通过多头交叉注意力机制，将 CLIP 编码的文本词嵌入与视觉特征进行交互，结合全局空间学习层的空间注意力机制，生成融合后的特征图。该设计实现了文本语义对视觉特征的精准引导。

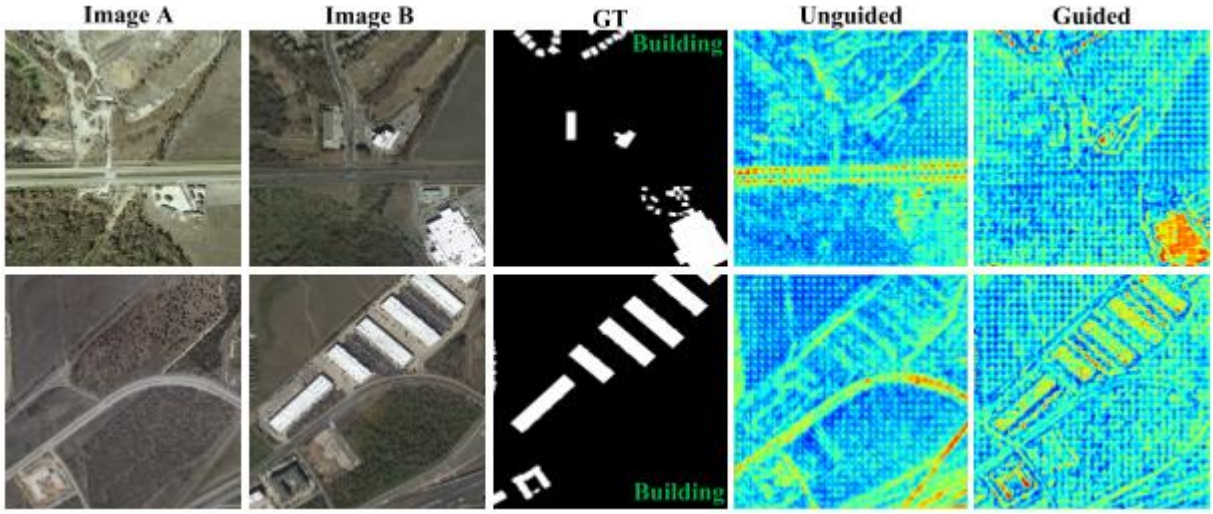
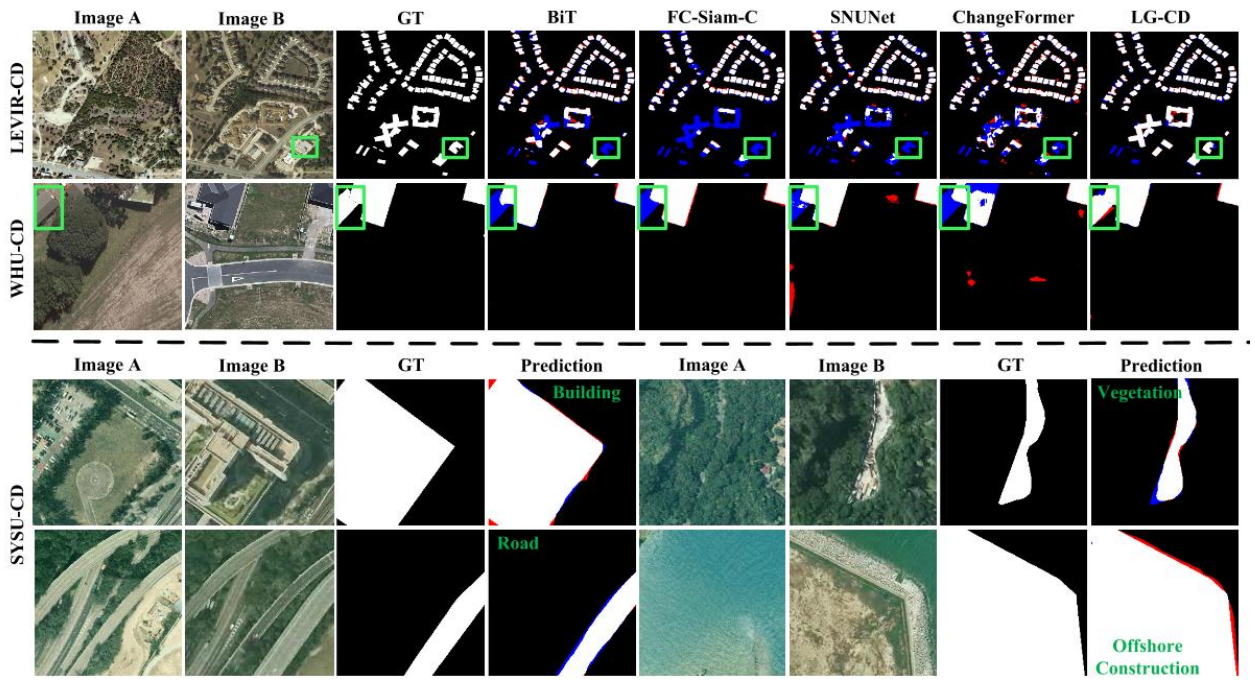


图 3：可视化分析对比了 LG-CD 与其他 6 种 SOTA 方法在 LEVIR-CD 和 WHU-CD 数据集上的检测结果。结果显示 LG-CD 在小目标区域（绿色框标注）具有更低的误检率（蓝色区域）和漏检率（红色区域），验证了其在复杂场景下的鲁棒性。此外，不同文本提示引导下的检测结果展示了模型的语义泛化能力。



论文总结

文章优点

该论文提出了一种新颖的语言引导变化检测模型（LG-CD），通过整合视觉和语义信息来显著提高远程感知变化检测的准确性和鲁棒性。该模型利用预训练的视觉基础模型（SAM2）作为其骨干网络，并使用多层适配器进行微调，使其能够快速适应远程感知变化检测任务。此外，设计了基于注意力的模块 TFAM 和 V-LFD，以对齐并深度融合视觉和语言特征，从而有效地捕捉变化模式并准确地生成变化检测掩模。在三个变化检测数据集上的实验结果充分验证了 LG-CD 的有效性，证明其优于其他方法。

方法创新点

该论文的方法创新点在于将自然语言提示与视觉信息相结合，提高了变化检测的准确性。具体来说，该模型利用 CLIP 作为文本特征提取器，利用预训练的 SAM2 作为共享视觉特征提取器，并设计了一个文本融合注意模块来对齐文本和视觉特征。此外，还设计了一个视觉语义融合解码器来生成高精度的变化检测掩模。这些创新点使得该模型能够更好地理解复杂场景和变化模式，从而实现对多种目标的可靠检测。

未来展望

未来研究可以进一步扩展该框架以容纳更多种类的语义场景，实现通用化的语言引导变化检测。此外，还可以探索如何将其他类型的数据，如声音或传感器数据，与视觉和语言信息相结合，以实现更全面的变化检测。