

基于 Cross-Transformer 差异特征融合的高分辨率光学遥感影像变化检测

彭代锋^{1,2}, 周顶蔚¹, 管海燕¹

1. 南京信息工程大学 遥感与测绘工程学院, 南京 210044;

2. 自然资源部遥感导航一体化应用工程技术创新中心, 南京 210044

摘要: 针对卷积网络在变化检测中面临的感受野有限、跨层次特征交互不足及长距离上下文依赖建模困难等问题, 本文提出基于 Cross-Transformer 差分特征融合的变化检测模型 CTDFNet (Cross-Transformer-based difference feature fusion network for change detection)。首先, 利用 ResNet18 网络构建多尺度局部特征表达, 并结合 Cross-Transformer 模块学习差异特征表示以实现全局上下文感知和多层次特征交互, 增强模型对复杂场景变化的敏感度; 然后, 解码部分利用密集连接结构实现特征的多路径传播和信息的充分利用, 同时抑制梯度消失; 最后, 结合自适应通道增强模块 ACEM (adaptive channel enhancement module) 动态调整通道权重以增强关键信息特征表达, 并利用 Sigmoid 层生成最后变化图。为验证本文方法的有效性, 采用 LEVIR-CD (LEVIR building Change Detection) 数据集、CDD (Change Detection Dataset) 和 WHU-CD (WHU building Change Detection) 数据集等 3 个变化检测数据集进行实验。定量结果表明, 与对比方法相比, CTDFNet 在 3 个数据集均取得最优精度指标, 其 F1 分数分别达到 91.52%、95.83% 和 93.02%。此外, 目视结果表明, CTDFNet 能有效抵抗复杂背景、阴影和光照变化等因素干扰, 可显著增强网络对不同尺度目标的检测识别效果。

关键词: 变化检测, 深度学习, Cross-Transformer, 交叉注意力机制, 自适应通道增强

中图分类号: P237/P2

引用格式: 彭代锋, 周顶蔚, 管海燕. 2026. 基于 Cross-Transformer 差异特征融合的高分辨率光学遥感影像变化检测. 遥感学报, 30(6): 1825-1840

Peng D F, Zhou D W and Guan H Y. 2026. High-resolution optical remote sensing image change detection based on Cross-Transformer difference feature fusion. National Remote Sensing Bulletin, 30(6): 1825-1840 [DOI: 10.11834/jrs.20264292]

1 引言

随着传感器技术的不断发展, 空天地一体化对地观测系统不断完善, 多模态、多时相、多分辨率遥感数据获取日益方便, 为地表时空分析与理解提供了海量数据源。然而, 如何有效挖掘海量数据中蕴含的知识、实现遥感影像智能化解译仍有待研究。变化检测技术可结合地物几何物理特性及遥感成像机制, 利用覆盖同一地区的多时相遥感影像提取感兴趣地物的位置、范围、属性等变化信息, 对揭示地表时空变化规律、挖掘遥感影像潜能具有重要意义, 在国土资源调查、生态保护红线监测、生态环境评估、城市扩张监测、应

急抢险救援等领域得到广泛应用 (Pekel 等, 2016; Tewkesbury 等, 2015; Song 等, 2018; Dalponte 等, 2019)。

根据解译单元的大小, 传统变化检测方法主要分为像素级与对象级两类 (眭海刚 等, 2018)。像素级方法以单个像元为基本分析单元, 通过直接比较不同时相影像对应位置的光谱特征差异来识别变化。该方法原理简单、计算高效, 尤其适用于中低分辨率遥感影像的大范围变化监测, 具体包括图像代数法 (Marin 等, 2014)、图像变换法 (Deng 等, 2008)、图像分类法 (Wu 等, 2017) 等。然而, 上述方法无法有效利用空间上下文信

收稿日期: 2024-07-15; 预印本: 2026-03-31

基金项目: 国家自然科学基金(编号:42371449, 41801386)

第一作者简介: 彭代锋, 研究方向为遥感影像智能解译。E-mail: daifeng@nuist.edu.cn

息, 容易导致大量虚检。为此, 马尔科夫随机场 (Benedek 等, 2015)、条件随机场 (Lv 等, 2018) 等概率图模型被广泛引入到变化检测领域。随着影像空间分辨率的提高, “异物同谱”和“同谱异物”现象加剧, 导致像素级变化检测无法有效提取精细尺度变化信息。为此, 研究人员提出对象级变化检测方法。该方法首先将双时相影像分割成互不重叠的影像对象, 进而根据对象级特征实现对象级特征比较 (张良培和武辰, 2017) 或分类后比较 (柳思聪 等, 2023; 佟国峰 等, 2015)。总体而言, 对象级变化检测方法可充分利用空间上下文信息以及丰富的对象级特征, 较好地提高了高分辨率遥感影像变化检测的准确性和可靠性。然而, 由于地物的多尺度效应及时空分布的复杂性, 分割尺度往往难以确定。

近年来, 由于强大的深度特征表示能力和非线性问题建模能力, 深度学习技术已在多个关键领域取得突破性进展, 并被广泛引入到遥感领域, 如目标检测 (Zhao 等, 2019)、影像分类 (Wang 等, 2021)、超分重建 (Dong 等, 2022b) 等。与传统变化检测方法相比, 深度学习变化检测方法可学习判别能力强的特征表达且自动化程度更高, 在精度和效率方面取得了传统方法无法比拟的优势。特别地, 基于深度学习的变化检测方法可自动学习多层次判别性特征表示, 且能够以端到端方式学习变化模式, 从而极大地简化了处理流程, 并展现出更优的鲁棒性及场景泛化能力。早期的深度学习变化检测方法往往利用预训练卷积神经网络 CNN (convolutional neural network) 提取深度特征以生成差异图 (El Amin 等, 2017; Zhang 等, 2016), 进而根据阈值分割或聚类差异影像分析方法得到变化图。为实现变化信息端到端提取, Daudt 等 (2018) 提出 3 种基于 UNet 结构的全卷积变化检测网络, 即 FC-EF (Fully Convolutional Early Fusion)、FC-Siam-Conc (Fully Convolutional Siamese Concatenation) 和 FC-Siam-Diff (Fully Convolutional Siamese Difference)。其中, FC-Siam-Conc 和 FC-Siam-Diff 采用经典的孪生架构, 在特征提取阶段利用共享权重的编码器提取多时相影像特征, 然后分别使用特征级联和差值方式融合多时相特征。类似地, Zheng 等 (2021) 提出了一种新的跨层卷积网络 CLNet (cross-layer convolutional neural network), 该网络通过引入跨层特征聚合模块自适应地整合多尺度特征与多层次上下文信息。

Peng 等 (2019) 在变化检测网络中引入密集连接机制, 可直接学习多时相影像变化模式, 且融合多个中间层输出以有效捕捉多尺度变化信息。然而, 由于卷积操作固有的有限感受野, 上述方法难以建模长距离的全局上下文依赖。为此, 注意力机制被广泛引入以增强模型的上下文感知能力 (梁哲恒 等, 2022; 刘英 等, 2024; Chen 和 Shi, 2020)。例如 Dong 等 (2022a) 联合通道和空间注意力机制实现多尺度上下文信息融合; Liu 等 (2022) 提出局部特征金字塔模块和全局空间金字塔模块, 较好地提取了多时相影像中不同尺度建筑物的变化特征; Chen 等 (2021) 提出了一种基于深度度量学习和双注意力机制的变化检测网络, 有效提高了特征的判别性。

值得注意的是, 基于自注意力机制的 Transformer 可有效捕捉全局上下文信息, 在视觉任务中得到广泛应用, 并被成功引入到变化检测领域 (Zhang 等, 2023; Liu 等, 2023)。例如 Chen 等 (2022) 提出一种基于 Transformer 的变化检测网络, 通过引入语义 Token 提高多时相影像深度特征表达能力, 有效抑制了背景噪声; Bandara 和 Patel (2022) 引入基于 Transformer 编码器与多层感知机 MLP (multilayer perceptron) 的孪生变化检测网络, 可高效地提取多尺度长距离细节信息; Liu 等 (2024) 通过结合多尺度特征信息和 Transformer 自注意力机制有效抑制了伪变化干扰; Feng 等 (2023) 通过线性交叉注意力机制融合基于 CNN 和 Transformer 的双分支编码器特征, 增强了模型对变化特征的表达能力; He 等 (2024) 提出渐进特征交互模块和跨尺度特征融合模块, 能够有效缓解变化检测中存在的语义模糊与空间细节损失问题, 从而提升变化区域的定位精度。

然而, 现有变化检测网络仍存在以下问题: (1) 多时相差异信息融合往往依赖差分影像, 容易导致变化信息混叠, 降低了变化目标特征表达的可靠性; (2) 受复杂背景、阴影、光照变化等因素干扰, 变化区域细节边缘信息容易缺失, 存在边缘模糊和漏检等问题; (3) 现有变化检测方法缺乏对跨层次特征交互的显式建模, 难以充分挖掘和利用不同语义层次特征间的依赖关系与互补性, 导致对变化区整体结构与深层语义捕捉能力不足。为此, 本文提出了一种基于 Cross-Transformer

差异特征融合的变化检测模型 CTDFNet (Cross-Transformer-based difference feature fusion network for change detection), 有效实现了双时相特征交互增强与变化信息鲁棒提取。首先, 针对不同层次特征间交互受限的问题, 引入 Cross-Transformer 模块 CTM (Cross-Transformer module) 以更好地捕捉到序列内部和序列间的关联性, 实现更全面的信息交互。其次, 采用多尺度特征融合策略, 结合多阶段 CTM 增强特征判别能力以有效抑制复杂背景干扰。此外, 在解码器中采用密集连接结构以充分融合多尺度信息、抑制语义差距。为进一步优化输出特征质量, 在输出端引入自适应通道增

强模块 ACEM (Adaptive Channel Enhancement Module), 通过对特征通道进行自适应加权, 抑制由语义鸿沟与空间错位引起的特征紊乱, 从而显著提升最终变化检测结果的可靠性与一致性。

2 研究方法

2.1 CTDFNet 模型结构

本文提出的 CTDFNet 模型整体结构如图 1 所示, 该网络融合了 CNN 的局部感知能力与 Transformer 的全局建模优势, 主要由编码器、CTM 模块、ACEM 模块和解码器组成。

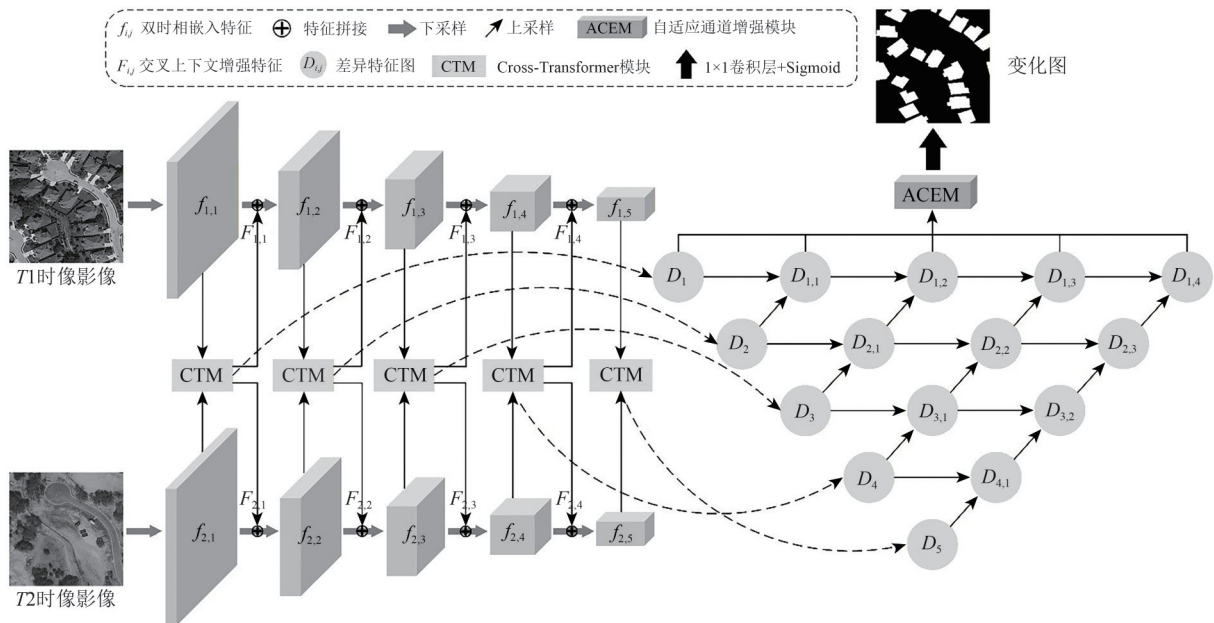


图1 CTDFNet模型整体结构

Fig. 1 Overall framework of CTDFNet

首先, 编码器提取的多尺度特征被并行送入 CTM 模块, 通过全局特征交互学习不同位置长程上下文特征, 同时引入双分支融合前馈神经网络 BF-FFN (bi-branch fusion feed forward network) 以进一步捕捉变化区域的局部细节信息。其次, 将前层输出下采样后与后层输入特征在通道维度上进行级联, 实现全局-局部特征融合, 并对每一层融合后特征进行逐元素差分以生成差异特征图。特别地, 与单一跳跃连接相比, 密集连接机制可在不同层次特征图间引入更丰富的跳跃连接, 不仅能在前向传播中有效传递并融合多尺度特征, 显著增强特征表达能力, 还能在反向传播中缓解梯度消失, 从而加快网络收敛。因此, 本文在解

码器中集成密集连接策略, 并通过一系列上采样操作生成多个与输入图像尺寸一致的中间层特征图。最后, 引入 ACEM 对特征图进行通道自适应的权重调整, 有效增强变化区的特征表征能力, 并利用 1×1 卷积层和 Sigmoid 层输出最终变化图。

2.2 主干网络

由于 ResNet18 网络参数量较少且具备强局部特征表示能力, 本文利用 ResNet18 作为 CTDFNet 主干网络。具体地, ResNet18 网络基于残差卷积单元构建 (如图 2 所示, 残差单元依次包含 3×3 卷积层、跳跃连接以及 ReLU 激活层), 通过引入残差连接机制抑制梯度消失, 同时加速模型收敛、优化训练过程。

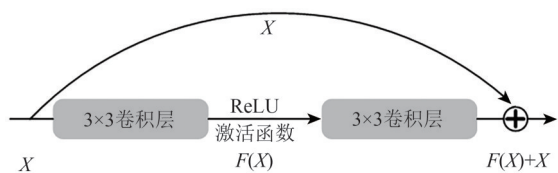


图2 残差块结构

Fig. 2 The structure of residual block

2.3 Cross-Transformer 模块

为抑制变化检测中复杂背景干扰和特征表达不足导致的变化区域内部破碎、边界模糊等问题, 本文提出用CTM模块实现双时相影像特征的显示交互与融合, 其核心创新之处在于: (1) 提出一种层级化交叉注意力机制, 通过将上一尺度注意力输出作为输入传递至当前尺度交叉注意力模块

进行处理, 实现了多尺度上下文信息的逐级汇聚与融合, 可有效抑制现有注意力机制在复杂场景下产生的背景响应, 提升特征表达的鲁棒性; (2) 设计了一种BF-FFN, 通过结构解耦与自适应融合增强了特征的局部表达力和全局一致性, 有效提升了变化区域的内部完整性与边界清晰度。CTM模块的核心结构如图3所示, 主要由位置编码、多头交叉注意力MHCA (multi-head cross attention) 模块以及BF-FFN依次串联构成。其中, CTM模块通过MHCA模块学习特征序列内和序列间的长距离依赖关系, 有效增强不同层次特征交互能力并抑制复杂背景干扰, 进而利用BF-FFN模块进一步增强网络的局部特征表达能力, 从而有效改善变化区的内部完整性和边缘定位能力。

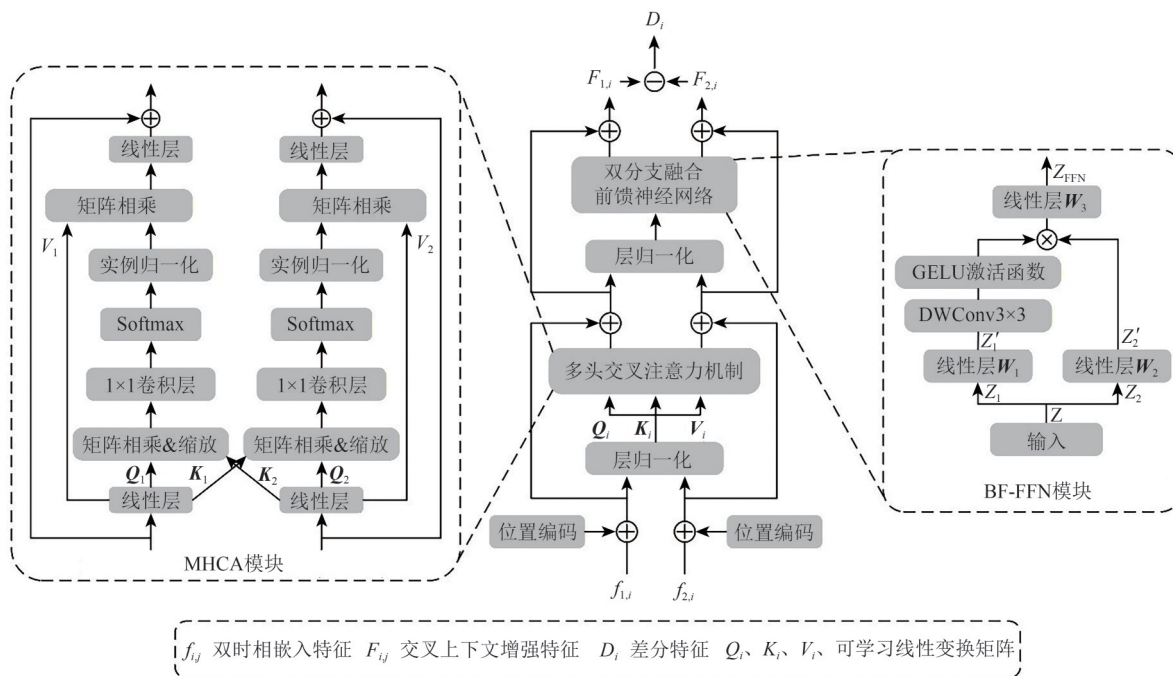


图3 CTM结构

Fig. 3 The structure of CTM

2.3.1 多头交叉注意力机制

MHCA模块整体结构如图3左侧小图所示, 主要由线性层、矩阵相乘层、 1×1 卷积层、Softmax层和实例归一化层构成。

首先, 对两时期输入特征 $f_1 \in \mathbb{R}^{C \times H \times W}$ 与 $f_2 \in \mathbb{R}^{C \times H \times W}$ (C 、 H 、 W 分别表示特征图通道数、高度和宽度) 构建输入Token序列, 即通过棋盘分割方式生成尺寸为 $N \times N$ 的非重叠图像块, 并展平为向量以生成维度为 $C \times N^2$ 的图像块序列 f^p 。其次, 通过

层归一化对序列进行标准化操作, 并加入位置编码以保留图像中的空间结构信息, 得到特征嵌入 f^{pe} :

$$f^{pe} = \text{LN}(f^p) + E_{\text{pos}} \quad (1)$$

式中, $\text{LN}(\cdot)$ 表示层归一化操作, E_{pos} 为可学习的位置编码。

如图3所示, 将 f_t^{pe} ($t = 1, 2$) 输入到MHCA模块学习双时相特征长程上下文关联。该模块首先通过一组可学习线性变换矩阵 W_t^Q 、 W_t^K 、 W_t^V 生成查询向量 Q_t 、键向量 K_t 和值向量 V_t :

$$Q_t = f_t^{\text{pe}} W_t^Q \quad (2)$$

$$K_t = f_t^{\text{pe}} W_t^K \quad (3)$$

$$V_t = f_t^{\text{pe}} W_t^V \quad (4)$$

其次, 通过将来自不同序列的查询向量与键向量相乘得到相似度矩阵; 接着, 通过一个 1×1 卷积层进一步建模特征交互, 并经 Softmax 归一化操作生成注意力得分矩阵; 最后, 将注意力得分矩阵与对应值向量进行加权聚合, 且引入实例归一化操作提升不同位置综合关注能力。假设上述交叉注意力表示为 $\text{CA}(Q_m, K_n, V_m)(m, n \in \{1, 2\}, m \neq n)$, 则有

$$\text{CA}(Q_m, K_n, V_m) = \text{IN} \left(\text{softmax}(\text{Conv}_{1 \times 1}(\frac{Q_m K_n^T}{\sqrt{d_k}})) V_m \right) \quad (5)$$

式中, $\sqrt{d_k}$ 表示缩放因子, $\text{IN}(\cdot)$ 表示实例归一化操作。

为充分利用不同表示子空间的注意力信息, CTM 采用多个交叉注意力头独立计算, 以生成不同交叉注意力表示, 并利用投影层融合拼接后的多头注意力表示。若 L 个交叉注意力头表示为 $\text{head}^l (l = 1, 2, \dots, L)$, 则有

$$\text{MHCA}(f^{\text{pe}}) = \text{Concat}(\text{head}^1, \text{head}^2, \dots, \text{head}^L) W^0 \quad (6)$$

式中, $\text{Concat}(\cdot)$ 表示级联操作, W^0 表示投影变换矩阵。

2.3.2 双分支融合前馈神经网络

传统前馈神经网络因采用全局全连接操作而缺乏位置敏感性, 难以有效顾及局部细节。鉴于

此, 本文提出一种 BF-FFN, 通过显式地挖掘每个 Token 局部依赖关系, 有效提升细节表达能力, 其结构见图 3 右侧小图。对于输入特征 Z , 将其沿通道维度平均切分为 Z_1 和 Z_2 , 并分别使用两个独立的线性变换矩阵 W_1 与 W_2 进行映射生成 Z'_1 和 Z'_2 :

$$Z'_1 = Z_1 W_1 \quad (7)$$

$$Z'_2 = Z_2 W_2 \quad (8)$$

接着, 对 Z'_1 进行深度可分离卷积操作, 在降低参数量的同时增强局部感受野, 进而引入激活操作并与 Z'_2 相乘。最后, 通过线性变换矩阵 W_3 生成输出特征 Z_{FFN} :

$$Z_{\text{FFN}} = \delta(\text{DWConv}(Z'_1)) Z'_2 W_3 \quad (9)$$

式中, $\delta(\cdot)$ 表示 GELU 激活函数, $\text{DWConv}(\cdot)$ 表示深度可分离卷积操作。

2.4 自适应通道增强模块

解码器中浅层特征包含丰富的细节与精确定位信息, 而深层特征则保留了高级语义信息。为融合二者优势, 本文提出 ACEM, 如图 4 所示。该模块通过动态校准通道权重, 自适应地融合不同解码层的输出特征, 从而有效抑制语义与定位差异。首先, 对解码器输出特征 $\{D_1, D_{1..i}, i \in \{1, 2, 3, 4\}\}$ 进行拼接得到特征 X 。其次, 通过并行的 3×3 卷积层和 5×5 卷积层提取多尺度特征, 并经过全局平均池化层后相加融合, 得到通道权重向量 $S (S \in \mathbb{R}^{C \times 1 \times 1})$ 。为保留更多空间信息, 进一步引入 1×1 卷积和 ReLU 激活函数, 并利用 Softmax 层生成通道权重 S' :

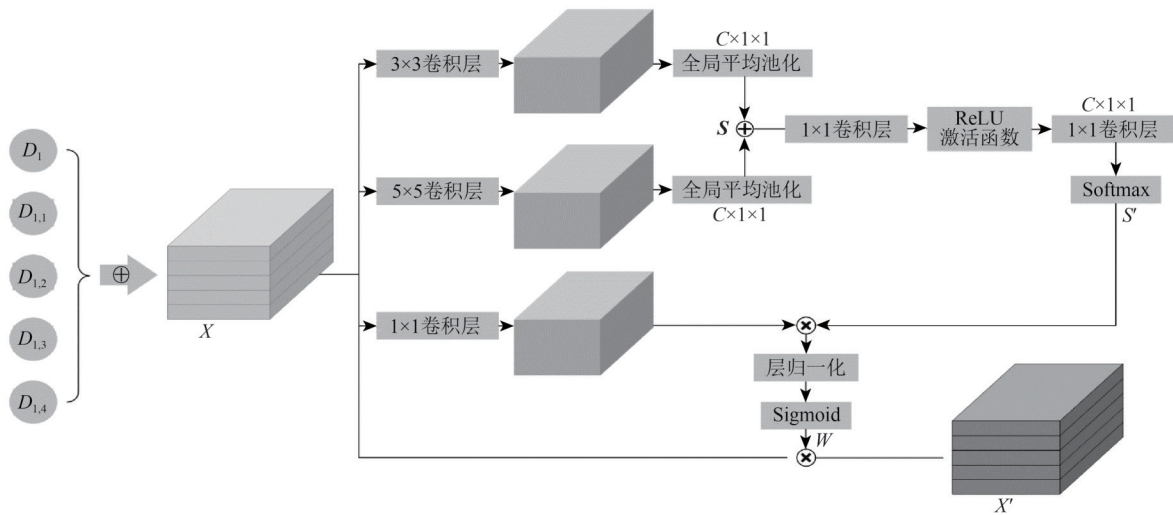


图4 ACEM 结构

Fig. 4 The structure of ACEM

$$S = \text{GAP}(\text{Conv}_{3 \times 3}(X)) + \text{GAP}(\text{Conv}_{5 \times 5}(X)) \quad (10)$$

$$S' = \text{Softmax}(\text{Conv}_{1 \times 1}(\varphi(\text{Conv}_{1 \times 1}(S)))) \quad (11)$$

式中, $\text{GAP}(\cdot)$ 表示全局平均池化操作, $\varphi(\cdot)$ 表示 ReLU 激活函数。

此外, 为建模通道间的交互关系, 对特征 X 进行 1×1 卷积变换并与权重 S' 点乘。最后, 利用层归一化 LN (layer normalization) 与 Sigmoid 激活函数校准后生成权重 W , 用于对特征 X 进行自适应加权, 以输出增强特征 X' :

$$W = \text{Sigmoid}(\text{LN}(S' \odot \text{Conv}_{1 \times 1}(X))) \quad (12)$$

$$X' = X \odot W \quad (13)$$

表 1 不同变化检测数据集参数统计信息

Table 1 Parameter statistics information for different change detection datasets

数据集	变化类型	影像来源	影像对尺寸	分辨率	训练集对数/验证集对数/测试集对数
LEVIR-CD	建筑物	Google Earth	256像素 × 256像素	0.5 m	7120/1024/2048
CDD	建筑物、道路、汽车	Google Earth	256像素 × 256像素	0.03—1 m	10000/3000/3000
WHU-CD	建筑物	航空影像	256像素 × 256像素	0.3 m	6096/762/762

(1) LEVIR-CD 数据集: 该数据集提供了 0.5 m 高分辨率的建筑变化样本, 包含 637 对尺寸为 1024 像素 × 1024 像素的双时相影像, 时间跨度为 5—14 年, 涵盖超过 3 万个独立建筑变化实例。为适应模型训练, 将 445 对训练集影像、64 对验证集影像和 128 对测试集影像统一裁剪为 256 像素 × 256 像素的非重叠图块, 生成 7120 对训练集、1024 对验证集与 2048 对测试集。

(2) CDD 数据集: 该数据集主要刻画季节性变化带来的地物差异, 包含 11 对高分辨率 (0.03—1 m) 影像, 涵盖建筑、车辆、道路等多种变化类型。实验中将影像统一裁剪为 256 像素 × 256 像素的影像块, 并按比例生成 10000 对训练集、3000 对验证集与 3000 对测试集。

(3) WHU-CD 数据集: 该数据集记录了新西兰基督城地震前后的建筑物变化, 由一对超大尺寸 (32507 像素 × 15354 像素) 的 0.3 m 分辨率的航拍图像构成。为构建训练样本, 采用滑动窗口法将其裁剪为 256 像素 × 256 像素的图块, 并按 8:1:1 的比例随机划分为训练集、验证集和测试集。

3.2 精度评估指标

为评估 CTDFNet 模型的综合性能, 本文采用精确度 P (Precision)、召回率 R (Recall)、 $F1$ 分数 $F1$ ($F1$ -score)、交并比 IoU (intersection over

式中, $\odot$ 表示逐元素相乘操作。

3 实验和结果分析

3.1 实验数据

本研究选用 LEVIR-CD (LEVIR building Change Detection) 数据集 (Chen 和 Shi, 2020)、CDD (Change Detection Dataset) (Lebedev 等, 2018) 和 WHU-CD (WHU building Change Detection) 数据库 (Ji 等, 2019) 等 3 个公开遥感变化检测数据集对所提方法的有效性进行综合验证。不同变化检测数据集参数统计信息如表 1 所示。

union) 和总体精度 OA (overall accuracy) 等 5 个精度指标。 P 衡量检测结果中变化像元的可靠程度, R 衡量模型对真实变化像元的覆盖能力, $F1$ 为精确度与召回率的调和平均数, IoU 是衡量预测变化区域与真实变化区域的空间重叠度, OA 反映模型整体分类正确的比例。 $F1$ 、IoU 和 OA 值越高, 则总体变化检测效果越好。5 个精度指标的公式分别为

$$P = \frac{TP}{TP + FP} \quad (14)$$

$$R = \frac{TP}{TP + FN} \quad (15)$$

$$F1 = \frac{2PR}{P + R} \quad (16)$$

$$\text{IoU} = \frac{TP}{TP + FN + FP} \quad (17)$$

$$\text{OA} = \frac{TP + TN}{TP + TN + FP + FN} \quad (18)$$

式中, TP (True Positive) 和 TN (True Negative) 分别表示正确预测的变化样本数量与未变化样本数量, FP (False Positive) 和 FN (False Negative) 分别表示错误预测的变化类样本数量与未变化类样本数量。

3.3 实验环境与设置

本文模型使用 PyTorch 深度学习框架实现, 模型训练与推理中均使用一块 24 GB 显存的 NVIDIA

GeForce RTX 3090 显卡。训练中使用AdamW 优化器（权重衰减参数为0.001，初始学习率为0.0001，并采用线性学习率衰减策略），批大小为4。为进一步提升模型的泛化能力，训练时对输入数据进行多种增强操作，包括水平翻转、垂直翻转、随机旋转、随机裁剪及高斯模糊。同时，训练过程中采用二元交叉熵损失函数 $\mathcal{L}_{\text{bce}}$ 优化模型参数：

$$\mathcal{L}_{\text{bce}} = -\frac{1}{N} \sum_{i=1}^N (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \quad (19)$$

式中， $\log$ 代表自然对数（即以e为底）， N 表示训练集中的样本数量， y_i 表示第*i*个样本的真值标签图， $\hat{y}_i$ 表示对应的预测概率图。

3.4 实验结果和分析

为评估CTDFNet网络的性能优势，本文选择了11种先进变化检测模型作为基准进行全面对比。

(1) FC-EF (fully convolutional early fusion) (Daudt等, 2018)：采用UNet结构，直接将双时相影像在通道维度拼接作为输入。

(2) FC-Siam-Conc (fully convolutional siamese concatenation) (Daudt等, 2018)：采用孪生架构对两期影像分别提取多尺度特征，并通过级联方式实现特征融合。

(3) FC-Siam-Diff (fully convolutional siamese difference) (Daudt等, 2018)：采用孪生架构分别提取多尺度特征，并通过差值方式实现特征融合。

(4) STANet (spatial-temporal attention neural network) (Chen和Shi, 2020)：利用时空注意力机制建模像素间长程依赖关系，生成更具判别性的特征表示。

(5) SNUNet (siamese nested U-Net) (Fang等, 2022)：采用孪生编码器与密集编码器结构，有效抑制解码器深层位置信息丢失。

(6) ChangeFormer (Bandara等, 2022)：结合层次化Transformer编码器与MLP解码器，有效捕捉多尺度长距离依赖信息。

(7) FTNet (fully transformer network) (Yan等, 2022)：利用Transformer全局上下文建模优势学习具有区分度的全局特征，引入金字塔结构实现多级视觉特征融合。

(8) BIT (bitemporal image transformer) (Chen等, 2022)：使用Transformer编码器挖掘多时相影

像语义信息，进而利用Transformer解码器增强原始特征。

(9) ICIFNet (intra-scale cross-interaction and inter-scale feature fusion network) (Feng等, 2022)：通过在尺度内和尺度间维度联合CNN和Transformer特征实现全局一局部特征表达。

(10) VcT (visual change transformer) (Jiang等, 2023)：引入RTM (reliable Token mining) 模块、自注意力—交叉注意力模块和锚点—主注意力机制模块增强原始深度特征表达。

(11) BiFA (bitemporal feature alignment) (Zhang等, 2024)：引入双时相特征对齐模块提高深度特征提取能力，并利用差分光流对齐模块构建多尺度差异特征，最后基于隐式神经对齐解码器生成变化图。

特别地，为系统比较不同对比方法，根据模型架构的不同进行分组，即分为CNN方法、Transformer方法和CNN-Transformer方法。3个变化检测数据集上不同方法的定量统计结果如表2所示。

3.4.1 LEVIR-CD数据集

从表2可以看出，相较于FC-EF、FC-Siam-Conc和FC-Siam-Diff等3个基础网络，基于密集连接的孪生网络SNUNet采用多个子解码器以增强特征传递能力，有效抑制了解码器深层位置信息丢失，进而提高了变化检测精度，其F1分数达到了88.79%。此外，STANet利用时空注意力模块增强了模型对显著性特征的捕捉能力，从而提升了检测性能。特别地，采用纯Transformer结构的FTNet和BiFA取得了对比模型中最高的精度指标，其F1分数分别达到90.82%和90.72%。BIT、ICIFNet和VcT通过结合CNN与Transformer的优势以融合局部特征和全局特征，可有效改善变化检测精度，其F1分数分别达到了89.43%、89.96%和90.11%。值得注意的是，CTDFNet在F1分数、IoU和OA等精度指标上均取得最高值，这是由于CTDFNet有效融合了CNN和Transformer网络的优势：通过CNN网络增强特征图细节信息表达能力，同时采用跨序列信息交互机制和多级特征融合策略以精准学习全局依赖关系，从而提高检测结果的准确性与鲁棒性。

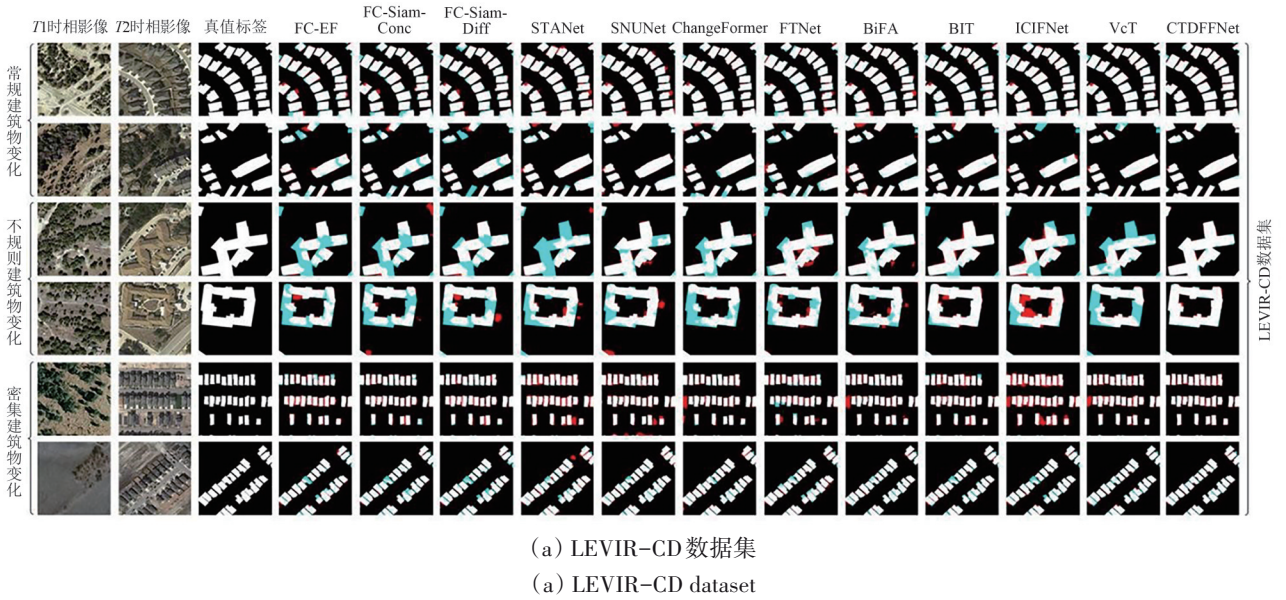
表2 3个变化检测数据集上不同方法的定量统计结果
Table 2 Quantitative statistical results of different methods on three change detection datasets

		/%		
方法		LEVIR-CD数据集	CDD数据集	WHU-CD数据集
		P/R/F1/IoU/OA	P/R/F1/IoU/OA	P/R/F1/IoU/OA
CNN方法	FC-EF	87.44/82.59/84.94/73.83/98.51	76.28/71.01/73.55/58.17/93.98	85.85/67.40/75.52/60.66/98.26
	FC-Siam-Conc	90.81/83.65/87.08/77.12/98.74	86.21/82.36/84.24/72.77/96.36	87.13/74.39/80.25/67.02/98.54
	FC-Siam-Diff	91.03/83.05/86.85/76.76/98.72	83.88/79.27/81.51/68.79/95.76	85.81/74.62/79.82/66.42/98.50
	STANet	84.73/92.60/88.75/79.78/98.87	88.00/94.32/91.10/83.60/97.83	84.92/88.57/86.71/76.53/98.95
	SNUNet	89.19/87.17/88.79/79.83/98.88	92.39/96.04/94.18/89.00/98.63	86.65/86.25/86.45/76.13/98.63
Transformer方法	ChangeFormer	92.05/88.81/90.40/82.48/99.04	93.45/96.08/94.54/89.65/98.71	91.83/88.02/89.88/81.63/99.12
	FTNet	92.38/89.32/90.82/83.17/99.08	88.00/93.98/90.89/83.30/97.93	96.31/88.06/92.00/85.19/99.21
	BiFA	91.54/89.91/90.72/83.01/99.06	94.90/94.80/94.90/90.29/98.80	93.58/91.13/92.34/85.77/99.24
CNN-Transformer方法	BIT	92.18/86.84/89.43/81.12/98.93	95.15/92.41/93.76/88.25/98.55	89.40/90.03/89.72/81.35/99.10
	ICIFNet	91.32/88.64/89.96/81.75/98.99	95.21/92.58/93.87/88.56/98.60	93.01/85.47/88.90/80.11/98.99
	VcT	92.65/87.71/90.11/82.01/99.02	95.54/91.61/93.54/87.86/98.51	93.75/91.82/92.78/86.53/99.28
本文方法	CTDFNet	92.63/90.45/91.52/84.37/99.15	95.19/95.78/95.83/92.00/98.97	95.99/90.23/93.02/86.95/99.32

注：加粗黑体数字和下划线数字分别表示最优结果和次优结果。

为定性评价各模型性能，选取3个代表性场景进行分析。其中，图5（a）前两行显示了常规建筑物变化场景检测效果，可视化结果显示，大多数对比算法在处理边缘细节时效果不佳，而CTDFNet模型可更好地识别和定位常规建筑物变化，清晰地检测出建筑物实体。图5（a）第三、第四行显示的不规则建筑物变化场景中，大多数对比方法的检测结果容易出现主体内部不连续现象，且边界处存在较多伪变化像素。与此相对，CTDFNet模型通过引入CTM模块在特征空间中建立高效多级特征交融机制，显著增强了网络对变

化前景与未变化背景的差异化判别能力，有效降低了变化检测漏检率和误检率，且建筑物变化区域边缘提取效果较好。图5（a）第五、第六行显示了密集建筑物变化场景。可以看出，CTDFNet模型可准确提取建筑物边缘变化信息，且显著降低了不同变化区域之间的粘连情况。这是由于CTDFNet模型基于CTM模块实现了双时相多尺度特征的层级化交互与融合，在实现长程上下文信息建模的同时较好地增强了局部特征表达，进而显著增强了模型抑制复杂背景干扰的能力，有效提升了变化区域的内部完整度和边界定位精度。



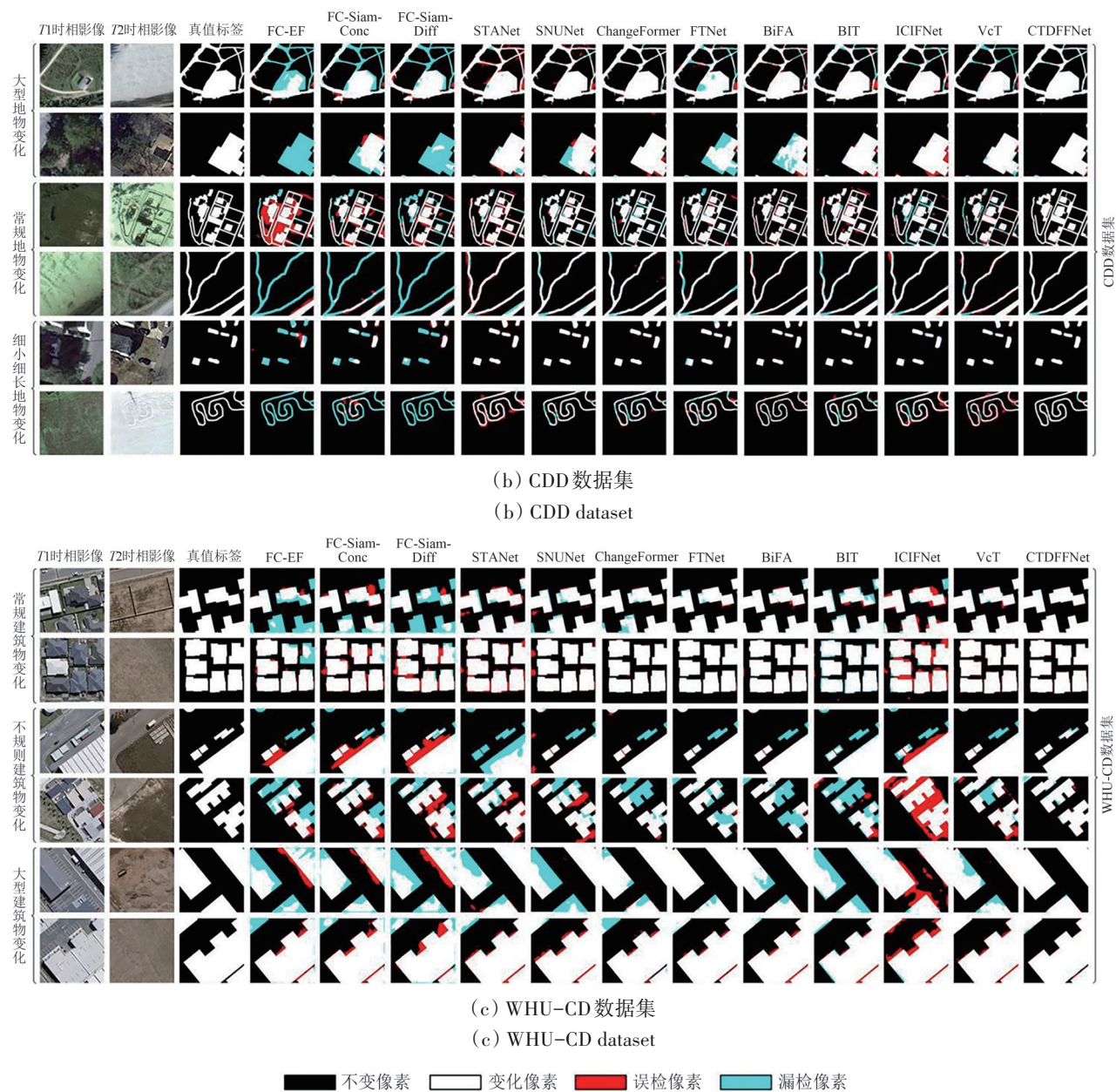


图5 不同数据集上各模型可视化对比结果

Fig. 5 Visual comparison results of various models on different datasets

3.4.2 CDD数据集

从表2可以看出，由于仅采用简单的跳跃连接操作融合编码器和解码器特征，前3个简单UNet架构模型表现不佳，STANet模型通过引入时空注意力机制，提高了融合特征的可靠性，变化检测精度指标得到有效改善。SNUNet模型在解码器中引入密集连接操作，同时使用通道注意力增强融合特征，变化检测精度得到进一步提高。此外，由于Transformer模型可较好地建立长距离依赖关系，基于纯Transformer结构的ChangeFormer、FTNet和BiFA模型，以及基于CNN-Transformer混合的

BIT、ICIFNet和VcT模型都不同程度地提高了变化检测精度。特别地，本文提出的CTDFNet模型利用CTM实现多层次特征交互，有效融合全局—局部信息，且利用ACEM抑制了语义差异、减少了信息损失，变化检测精度指标取得显著提升，与3个基础变化检测网络（FC-EF、FC-Siam-Conc、FC-Siam-Diff）相比，其F1分数分别提高了22.28%、11.59%和14.32%。

为定性分析不同方法效果，选择3个典型场景进行可视化分析。图5（b）第一、第二行展示了大型地物变化的情况，前3个简单UNet架构模

型检测结果中包含大面积漏检像素, 而引入 Transformer 模块的 ChangeFormer、FTNet、BiFA、BIT、ICIFNet 和 VcT 等对比模型中漏检率得到不同程度降低, STANet 和 SNUNet 模型可有效降低漏检率, 但误检像素较多。与之不同, 本文提出的 CTDFNet 在抑制漏检与误检方面均表现出显著优势, 生成的变化检测图与真值标签最接近。图 5 (b) 第三、第四行显示了建筑物、道路等常规地物变化场景。可以看出, 对比方法中存在较多漏检和误检像素, 而 CTDFNet 模型漏检率和误检率较低, 建筑物和道路结构提取完整度和准确度较高, 目视效果最佳。图 5 (b) 第五、第六行显示了细小细长地物变化场景。结果表明, 前 3 个简单 UNet 架构模型对这类细微变化特征的提取能力较弱, 难以准确识别道路的连续结构及车辆的完整轮廓。尽管 SNUNet、BIT 及 ICIFNet 等改进模型在捕捉离散小目标方面有所提升, 但在处理具有连续空间特征的道路目标时仍存在明显不足。ChangeFormer、STANet、FTNet、BiFA 和 VcT 等模型进一步提升了汽车等小目标检测效果, 但其边缘定位精度仍有不足, 存在明显的误检现象。与之不同, CTDFNet 模型通过 CTM 对全局依赖关系与局部细粒度信息进行高效建模融合, 在细小细长地物变化检测中展现出显著优势, 可准确地识别变化区的形状和结构信息, 取得最佳目视效果。总体而言, 由于缺乏多尺度差异特征的有效融合, FC-EF、FC-Siam-Conc 和 FC-Siam-Diff 等模型在检测建筑物、道路等实体时, 容易出现目标内部不完整或缺失等严重漏检现象; STANet 和 SNUNet 通过引入注意力机制和密集连接机制, 可较好地抑制漏检像素; 而其他对比模型 (ChangeFormer、FTNet、BiFA、BIT、ICIFNet 和 VcT) 通过引入 Transformer 模块建模全局上下文信息, 漏检率得到显著降低。本文方法通过引入层级化交叉注意力机制和双分支融合前馈网络, 有效增强了多尺度上下文信息建模与局部特征表达能力, 且利用 ACEM 模块进一步提升变化区定位精度, 显著增强了模型抑制复杂背景噪声与保持边缘定位精度能力。

3.4.3 WHU-CD 数据集

如表 2 所示, 相较于对比模型, CTDFNet 在 5 个定量指标上均取得了大幅度提升, 取得了最优

总体精度。相比之下, FC-EF、FC-Siam-Conc 和 FC-Siam-Diff 3 个简单 UNet 架构模型难以充分提取深层语义特征, 其 $F1$ 分数仅达到了 75.52%、80.25% 和 79.82%。SNUNet 和 STANet 通过引入密集连接机制和注意力机制可较好保留细节信息、减少信息损失, 其 $F1$ 分数达到了 86.45% 和 86.71%。BIT、ICIFNet 和 VcT 通过融合 CNN 局部特征表达优势与 Transformer 的长距离依赖建模能力, 变化检测精度进一步提高, 其 $F1$ 分数达到了 89.72%、88.90% 和 92.78%。此外, 基于纯 Transformer 结构的 ChangeFormer、FTNet 和 BiFA 可进一步提高变化检测精度, 其 $F1$ 分数达到了 89.88%、92.00% 和 92.34%。与之相比, 本文提出的 CTDFNet 模型可取得更优精度指标。

为比较各模型在不同场景下的性能差异, 选取代表性样本进行可视化分析。图 5 (c) 前两行显示了常规建筑物变化场景, 与对比方法相比, CTDFNet 模型能够较好地识别出目标基本轮廓, 且边缘定位精度较高。在图 5 (c) 第三、第四行展示的不规则建筑物变化场景中, 本文提出的 CTDFNet 模型在变化区域的完整性与边界清晰度方面均表现更优。此外, 在图 5 (c) 第五、第六行展示的大型建筑物变化场景中, CTDFNet 模型的建筑物边缘区域提取精度更高、细节保持能力较强, 取得了明显更优的结果。这是由于, 本文方法通过引入 CTM 实现了双时相特征的长程上下文信息建模与局部特征增强, 有效抑制了复杂背景噪声干扰, 同时提高了变化区的内部完整性与边缘定位精度, 且利用 ACEM 进一步提升了变化检测结果的可靠性。

4 讨论

4.1 模型复杂度

为全面评估模型的综合性能, 对 LEVIR-CD 数据集中各模型的精度 ($F1$ 分数) 与复杂度 (参数量、单张图像测试耗时) 进行了对比分析, 结果汇总于表 3。由表 3 可知, 3 个简单 UNet 模型 (FC-EF、FC-Siam-Conc、FC-Siam-Diff) 在参数量和推理速度上占优, 但其精度指标处于明显劣势。与之不同, 引入注意力机制的 STANet 与基于密集连接结构的 SNUNet 虽能显著提升检测精度, 但其参数量与推理耗时也相应大幅增加。此外,

由于 Transformer 结构基于自注意力机制单元构建，其模型参数量通常显著高于卷积网络。其中，基于纯 Transformer 结构的 ChangeFormer 和 FTNet 模型表现出较优的检测效果，但参数量显著增加，分别达到 41.01 M 和 164.45 M。BiFA 模型由于引入了轻量级 Transformer 结构，网络参数量显著降低。基于 CNN-Transformer 结构的 BIT、ICIFNet 和 VcT 模型可有效降低模型参数量，但其精度指标相较于 Transformer 方法也呈现出不同程度地降低。综合而言，CTDFFNet 在适度增加模型复杂度和测试时间的基础上显著提升了精度指标，从而在性能与效率之间实现了较好的平衡。

4.2 消融实验

为验证本文提出的不同模块对 CTDFFNet 模型性能的提升效果，在 3 个数据集上进行了一系列消融实验，结果如表 4 所示。可以看出，引入 MHCA 模块后，LEVIR-CD、CDD 和 WHU-CD 3 个数据集上的精度指标都得到不同程度提升。进一步添加 BF-FFN 模块后，3 个数据集的 F1 分数分别提升了

0.12%、0.59% 和 0.26%，IoU 分别提升了 0.24%、0.97% 和 0.43%。最后，加入 ACEM 模块进行特征增强后，本文提出的 CTDFFNet 在所有数据集上均取得了最优的综合精度。

表 3 LEVIR-CD 数据集中不同模型精度与复杂度对比
Table 3 Comparison of accuracy and complexity of different models on LEVIR-CD dataset

模型	F1/%	参数量/M	测试时间/s
FC-EF	84.94	1.35	34
FC-Siam-Conc	87.08	1.55	38
FC-Siam-Diff	86.85	1.35	38
STANet	88.75	16.93	64
SNUNet	88.79	12.03	245
ChangeFormer	90.40	41.01	140
FTNet	90.82	164.45	375
BiFA	90.72	5.58	91
BIT	89.43	3.49	63
ICIFNet	89.96	23.84	88
VcT	90.11	3.57	107
CTDFFNet	91.52	113.26	249

表 4 不同模块消融实验结果
Table 4 Ablation study results of different modules

模块			LEVIR-CD 数据集		CDD 数据集		WHU-CD 数据集	
MHCA	BF-FFN	ACEM	F1/%	IoU/%	F1/%	IoU/%	F1/%	IoU/%
×	×	×	90.57	82.79	92.70	86.38	91.48	84.27
√	×	×	91.08	83.60	94.26	89.19	92.35	85.80
√	√	×	91.20	83.84	94.85	90.16	92.61	86.23
√	√	√	91.52	84.37	95.83	92.00	93.02	86.95

图 6 显示了本文方法在 3 个数据集上的消融实验目视效果，Baseline 表示未添加任何改进模块的基础网络。可以看出，基础网络（第四行）在各类地物变化提取结果中均表现不佳。引入 MHCA 模块与 BF-FFN 模块后（第五、第六行），建筑物和道路变化区漏检像素显著减少、连续性明显改善。这主要得益于 MHCA 模块增强了长距离时空依赖建模能力，同时 BF-FFN 模块有效提升了局部细节表征能力，二者共同优化了变化区域的完整度与精确性。进一步地，由于较好地抑制了特征对齐中的语义与空间偏差，引入 ACEM 模块后（第七行）的检测效果达到最优，变化区域的内部均质性与边界准确性均得到有效提升。

此外，本文方法在 5 个不同编码器输出层均使

用 CTM 模块以实现多尺度多层次特征交互。为验证 CTM 模块个数对变化检测精度的影响，分别在不同编码器层添加 CTM 模块，并统计相应设置下 3 个数据集的变化检测精度，如表 5 所示。其中，CTM45 表示在第 4 层和第 5 层编码器添加 CTM 模块，CTM345 表示在第 3 层、第 4 层和第 5 层编码器添加 CTM 模块。可以看出，本文方法在 5 层编码器特征中均引入 CTM 模块，以提高多层次特征交互能力和全局特征表达能力，其效果优于仅在部分深层编码器特征中添加 CTM 模块（如 CTM45、CTM345）。这是由于 CTM 模块不仅能够提高全局特征表达能力，同时还能增强网络的多层次特征交互与融合能力，使得在不同编码器层均引入 CTM 模块实现了更好的特征表达增强与优化，从而有效提升了变化检测效果。

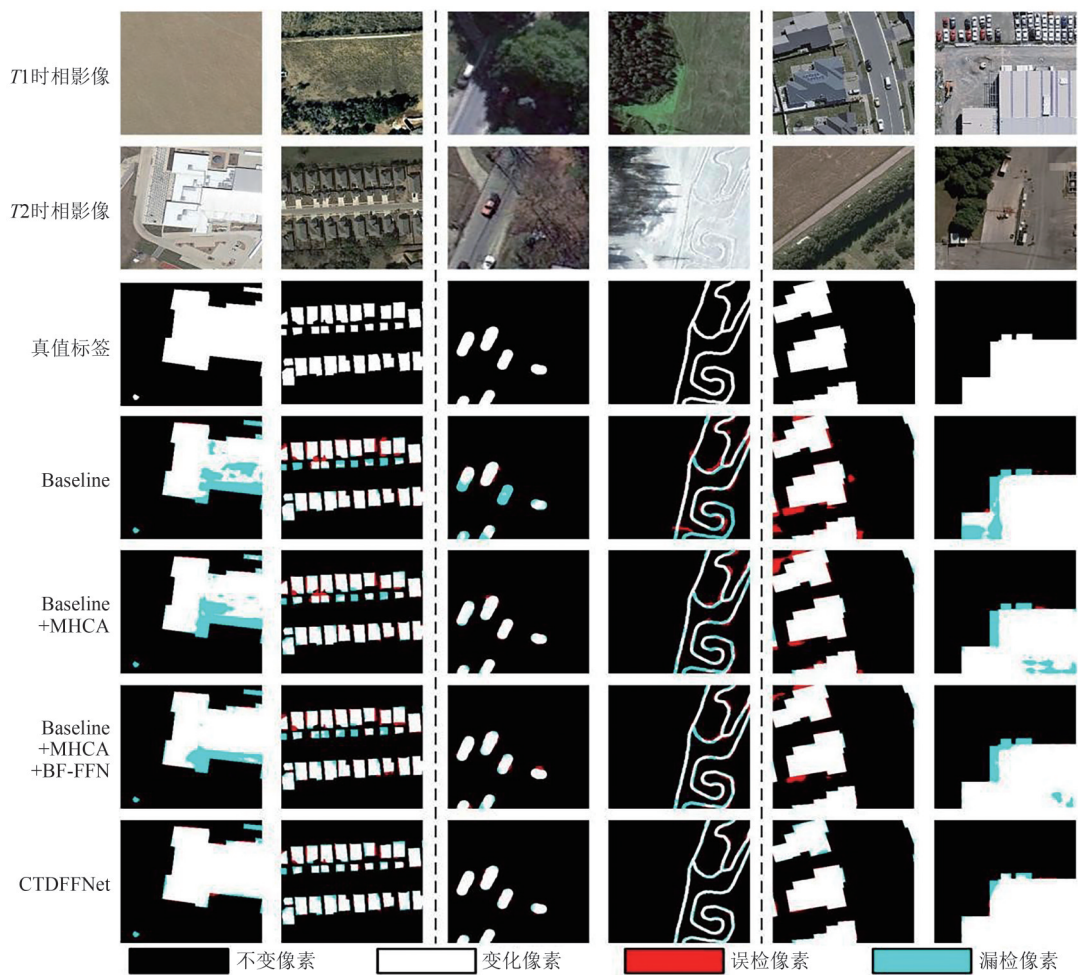


图6 消融实验目视结果
Fig. 6 Visual results of ablation study

表5 CTM个数对变化检测效果的影响
Table 5 The effect of the number of CTM on change detection performance

CTM组合	LEVIR-CD数据集		CDD数据集		WHU-CD数据集	
	F1	IoU	F1	IoU	F1	IoU
CTM45	90.18	82.11	94.69	89.91	92.32	85.74
CTM345	90.48	82.62	94.81	90.14	92.41	85.88
本文方法	91.52	84.37	95.83	92.00	93.02	86.95

特别地，为验证CTM模块中交叉注意力头个数对模型性能的影响，本文分别选择了不同的交叉注意力头个数 ($l = \{4, 8, 16\}$) 进行实验，并统计对应变化检测结果 $F1$ 分数和 IoU 指标，如表6所示。可以看出，当交叉注意力头数量设为4时，本文方法在LEVIR-CD和WHU-CD数据集上表现最优；而在CDD数据集上，最优数量为16。这是由于LEVIR-CD和WHU-CD数据集仅包含建筑物变化，使用少量交叉注意力头即可实现多时相特

征的有效交互与融合，而使用较多注意力头时由于不同方向注意力产生的信息冗余导致模型性能降低。相比之下，CDD数据集涵盖建筑物、道路、汽车等更多地物变化类型，特征交互更为复杂，增加交叉注意力头有助于捕捉更丰富的变化模式，模型性能可进一步提升。因此，为取得更优实验结果，对于LEVIR-CD和WHU-CD数据集设置交叉注意力头个数为4，而对于CDD数据集，设置交叉注意力头个数为16。

表6 CTM中交叉注意力头个数对变化检测结果影响
Table 6 The effect of the number of cross-attention heads of CTM on change detection performance

交叉注意力头个数	LEVIR-CD		CDD		WHU-CD	
	F1/%	IoU/%	F1/%	IoU/%	F1/%	IoU/%
4	91.52	84.37	95.37	91.22	93.02	86.95
8	91.16	83.77	95.33	91.14	91.43	84.29
16	91.35	84.09	95.83	92.00	92.06	85.34

4.3 训练过程可视化

为进一步验证 MHCA 模块、BF-FFN 模块和 ACEM 模块的有效性，本文绘制了 CTDFNet 中间层特征图可视化结果，如图 7 所示（图中 3 组 T1/T2 时相影像表示 3 组不同的双时相遥感图像对；小图左上角标注的“1”和“2”分别表示对应来自两个不同阶段的特征图可视化结果。）。结果显

示，MHCA 模块的响应集中于变化区域的整体结构，通过融合低层细节与高层语义特征来捕获长程依赖关系；BF-FFN 模块则通过建模局部特征关联，以提升细节特征的表达能力。此外，ACEM 模块通过对特征通道进行自适应校准，有效增强了变化区边界的定位精度，使最终特征图中的目标轮廓得到显著细化。

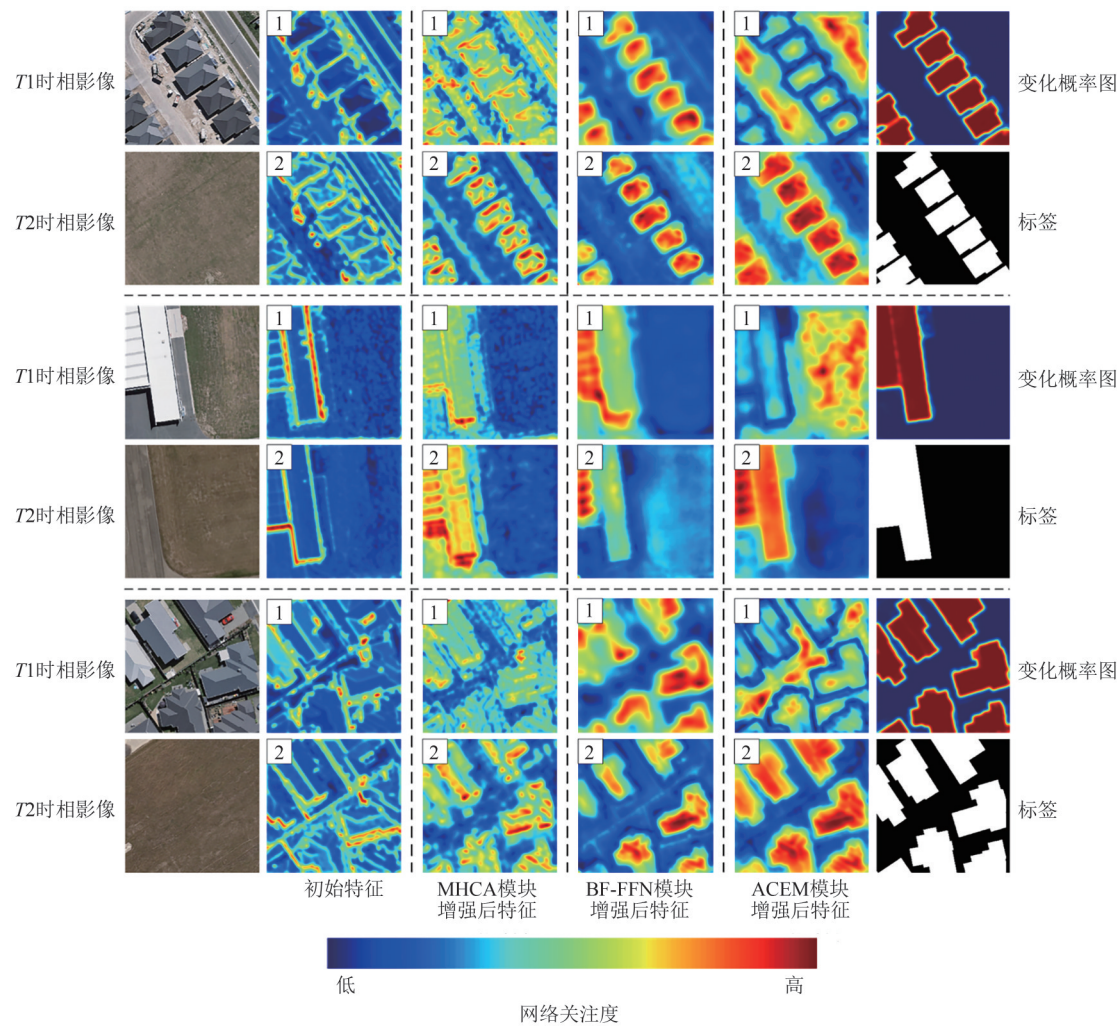


图 7 CTDFNet 中间层特征图可视化结果

Fig. 7 Visual results of intermediate feature maps in CTDFNet

5 结 论

针对卷积变化检测网络普遍存在的感受野有限、特征交互不足及长程依赖建模能力弱等问题，本文在引入 CTM 模块和 ACEM 模块的基础上构建了 CTDFNet 模型。在编码阶段，利用 CTM 模块实现双时相特征的多层级显示交互与融合，有效提升了模型在复杂场景下的检测性能。此外，解码部分利用密集连接结构显著增强了多维度特征利

用率，并强化了细节恢复能力。进一步地，ACEM 模块能够动态校准通道权重，融合解码器中不同层次的特征信息，从而有效缓解因语义鸿沟与空间错位导致的特征紊乱，提高网络对边缘结构的捕捉能力。实验结果表明，CTDFNet 在 3 个公开数据集上均取得最优精度指标，并在多种典型场景下展现出最优目视效果，验证了其方法的优越性。同时，通过消融实验验证了网络不同模块及策略的有效性。特别地，通过引入 CTM 和 ACEM

模块, 本文提出的CTDFFNet可有效建模长程上下文关联、增强局部特征表达, 实现了对复杂背景噪声(如阴影、光照)的有效抑制, 并同步优化了变化目标的内部完整性与边缘精度。这一设计使其在复杂地物交界区及城市建筑物密集区等高异质性区域的变化检测性能显著提升, 为应对城区复杂场景变化检测任务提供了可靠的技术支撑。然而, 本文方法仅考虑光学遥感影像信息, 缺少多源信息融入, 且网络计算复杂度较大, 对标记样本依赖性强。未来研究将结合多源数据融合方法及自监督与半监督学习算法, 进一步提高变化检测算法在复杂场景中的鲁棒性和普适性。

参考文献(References)

- Bandara W G C and Patel V M. 2022. A transformer-based siamese network for change detection//IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium. Kuala Lumpur, Malaysia: IEEE: 207-210 [DOI: 10.1109/IGARSS46834.2022.9883686]
- Benedek C, Shadaydeh M, Kato Z, Szirányi T and Zerubia J. 2015. Multilayer Markov random field models for change detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 107: 22-37 [DOI: 10.1016/j.isprsjprs.2015.02.006]
- Chen H, Qi Z P and Shi Z W. 2022. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 5607514 [DOI: 10.1109/TGRS.2021.3095166]
- Chen H and Shi Z W. 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10): 1662. [DOI: 10.3390/rs12101662]
- Chen J, Yuan Z Y, Peng J, Chen L, Huang H Z, Zhu J W, Liu Y and Li H F. 2021. DASNet: dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 1194-1206 [DOI: 10.1109/JSTARS.2020.3037893]
- Dalponte M, Jucker T, Liu S C, Frizzera L and Gianelle D. 2019. Characterizing forest carbon dynamics using multi-temporal lidar data. *Remote Sensing of Environment*, 224: 412-420 [DOI: 10.1016/j.rse.2019.02.018]
- Daudt R C, Le Saux B and Boulch A. 2018. Fully convolutional siamese networks for change detection//2018 25th IEEE International Conference on Image Processing (ICIP). Athens, Greece: IEEE: 4063-4067 [DOI: 10.1109/ICIP.2018.8451652]
- Deng J S, Wang K, Deng Y H and Qi G J. 2008. PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data. *International Journal of Remote Sensing*, 29(15/16): 4823-4838 [DOI: 10.1080/01431160801950162]
- Dong J, Zhao W F and Wang S. 2022a. Multiscale context aggregation network for building change detection using high resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19: 8022605 [DOI: 10.1109/LGRS.2021.3121094]
- Dong R M, Zhang L X and Fu H H. 2022b. RRSAN: Reference-based super-resolution for remote sensing image. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 5601117 [DOI: 10.1109/TGRS.2020.3046045]
- El Amin A M, Liu Q J and Wang Y H. 2017. Zoom out CNNs features for optical remote sensing change detection//2017 2nd International Conference on Image, Vision and Computing (ICIVC). Chengdu, China: IEEE: 812-817 [DOI: 10.1109/ICIVC.2017.7984667]
- Fang S, Li K Y, Shao J Y and Li Z. 2022. SNUNet-CD: a densely connected Siamese network for change detection of VHR images. *IEEE Geoscience and Remote Sensing Letters*, 19: 8007805 [DOI: 10.1109/LGRS.2021.3056416]
- Feng Y C, Jiang J W, Xu H H and Zheng J W. 2023. Building change detection using cross-temporal feature interaction network//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.1109/ICASSP49357.2023.10096120]
- Feng Y C, Xu H H, Jiang J W, Liu H and Zheng J W. 2022. ICIF-Net: intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 4410213 [DOI: 10.1109/TGRS.2022.3168331]
- He X Z, Wang Y, Sun Q L and Zhou F X. 2024. PFCF-Net: a network based on progressive feature interaction and cross-scale feature fusion for remote sensing change detection//2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea, Republic of: IEEE: 4330-4334 [DOI: 10.1109/ICASSP48485.2024.10448442]
- Ji S P, Wei S Q and Lu M. 2019. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1): 574-586 [DOI: 10.1109/TGRS.2018.2858817]
- Jiang B, Wang Z T, Wang X X, Zhang Z Y, Chen L, Wang X and Luo B. 2023. VcT: visual change transformer for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 2005214 [DOI: 10.1109/TGRS.2023.3327139]
- Lebedev M A, Vizilter Y V, Vygolov O V, Knyaz V A and Rubis A Y. 2018. Change detection in remote sensing images using conditional adversarial networks. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2(1682/1750): 565-571 [DOI: 10.5194/isprs-archives-XLII-2-565-2018]
- Liang Z H, Li X, Deng P, Sheng S and Jiang F Q. 2022. Remote sensing image change detection fusion method integrating multi-scale feature attention. *Acta Geodaetica et Cartographica Sinica*, 51(5): 668-676 (梁哲恒, 黎宵, 邓鹏, 盛森, 姜福泉. 2022. 融合多尺度特征注意力的遥感影像变化检测方法. *测绘学报*, 51(5): 668-676) [DOI: 10.11947/j.AGCS.2022.20200540]

- Liu S C, Du K C, Zheng Y J, Chen J, Du P J and Tong X H. 2023. Remote sensing change detection technology in the era of artificial intelligence: inheritance, development and challenges. *Journal of Remote Sensing* (in Chinese), 27(9): 1975-1987 (柳思聪, 都科丞, 郑永杰, 陈晋, 杜培军, 童小华. 2023. 人工智能时代的遥感变化检测技术: 继承、发展与挑战. *遥感学报*, 27(9): 1975-1987) [DOI: 10.11834/jrs.20222199]
- Liu T F, Gong M G, Lu D, Zhang Q F, Zheng H H, Jiang F L and Zhang M Y. 2022. Building change detection for VHR remote sensing images via local - global pyramid network and cross-task transfer learning strategy. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 4704817 [DOI: 10.1109/TGRS.2021.3130940]
- Liu W, Lin Y Y, Liu W J, Yu Y T and Li J. 2023. An attention-based multiscale transformer network for remote sensing image change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202: 599-609 [DOI: 10.1016/j.isprsjprs.2023.07.001]
- Liu Y, He X, Li D Y, Yue H and Wei J L. 2024. CBAM UNet++: attention mechanism to guide change detection studies of full-scale connected networks. *National Remote Sensing Bulletin*, 28(4): 1052-1065 (刘英, 何雪, 李单阳, 岳辉, 魏嘉莉. 2024. 注意力引导全尺度连接网络的高分辨率影像变化检测. *遥感学报*, 28(4): 1052-1065) [DOI: 10.11834/jrs.20221548]
- Liu Y Q, Chai Y and Lin K Z. 2024. Multi-scale feature fusion network with transformer for change detection of remote sensing images//2024 International Symposium on Autonomous Systems (ISAS). Chongqing, China: IEEE: 1-6 [DOI: 10.1109/ISAS61044.2024.10552528]
- Lv P Y, Zhong Y F, Zhao J and Zhang L P. 2018. Unsupervised change detection based on hybrid conditional random field model for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 56(7): 4002-4015 [DOI: 10.1109/TGRS.2018.2819367]
- Marin C, Bovolo F and Bruzzone L. 2015. Building change detection in multitemporal very high resolution SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5): 2664-2682 [DOI: 10.1109/TGRS.2014.2363548]
- Pekel J F, Cottam A, Gorelick N and Belward A S. 2016. High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633): 418-422 [DOI: 10.1038/nature20584]
- Peng D F, Zhang Y J and Guan H Y. 2019. End-to-end change detection for high resolution satellite images using improved UNet++. *Remote Sensing*, 11(11): 1382 [DOI: 10.3390/rs11111382]
- Song X P, Hansen M C, Stehman S V, Potapov P V, Tyukavina A, Vermote E F and Townshend J R. 2018. Global land change from 1982 to 2016. *Nature*, 560(7720): 639-643 [DOI: 10.1038/s41586-018-0411-9]
- Sui H G, Feng W Q, Li W Z, Sun K M and Xu C. 2018. Review of change detection methods for multi-temporal remote sensing imagery. *Geomatics and Information Science of Wuhan University*, 43(12): 1885-1898 (眭海刚, 冯文卿, 李文卓, 孙开敏, 徐川. 2018. 多时相遥感影像变化检测方法综述. *武汉大学学报(信息科学版)*, 43(12): 1885-1898) [DOI: 10.13203/j.whugis.20180251]
- Tewkesbury A P, Comber A J, Tate N J, Lamb A and Fisher P F. 2015. A critical synthesis of remotely sensed optical image change detection techniques. *Remote Sensing of Environment*, 160: 1-14. [DOI: 10.1016/j.rse.2015.01.006]
- Tong G F, Li Y, Ding W L and Yue X Y. 2015. Review of remote sensing image change detection. *Journal of Image and Graphics*, 20(12): 1561-1571 (佟国峰, 李勇, 丁伟利, 岳晓阳. 2015. 遥感影像变化检测算法综述. *中国图象图形学报*, 20(12): 1561-1571) [DOI: 10.11834/jig.20151201]
- Wang P, Fan E and Wang P. 2021. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Letters*, 141: 61-67 [DOI: 10.1016/j.patrec.2020.07.042]
- Wu C, Du B, Cui X H and Zhang L P. 2017. A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion. *Remote Sensing of Environment*, 199: 241-255 [DOI: 10.1016/j.rse.2017.07.009]
- Yan T Y, Wan Z F and Zhang P P. 2022. Fully transformer network for change detection of remote sensing images//Proceedings of the 16th Asian Conference on Computer Vision. Macao, China: Springer: 1691-1708 [DOI: 10.1007/978-3-031-26284-5_5]
- Zhang B, Chen Y X, Rong Y, Xiong S W and Lu X Q. 2023. MATNet: a combining multi-attention and transformer network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 5506015 [DOI: 10.1109/TGRS.2023.3254523]
- Zhang H, Gong M G, Zhang P Z, Su L Z and Shi J. 2016. Feature-level change detection using deep representation and feature change analysis for multispectral imagery. *IEEE Geoscience and Remote Sensing Letters*, 13(11): 1666-1670 [DOI: 10.1109/LGRS.2016.2601930]
- Zhang H T, Chen H, Zhou C Y, Chen K Y, Liu C Y, Zou Z X and Shi Z W. 2024. BiFA: remote sensing image change detection with bi-temporal feature alignment. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 5614317 [DOI: 10.1109/TGRS.2024.3376673]
- Zhang L P and Wu C. 2017. Advance and future development of change detection for multi-temporal remote sensing imagery. *Acta Geodaetica et Cartographica Sinica*, 46(10): 1447-1459 (张良培, 武辰. 2017. 多时相遥感影像变化检测的现状与展望. *测绘学报*, 46(10): 1447-1459) [DOI: 10.11947/j.AGCS.2017.20170340]
- Zhao Z Q, Zheng P, Xu S T and Wu X D. 2019. Object detection with deep learning: a review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11): 3212-3232 [DOI: 10.1109/TNNLS.2018.2876865]
- Zheng Z, Wan Y, Zhang Y J, Xiang S Z, Peng D F and Zhang B. 2021. CLNet: cross-layer convolutional neural network for change detection in optical remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175: 247-267 [DOI: 10.1016/j.isprsjprs.2021.03.005]

High-resolution optical remote sensing image change detection based on Cross-Transformer difference feature fusion

PENG Daifeng^{1,2}, ZHOU Dingwei¹, GUAN Haiyan¹

1. School of Remote Sensing and Surveying Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China;

2. Technology Innovation Center for Integration Applications in Remote Sensing and Navigation, Ministry of Natural Resources, Nanjing 210044, China

Abstract: Deep learning-based Change Detection (CD) methods, especially Convolutional Neural Networks (CNNs), are now widely used for high-resolution optical remote sensing imagery. Nevertheless, CNN-based architectures retain important structural limitations. Their restricted receptive fields make it difficult to capture global contextual cues that are essential for modeling long-range spatial dependencies and interpreting complex geographic scenes. In addition, insufficient interaction among multi-level features weakens the integration of high-level semantic representations with low-level spatial details, thereby limiting the discrimination of subtle and irregular changes.

To overcome these limitations, this paper presents the Cross-Transformer-based Difference Feature Fusion Network (CTDFFNet), a unified CD framework that combines global context perception, cross-level feature interaction, and adaptive feature refinement. CTDFFNet follows an encoder-decoder architecture. In the encoder, a ResNet18 backbone pre-trained on ImageNet extracts multi-scale local feature representations from bi-temporal remote sensing images. To compensate for the limited receptive field of standard convolutions, a Cross-Transformer module is designed to learn discriminative difference features. Through a hierarchical cross-attention mechanism, the module captures long-range spatial dependencies and strengthens interactions across feature levels, thereby improving sensitivity to changes under complex conditions, including illumination variation, shadow effects, and seasonal differences. In the decoder, dense connections enable each layer to receive information from all preceding layers, which promotes feature reuse, improves the use of hierarchical information, alleviates gradient vanishing, and supports more stable training. An Adaptive Channel Enhancement Module (ACEM) is further incorporated to dynamically recalibrate channel-wise feature responses by learning attention weights, emphasizing change-salient features while suppressing irrelevant background noise. A Sigmoid layer is then used to produce the final change map.

Extensive experiments on three publicly available CD datasets, namely LEVIR-CD, CDD, and WHU-CD, were conducted to evaluate the effectiveness and generalization capability of CTDFFNet. Quantitative results demonstrate that CTDFFNet consistently outperforms state-of-the-art methods on all three datasets, achieving the highest F1 scores of 91.52%, 95.88%, and 93.02% on LEVIR-CD, CDD, and WHU-CD, respectively. These gains confirm the effectiveness of incorporating Transformer-based mechanisms into CD frameworks. Visual comparisons further indicate that CTDFFNet effectively suppresses interference caused by complex backgrounds and environmental variations, producing change maps with substantially fewer false alarms and missed detections while maintaining high precision in delineating changed objects at different scales.

In summary, CTDFFNet addresses key limitations of CNN-based CD frameworks through three main innovations: (1) the Cross-Transformer module captures long-range dependencies and promotes cross-level feature interaction, enhancing sensitivity to complex scene changes; (2) the densely connected decoder improves multi-scale feature use and alleviates gradient vanishing, contributing to stable training; and (3) ACEM dynamically emphasizes change-salient features and suppresses background interference. Experiments on multiple datasets demonstrate that CTDFFNet outperforms existing methods and exhibits strong generalization and robustness under challenging conditions. These findings support its potential for a wide range of remote sensing CD applications. Future work will extend the framework to multi-modal CD scenarios and investigate label-efficient learning strategies to reduce dependence on large-scale annotated data.

Key words: change detection, deep learning, Cross-Transformer, cross attention mechanism, adaptive channel enhancement

Supported by National Natural Science Foundation of China (No. 42371449, 41801386)