中国科学: 信息科学 2022 年 第 52 卷 第 11 期: 2058–2074
SCIENTIA SINICA Informationis
论文

基于全局结构差异与局部注意力的变化检测

梅杰, 程明明*

南开大学计算机学院天津市媒体计算技术工程研究中心, 天津 300350
* 通信作者. E-mail: cmm@nankai.edu.cn

收稿日期: 2022–03–01; 修回日期: 2022–06–04; 接受日期: 2022–06–27; 网络出版日期: 2022–11–09

科技创新 2030 "新一代人工智能" 重大项目 (批准号: 2018AAA0100400) 和国家自然科学基金优秀青年科学基金 (批准号: 61922046) 资助项目

摘要 检测由自然灾害造成的不同变化, 对于有效地指导人道主义援助和灾难响应行动来说至关重要. 但是灾害发生的地区通常面积大、地面环境复杂, 导致检测其变化具有较大的挑战性. 现有的评估方法通常依靠人工来进行判别, 不适用于多种灾害的检测. 本文提出了一种新颖的变化检测模型 (change transformer, CHTR), 基于双时序遥感图像来同时进行建筑分割和多级变化检测两个任务. 本文结合卷积神经网络擅长学习局部细节特征和 Transformer 可以建模长程依赖关系的优势, 采用混合卷积神经网络和 Transformer 的架构作为编码器. 考虑到自然灾害通常会对复杂环境中的建筑物造成不同程度的破坏, 本文提出了一种全局差异模块, 以捕获全局变化模式, 提高对双时序图像之间变化的整体认识. 进一步设计了一种局部门控注意力模块, 以学习多级别变化之间的局部依赖性, 增强对不同变化的判别能力. 在目前最大的建筑物损毁评估数据集 (xBD) 上进行的大量实验表明, 本文提出的方法在建筑分割和变化检测两个任务上都取得了更好的结果.

关键词 建筑物分割, 变化检测, 遥感图像, 全局和局部结构, Transformer

1 引言

对土地覆盖的变化检测是区域和全球环境监测的一项重要技术. 遥感图像能够提供大规模和多时序信息, 促进了多种应用场景下的变化检测研究, 例如, 环境监测 [1]、资源管理 [2]、城市化评估 [3] 和灾害评估 [4] 等. 灾害检测可以看作是变化检测的一个子任务, 它指的是当自然灾害发生时, 及时地评估灾害损毁的位置和严重程度, 这对于开展有效的灾难响应活动至关重要. 由于地面评估存在风险且难以获得, 遥感图像已成为灾害变化检测和损毁评估的有力工具. 然而, 目前的评估流程主要是依靠人工分析员观察双时序遥感图像并识别损坏区域, 需要较多的人力和时间, 不适合于大尺度的区域. 因此, 能够减轻人力工作并加快评估过程的自动化变化检测方法, 近年来受到了更多的关注.

引用格式: 梅杰, 程明明. 基于全局结构差异与局部注意力的变化检测. 中国科学: 信息科学, 2022, 52: 2058–2074, doi: 10.1360/SSI-2021-0384
Mei J, Cheng M-M. Damage assessment with global differences and local attention (in Chinese). Sci Sin Inform, 2022, 52: 2058–2074, doi: 10.1360/SSI-2021-0384

图 1 (网络版彩图) 自然灾害后的多级别变化示意图. (a) 灾害前图像; (b) 灾害后图像; (c) 建筑真值; (d) 多级别变化真值. 其中的 4 种颜色代表 4 种损毁等级: 绿色、蓝色、橙红色和红色分别表示无损毁、轻微损毁、严重损毁和完全损毁

Figure 1 (Color online) Illustration of the multi-level changes after a natural disaster. (a) Pre-disaster image; (b) post-disaster image; (c) ground truth of buildings; (d) ground truth of multi-level changes. Four colors in (d) represent four damage scales: green, blue, orange, and red denote no damage, minor damage, major damage, and destroyed, respectively

早期的变化检测研究通常基于两个不同时期的图像, 使用感知像素差异的模型[5,6]. 这些方法通常是为特定数据设计的, 在处理其他区域的图像时效果较差[7]. 近年来, 基于卷积神经网络 (convolutional neural network, CNN) 的模型被大量提出并有效地应用于语义分割的任务中[8~10]. 一些研究[11,12] 也采用卷积神经网络, 基于分割来进行灾害的变化检测和损毁评估, 但这些方法通常只用于检测由单个灾害引起的变化. Gupta 等[13] 发布了一个名为 xBD 的大规模数据集, 其中包含来自 19 种自然灾害和不同损毁等级的遥感图像. 基于数据集 xBD, 文献 [13] 进一步采用 ResNet-50[14] 进行灾害损毁等级的分类, 并采用 U-Net[15] 模型来完成建筑分割的任务. 然而, 这个方法利用两个独立的模型来分别完成建筑物分割和变化检测任务, 使其无法从多任务学习中受益, 并且通常需要更多的时间和步骤来训练.

由于卷积运算的内在局部性, 基于 CNN 的方法缺乏构建长程依赖关系的能力. 为了克服这个限制, 一些研究[16,17] 在图像识别任务中为 CNN 构建了自注意力机制. Wang 等[18] 提出了非局部 (non-local) 的操作来捕获视频分类任务中任何位置之间的长程依赖关系. 另一方面, 为序列到序列预测任务而提出的 Transformer, 在学习长程依赖方面展示了卓越的能力. 最近, Transformer 结构在许多计算机视觉任务中得到了探索, 并获得了较好的性能[19,20].

本文依托具体的灾害检测任务对变化检测进行研究, 并提出了一种新颖的变化检测方法 (change transformer, CHTR), 基于双时序遥感图像同时进行建筑物分割和多级别变化检测. 本文结合 CNN 擅长学习局部细节特征和 Transformer 可以建模长程依赖关系的优势, 采用混合 CNN 和 Transformer 的架构作为编码器. 同时引入渐进式上采样策略作为解码器, 来生成建筑物分割和变化检测的输出. 一张高分辨率遥感图像通常覆盖较大的面积, 其中包含的建筑物, 在自然灾害发生后会受到不同程度的损毁, 如图 1 所示. 为了提高对多级别变化的理解, 本文提出全局差异 (global difference, GD) 模块和局部门控注意力 (local gated attention, LGA) 模块. 由于图像块的小尺寸会限制模型学习不同变化之间的长程依赖关系, 本文设计了全局差异模块来缓解这个问题并学习全局变化模式. 并且提出了局部门控注意力模块来获取局部变化差异并增强对双时序图像之间多级别变化的辨别力. 本文在大规模建筑损毁评估数据集 xBD 上进行了大量实验, 结果验证了所提出方法 CHTR 的有效性.

本文的主要贡献可总结如下.

• 提出了一种新颖的变化检测方法, 基于双时序遥感图像同时进行建筑物分割和多级变化检测任务. 其在大规模建筑损毁评估数据集 xBD 上取得了较好的结果.

• 开发了一种全局差异模块, 来捕捉全局变化模式并提高对双时序遥感图像之间变化的整体认识.

• 设计了一种局部门控注意力模块, 其通过探索多级别变化之间的局部依赖性提高识别不同变化的能力.

2 相关研究现状

2.1 变化检测

基于遥感图像的变化检测对于监测区域和全球环境来说越来越重要, 近年来吸引了很多的相关研究 [3,4]. 一些经典的分类算法, 例如, 多层感知器 [21]、极限学习机 [22]、支持向量机 [23], 以及一些无监督的方法, 如, 马尔可夫 (Markov) 随机场 [24]、变化向量分析 [25] 等, 都被用于变化检测任务. Im 等 [6] 基于高空间分辨率的多时序遥感图像, 提出了一种三通道邻域相关的方法. Tan 等 [26] 利用多尺度不确定性分析和多个分类器, 设计了一种基于对象的变化检测方法. 然而, 这些传统的方法通常应用基于像素差异的模型, 可能会产生遗漏的错误.

近年来, 深度学习在多个领域都取得了重大的进展 [27~30], 一些基于卷积神经网络的方法也被提出并用于变化检测任务. Daudt 等 [31] 提出了一种迭代学习方法来训练一个全卷积神经网络, 用于从噪声数据中检测变化. Papadomanolaki 等 [32] 结合循环神经网络 (recurrent neural network, RNN) 和全卷积神经网络, 使用多时序的高分辨率数据来进行城市的变化检测. 孪生神经网络可以评估两幅图像之间的相似性, 这使其在变化检测任务中非常有效. Liu 等 [33] 设计了一个深度卷积耦合网络, 用于从两幅图像中检测变化. Daudt 等 [34] 基于普通图像和多光谱图像, 提出了两个孪生扩展模块, 并将其融合到全卷积神经网络中来完成变化检测. Chen 等 [35] 结合 CNN 和 RNN 的优势, 提出了一种深度孪生卷积循环神经网络进行变化检测, 其可以用于同质和异质的高分辨率遥感图像. Du 等 [36] 利用两个对称的神经网络来提取双时序图像的特征, 之后采用慢特征分析 (slow feature analysis, SFA) 来突出转换特征中变化的部分. Wu 等 [37] 基于核主成分分析卷积, 提出了一种无监督的深度孪生卷积网络进行二类和多类别的变化检测. 然而, 孪生神经网络需要两个网络来处理不同时序的图像, 这会引入更多的参数量.

当自然灾害发生时, 快速检测变化信息以进行有效的灾害响应至关重要. 因此, 最近的变化检测研究更加关注由自然灾害引起的变化. Ji 等 [11] 采用卷积神经网络, 利用灾害后的遥感图像来识别倒塌的建筑物. Duarte 等 [38] 提出了一个具有空洞卷积和残差连接的 CNN 框架, 基于遥感图像对建筑物的损毁进行分类. Rudner 等 [12] 利用一个具有编码器 – 解码器架构的卷积神经网络, 将多分辨率、多传感器和多时序的遥感图像融合起来, 以得到被淹没建筑物的分割结果. 这些方法通常只用于检测由单一灾害引起的变化, 无法应对现实中复杂的自然灾害. 文献 [13] 提出了一个名为 xBD 的大规模数据集, 用于建筑物的损毁评估和变化检测, 其提供了来自 19 种不同自然灾害和 4 种损毁等级的遥感图像. 基于 xBD 数据集, Gupta 等 [13] 设计了一个基准模型, 其采用 ResNet-50 [14] 进行损毁等级的分类, 同时采用 U-Net 模型 [15] 用于建筑物的分割. Weber 等 [39] 使用 Mask R-CNN [40] 架构, 通过加入特征金字塔网络模块 [41] 和语义分割模块来进行建筑损毁评估. Bai 等 [42] 介绍了一种用于建筑物损毁评估的并发学习注意力网络. 然而, 这些方法通常将建筑物分割和损毁评估分为两个独立的阶段, 这使得它们无法从多任务学习中受益, 并且需要复杂的步骤来训练.

2.2 自注意力和 Transformer

针对机器翻译而提出的自注意力和 Transformer, 在许多自然语言处理任务中都取得了较好的性能 [43,44]. 考虑到 Transformer 可以学习到输入项之间长程依赖关系的能力, 最近有一些研究将 Transformer 结构用于计算机视觉任务中 [19,20,45].

具体来说, 非局部操作 [18] 通过计算任意两个位置之间的关联获取长程依赖关系. Criss-cross 网络 [16] 基于纵横路径从全图依赖中学习上下文信息, 更加高效和有效. Mei 等 [17] 将自注意力应用于肺结节检测并提出了基于分组切片的非局部模块. Wang 等 [46] 将二维自注意力分解为两个一维自注意力, 并提出了一种位置敏感的轴向注意力模块. 文献 [47] 考虑了两种自注意力的变体, 分别是成对自注意力和成块自注意力. Vision Transformer (ViT) [48] 直接使用 Transformer 处理图像块的序列, 来进行图像的分类. Touvron 等 [49] 利用基于输入项的蒸馏进一步扩展了 ViT. 在目标检测任务中, Carion 等 [19] 利用 Transformer 来推理全局图像上下文和目标的关系, 并直接输出最终的预测结果, 无需空间锚点或非极大值抑制. 文献 [50] 进一步提出了一个可变形注意力模块, 该模块关注一小组采样位置, 从而实现了快速收敛和更好的性能. 至于语义分割, Zheng 等 [20] 只应用 Transformer 将图像编码为图像块的序列, 结合一个简单的解码器来完成语义分割. 一些研究将自注意力和 Transformer 应用于变化检测的任务中. Zhang 等 [51] 利用通道注意力和空间注意力模块对双时序图像的特征从通道维度和空间维度上进行优化, 但是其缺乏长程建模的能力. Chen 等 [52] 利用自注意力机制来定位变化区域同时获取更具有判别性的特征, 但是其中的矩阵运算会带来较大的计算复杂度. Chen 等 [53] 将 Transformer 引入到变化检测任务中来更好地建模双时序图像的上下文信息, 然而这种方法将图像特征划分成特征块后再输入到 Transformer 中, 会带来特征块之间全局结构信息的损失. 本文尝试探索将整个特征图和划分后的特征块用于 Transformer, 其能够学习双时序图像之间全局和局部的结构信息以及变化模式, 增强对多级别变化的判别能力.

3 基于全局结构差异与局部注意力的变化检测方法

考虑到复杂的现实环境和灾害的多样性, 灾前和灾后的图像之间通常存在多种级别的变化. 为了学习不同变化之间的关系, 本文提出了一种新颖的变化检测方法 CHTR, 基于双时序遥感图像同时进行建筑物分割和多级变化检测, 如图 2 所示. 其中本文设计了一个全局差异 (GD) 模块来获取全局变化信息. 同时进一步提出了一个局部门控注意力 (LGA) 模块, 以学习多级别变化之间的局部依赖关系.

3.1 网络结构

本文结合 CNN 擅长学习局部细节特征和 Transformer 可以建模长程依赖关系的优势, 采用混合 CNN 和 Transformer 的架构作为编码器, 将 CNN 输出的降采样后的特征图作为 Transformer 的输入. 这种结构能够帮助学习全局变化模式, 提高对图像的整体认识. 对于网络输入的双时序遥感图像, 即一对灾前和灾后的图像, ResNet-50 [14] 首先被用作特征提取器来生成两个特征图. 基于双时序特征图, 本文设计了一个全局差异模块和一个局部门控注意力模块, 来获取不同时序图像之间的全局和局部变化信息.

本文采用渐进式上采样策略作为解码器, 其中包含 4 个上采样块来解码特征以达到输入图像的分辨率. 每个解码器块依次包含一个 3×3 卷积层、一个 BatchNorm 层、一个 ReLU 层和一个 $2\times/4\times$ 上采样算子. 在最后一个解码器块后有两个分支: 建筑分割分支和变化检测分支. 每个分支都包含一个 3×3 的卷积层, 分别输出通道数为 1 和通道数为 5 的结果. 来自全局差异模块和局部门控注意力

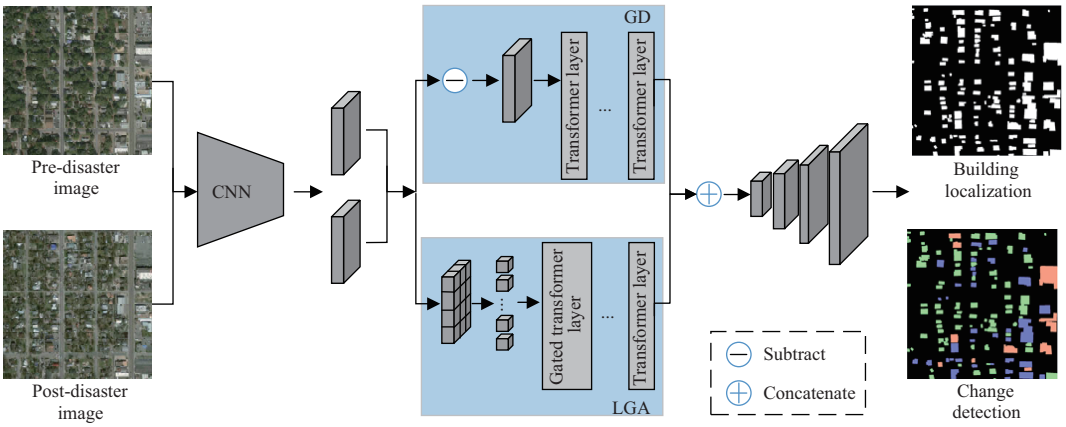

图 2 (网络版彩图) 本文所提出方法的总体架构图. 双时序遥感图像被输入到混合 CNN 和 Transformer 架构的编码器以提取特征, 然后采用渐进式上采样解码器来同时完成建筑物分割和多级别变化检测. 对于编码器中的 Transformer, 本文提出了两个模块: 全局差异模块和局部门控注意力模块, 以获取全局和局部的变化模式

Figure 2 (Color online) Overall architecture of the proposed change transformer (CHTR). The dual-temporal satellite images are fed to an encoder with CNN-Transformer architecture to extract features, which are followed by a progressive upsampling decoder to complete building localization and multi-level change detection simultaneously. For the transformer layers in the encoder, we propose two modules: global difference (GD) module and local gated attention (LGA) module, to capture the global and local change patterns

模块的特征在融合后被输入到解码器中, 从而输出建筑分割和多级别变化检测的结果. 此外, 本文在建筑分割任务中采用 Combo Loss [54], 在变化检测任务中采用加权交叉熵损失 (cross-entropy loss), 其中 Combo Loss 的定义为

$$L_{\text{Combo}} = \lambda_{c1} L_{\text{Dice}} + \lambda_{c2} L_{\text{Focal}}, \tag{1}$$

其中 λ_{c1} 和 λ_{c2} 是平衡系数. Combo Loss 是 Dice Loss [55] 和 Focal Loss [56] 的加权和, 其定义如下:

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N}(y_i \hat{y}_i)}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} \hat{y}_i{}^2}, \tag{2}$$

$$L_{\text{Focal}} = -\frac{1}{N}\sum_{i=1}^{N} \alpha y_i (1-\hat{y}_i)^{\gamma} \log(\hat{y}_i) - (1-\alpha)\hat{y}_i{}^{\gamma}(1-y_i)\log(1-\hat{y}_i), \tag{3}$$

其中 N 为特征图中像素的数量. y_i 是指示位置 i 处像素为建筑物或背景的真值, $\hat{y}_i$ 为本文方法对建筑分割任务相应的预测概率. α 是一个权重因子, γ 是一个可调节参数, 用于处理类别不平衡的问题.

加权交叉熵损失的定义为

$$L_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{n} w_c y_i^c \log(\hat{y_i^c}), \tag{4}$$

其中, n 为变化的等级数量, w_c 为每个类别的缩放权重. y_i^c 是指示位置 i 处像素的变化等级为 c 的真值, $\hat{y_i^c}$ 为本文方法预测 i 处像素的变化等级为 c 的概率.

本文方法的总体损失函数定义为

$$L_{\text{CHTR}} = \lambda_1 L_{\text{Combo}} + \lambda_2 L_{\text{CE}}, \tag{5}$$

其中 λ_1 和 λ_2 是用于平衡两个损失函数的常数.

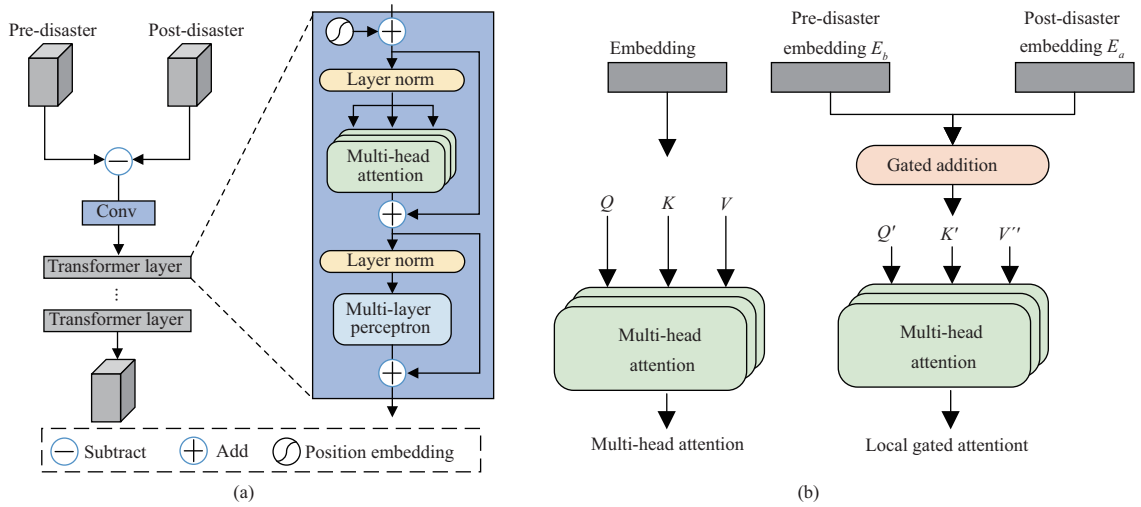

图 3 (网络版彩图) (a) 全局差异模块示意图; (b) 原本的多头自注意力和本文所提出的局部门控注意力的对比. 利用门控加和操作, 来自灾害前后的图像块特征可以自适应地融合

Figure 3 (Color online) (a) Illustration of the global difference module (GD); (b) the original multi-head attention vs. our local gated attention (LGA). With the gated addition operation, the pre-disaster and post-disaster patch embeddings are fused adaptively

3.2 全局差异模块

在图像块上训练 Transformer 有助于提高训练速度. 然而, 只利用图像块进行训练, 对于遥感图像的多级别变化检测任务是不够的. 一张高分辨率遥感图像通常覆盖较大的面积, 其中包含自然灾害后不同损毁程度的建筑物. 一个图像块的尺寸要小于整幅图像, 这限制了模型学习全局变化模式和不同变化之间依赖关系的能力. 为了学习全局变化模式并提高对图像的整体理解, 本文提出了一个全局差异 (GD) 模块.

全局差异模块的架构如图 3(a) 所示. 设置 $X_b, X_a \in \mathbb{R}^{H\times W\times C}$ 表示 CNN 输出的灾害前后图像的一对特征, 其被用于计算特征差异:

$$X_g = X_a - X_b, \tag{6}$$

其中 X_g 表示双时序特征之间的全局差异. 为了提高计算效率, X_g 被输入到一个卷积层来减少空间维度和特征通道数, 并获得其输出特征 $X'_g \in \mathbb{R}^{H'\times W'\times C'}$. 之后 X'_g 被输入到两层 Transformer 中, 其中每层 Transformer 包含层归一化 (layer normalization, LN)、多头自注意力 (multi-head self-attention, MSA) 和多层感知机 (multi-layer perceptron, MLP).

由于 Transformer 层需要输入一个序列, X'_g 被展平并输入一个可训练的线性层以获得特征嵌入序列 $e_g \in \mathbb{R}^{L\times C_g}$, 其中 L 是序列长度, C_g 是隐藏通道维数. 为了编码缺失的图像块空间信息, 本文学习了特定的位置序列 p_g, 并将其添加到特征序列 e_g 中以形成最终的序列输入 $E_g = e_g + p_g$. 这样, 虽然 Transformer 的自注意力本质是无序的, 但是空间信息还是可以得到保留的.

在单个自注意力模块中, 自注意力的输入 $E_g \in \mathbb{R}^{L\times C_g}$ 被线性变换为 3 部分, 即 Query $Q \in \mathbb{R}^{L\times d_k}$, Key $K \in \mathbb{R}^{L\times d_k}$, 以及 Value $V \in \mathbb{R}^{L\times d_v}$, d_k, d_v 分别是 Query (Key) 和 Value 的维数. 线性变换定义为

$$\text{Query } Q = E_g w_Q, \quad \text{Key } K = E_g w_K, \quad \text{Value } V = E_g w_V, \tag{7}$$

其中 w_Q, w_K 和 w_V 是 3 个线性变换函数的权重矩阵. Q, K, V 都源于输入特征 E_g 本身, Q 和 K 是

用于计算注意力权重的特征向量, V 的引入是为了保留输入的特征向量.

然后将 Scaled Dot-Product Attention 应用于 Q, K 和 V:

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}\right)V, \tag{8}$$

其中, "softmax" 是指 softmax 函数, $\sqrt{d_k}$ 是一个缩放因子. 式 (8) 中 $\text{softmax}(QK^{\text{T}}/\sqrt{d_k})$ 用来计算注意力权重, Q 和 K 两个不同特征向量的引入可以提高注意力权重矩阵的泛化能力, 之后再利用注意力权重将 V 映射到一个新的空间.

作为包含多个独立自注意力操作的扩展, 多头自注意力将 Query, Key 和 Value 进行多次拆分, 并行计算上述注意力函数, 然后将所有头的输出拼接起来并投射到最终的输出. 多头自注意力允许模型联合在不同位置的不同表征子空间的信息 [43], 其公式定义为

$$\text{MultiHead}(Q,K,V)=\text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O, \tag{9}$$

$$\text{head}_i=\text{Attention}(QW_i^Q,KW_i^K,VW_i^V), \tag{10}$$

其中, h 为多头自注意力的头数. W_i^Q, W_i^K 和 W_i^V 为第 i 个头的 3 个权重矩阵, W^O 为将拼接结果投射到输出的转换矩阵.

多头自注意力 MSA 的输出在经过层归一化后, 被输入到多层感知机中进行特征转换. 此外, 如图 3(a) 所示, 多头自注意力和多层感知机的输出中都应用了残差连接 [14]. 为了更加清晰地表明 Transformer 的内部结构, 本文将第 l 层 Transformer 的输出定义为

$$Z_l=\text{MLP}(\text{LN}(Z_l'))+Z_l', \tag{11}$$

其中,

$$Z_l'=\text{MSA}(\text{LN}(Z_{l-1}))+Z_{l-1}, \tag{12}$$

$\{Z_1,Z_2,\ldots,Z_l\}$ 表示 Transformer 层的特征. $\text{LN}(\cdot)$ 是层归一化操作, 其被用于实现更快的收敛和稳定的训练.

3.3 局部门控注意力模块

与灾前的遥感图像相比, 自然灾害通常会对灾后图像中的建筑物造成不同程度的损毁. 评估损毁的严重程度是开展人道主义援助和灾难响应的先决条件, 而了解双时序图像之间的局部依赖性有助于识别不同的损毁程度. 因此, 本文设计了一个局部门控注意力 (LGA) 模块来获取不同时序成对图像之间多级别变化的局部依赖性.

对于成对的灾前和灾后特征 $X_b,X_a\in\mathbb{R}^{H\times W\times C}$, 首先将它们分别划分成 $\frac{HW}{4^2}$ 个特征块. 对于获得的灾前和灾后特征的两组特征块, 将其展平并输入到两个线性层以获取特征序列. 两组特征序列与位置序列相加以生成最终的灾前序列 E_b 和灾后序列 E_a. 然后 E_b 和 E_a 被输入到一层 Gated Transformer 和一层 Transformer 中, 如图 2 所示.

在 Gated Transformer 层中, 参考文献 [57] 中的门控轴向注意力, 本文设计了一个局部门控注意力来代替原始的多头自注意力, 如图 3(b) 所示. 输入的 E_b 和 E_a 被线性变换, 以生成灾前特征序列的 Q_b, K_b, V_b 和灾后特征序列的 Q_a, K_a, V_a. LGA 利用了门控加法来融合 Q_b, K_b, V_b 和 Q_a, K_a, V_a, 然后将它们输入到多头自注意力中. 与式 (8) 不同, 本文的局部门控注意力被定义为

$$\text{LGA}(Q',K',V')=\text{softmax}\left(\frac{Q'K'^{\text{T}}}{\sqrt{d_k}}\right)V', \tag{13}$$

其中, $Q' = G_b^q Q_b + G_a^q Q_a$, $K' = G_b^k K_b + G_a^k K_a$, $V' = G_b^v V_b + G_a^v V_a$. G_b^q, G_a^q, G_b^k, G_a^k, G_b^v, G_a^v 是可学习的参数, 这些参数构建了一种门控机制, 用于控制学习到的灾前和灾后特征块之间的相应特征序列. 如果一个时序的特征序列被准确地学习并且对识别多级别变化更有利, 门控机制将为其分配更高的权重, 同时给另一个时序的特征序列分配较低的权重. 利用这种门控机制, 局部门控注意力模块能够关注灾前和灾后特征块中对变化敏感的特征, 进而将变化敏感的特征块输入到后续的多头自注意力中. 利用自注意力机制, 可以探索相邻变化敏感特征块之间的局部依赖性, 进而学习到双时序图像之间的局部变化差异, 并细化所识别出的多级别变化.

4 数据集和评估指标

4.1 数据集

本文的实验中采用了 xBD 数据集 [13], 它是目前最大的建筑物损毁评估数据集. 其中包含了 45362 平方公里的地理区域、850736 个标注的建筑物以及来自 19 种不同自然灾害的 22068 张遥感图像 (即 11034 对灾前和灾后的图像), 例如洪水、地震、野火和火山爆发等. 该数据集中图像的像素分辨率为 1024×1024, 空间分辨率为 0.3 m. 其介绍了四级损毁等级来评估多种灾害发生后的建筑物损毁, 包括无损毁、轻微损毁、严重损毁和完全损毁. xBD 数据集被划分为 4 个部分: 训练集、测试集、Holdout 和 Tier3. 本文选取包含 2799 对遥感图像的训练集和包含 933 对遥感图像的测试集, 用于本文实验的训练和测试.

4.2 评估指标

为了评估所提出方法在多级别变化检测任务中的性能, 本文使用了 xView2 挑战中的评测指标 [13]. 其是建筑物定分割 $F1$ 分数与多级别损毁检测 $F1$ 分数的加权平均值, 定义为

$$S_{\text{xView2}} = 0.3F1_{\text{loc}} + 0.7F1_{\text{damage}}, \tag{14}$$

$$F1_{\text{damage}} = \frac{n}{\frac{1}{F1_{\text{cls1}}} + \cdots + \frac{1}{F1_{\text{cls}\,n}}}, \tag{15}$$

其中 $F1_{\text{loc}}$ 是建筑物分割的 $F1$ 分数, 它评估了灾前图像中地面真值与预测值之间的一致性, $F1_{\text{damage}}$ 是变化检测的 $F1$ 分数, 它评估了灾后图像中像素的预测值与地面真值之间的一致性, $F1_{\text{cls1}}, \ldots, F1_{\text{cls}\,n}$ 表示 n 种损毁等级的多级变化检测 $F1$ 分数. 由于 xBD 数据集中的损毁等级分布严重不平衡, 并且评测指标 S_{xView2} 会惩罚对具有过多建筑物损毁等级的过度拟合, 因此这是一个具有挑战性的指标.

5 实验结果

本节将介绍本方法同现有变化检测方法的比较, 同时也会对本方法中两个模块的重要性进行详细分析.

5.1 实验配置

在本文所提出的变化检测方法 CHTR 中, Batch Size 设置为 16, 采用随机梯度下降 (stochastic gradient descent, SGD) 优化器. Momentum 和 Weight Decay Coefficient 分别设置为 0.9 和 5×10^{-4}, 初始学习率设置为 0.02, 训练的 Epoch 为 150. 此外, 在训练中本文采用了 “poly” 学习率策略, 通过

表 1　本文方法 CHTR 与目前较好的变化检测方法在 xBD 数据集上的定量比较 (%)

Table 1 Quantitative comparison of our proposed CHTR with some state-of-the-art change detection methods on the xBD dataset (%)[a)]

Method	$F1$ score	Localization $F1$	Damage $F1$	Undamaged $F1$	Minor $F1$	Major $F1$	Destroyed $F1$
Baseline [13]	28.41	80.48	6.09	65.79	7.08	2.16	26.40
Siamese-UNet (ResNext50) [34]	67.33	79.81	61.97	76.86	45.39	64.17	71.86
Siamese-UNet (DPN92) [34]	69.18	83.56	63.02	81.55	43.90	66.21	75.01
Dual-HRNet [59]	71.35	83.61	66.09	86.43	48.66	69.11	71.80
Dual-Temporal Fusion [39]	72.52	82.76	68.13	86.29	50.77	68.71	**77.71**
RescueNet [60]	70.23	84.09	63.94	86.09	45.72	62.76	76.15
CHTR	**73.65**	**84.14**	**69.16**	**88.42**	**51.21**	**71.50**	76.84

a) '$F1$ score' denotes the overall $F1$ score, i.e., S_{xView2} in (14); 'Localization $F1$' is the $F1$ score of building segmentation; 'Damage $F1$' represents the $F1$ score of change detection.

乘以 $(1-\frac{\text{iter}}{\text{maxiter}})^{\text{power}}$ 降低学习率, 其中 power = 3. 本方法使用了深度学习框架 PyTorch [58] 来实现. 实验是在一台具有 4 个 16 GB 显存的 NVIDIA Tesla V100 GPU 的服务器上运行的. 在训练期间应用随机缩放、水平翻转和高斯模糊等数据增强技术来提高本方法的泛化能力.

5.2　与现有方法的比较

本小节介绍了本方法 CHTR 与其他几种变化检测方法在 xBD 数据集 [13] 上的结果比较, 包括 Baseline [13], Siamese-UNet (ResNext50) [34], Siamese-UNet (DPN92) [34], Dual-HRNet [59], Dual-Temporal Fusion [39] 和 RescueNet [60]. 在文献 [13] 中提出的 Baseline 采用 U-Net [15] 架构进行建筑物分割, 同时使用在 ImageNet [61] 上预训练的 ResNet-50 [14] 网络利用灾后遥感图像完成灾害损毁评估. Siamese-UNet 架构广泛用于变化检测任务, 其在灾前图像上使用 U-Net 架构进行建筑物分割, 并利用具有共享编码器权重的 Siamese-UNet 在双时序图像上完成变化检测. 本文基于 ResNext50 [62] 和 DPN92 [63] 实现了两个 Siamese-UNet 架构. 在对双时序特征图进行融合时, 本文实现的 Siamese-UNet 采用的是特征堆叠的融合方式. 文献 [39] 中的 Dual-Temporal Fusion 模型采用了 Mask R-CNN [40], 并增加了一个语义分割分支和一个特征金字塔网络模块 [41] 来完成建筑物分割和变化检测. RescueNet [60] 设计了一种新颖的位置感知损失函数来完成这两个任务. 为了公平地进行比较, 本文重新实现了这些方法, 并在相同的数据集上进行了实验.

定量比较. 本方法 CHTR 和其他方法的定量化比较结果如表 1 所示. 本方法在 Damage $F1$ 上获得了 69.16%, 比 Dual-Temporal Fusion [39] 高 1.03%. 此外, CHTR 在 3 个损毁级别上都取得了最好的性能, 例如在 Undamaged $F1$ 上达到了 88.42%, 在 Minor $F1$ 上达到了 51.21%, 分别比 Dual-HRNet [59] 高了 1.99% 和 2.55%. 对于建筑物分割任务, 在 Localization $F1$ 上, CHTR 优于 Dual-HRNet [59] 0.53%, 优于 Dual-Temporal Fusion [39] 1.38%. 值得注意的是, 本方法在变化检测任务中有效地实现了比广泛使用的孪生网络更好的性能, 这使其成为灾害评估中更好的选择.

如表 2 所示, 本文分析了 CHTR 和其他变化检测方法的网络参数量、浮点运算数 (floating point operations, FLOPs) 和运行时间. 其中浮点运算数 FLOPs 可以度量一个模型的复杂度, 单位为 G, 1 G FLOPs = 10^9 FLOPs. 为了公平地比较, 表中列出了对一张 1024×1024 图像的推理时间, 并且所有比较方法的实验都是在具有 4 个 NVIDIA Tesla V100 GPU 的服务器上实现的. 与其他方法相比, 本文所提方法 CHTR 的参数量和 FLOPs 都相对较少, 参数量比 Dual-Temporal Fusion [39] 少 10.4 M, 比

表 2 本文方法 CHTR 与其他变化检测方法在网络参数量、浮点运算数 (floating point operations, FLOPs) 和运行时间方面的比较分析

Table 2 Comparison of our proposed CHTR and some state-of-the-art change detection methods in the parameter number, floating point operations (FLOPs), and running time[a)]

Method	Param (M)	FLOPs (G)	Time (s)
Baseline [13]	44.2	224.9	0.039
Siamese-UNet (ResNext50) [34]	69.1	142.0	0.068
Siamese-UNet (DPN92) [34]	94.8	243.6	0.437
Dual-HRNet [59]	59.5	91.3	0.055
Dual-Temporal Fusion [39]	43.9	201.0	0.035
RescueNet [60]	40.4	182.9	0.043
CHTR	33.5	71.7	0.022

a) 'Param' refers to the number of parameters; 'FLOPs' is the floating point operations with input size 512×512; 'Time' represents the inference time for one image.

Dual-HRNet [59] 少 26.0 M. 值得注意的是, Siamese-UNet 的参数数量是所有比较方法中最多的, 因为它采用了两个网络进行建筑物分割和变化检测任务. 此外, CHTR 也获得了最快的推理时间. 综上所述, 本方法 CHTR 的计算复杂度更小且推理时间更快, 同时又能够获得更好的性能, 因此更有利于实际中的应用.

从上述实验中可以看出, 孪生网络在引入较多参数量的同时, 对于建筑物分割和变化检测的性能却相对较差, 因此本文做了一系列实验探索其中可能的原因. 如果将本文方法混合 CNN 和 Transformer 架构的编码器替换到 Siamese-UNet 中, 即分成两个阶段完成两个任务, 可以获得实验结果: $F1$ score 为 72.78%、Localization $F1$ 为 83.65%、Damage $F1$ 为 68.12%, 这个结果要低于本文的方法 CHTR. 如果将 Siamese-UNet 改为单阶段的结构, 即利用孪生网络提取灾前和灾后图像的特征, 两个时序的特征在融合后, 利用一个解码器网络同时输出建筑分割和变化检测的结果. Siamese-UNet (ResNext50) 的结果: $F1$ score 为 69.06%、Localization $F1$ 为 82.91%、Damage $F1$ 为 63.13%. Siamese-UNet (DPN92) 的结果: $F1$ score 为 70.08%、Localization $F1$ 为 83.60%、Damage $F1$ 为 64.28%. 可以看出, 相比于双阶段的网络结构, 单阶段的结构 (即同时完成两个任务) 取得了更好的性能. 从这些实验中可以得出, 双阶段的孪生网络性能较低的原因为将两个任务分开进行训练, 使其无法从多任务学习中受益. 由于建筑物分割和变化检测两个任务所包含的知识是相似的且可以共享, 多任务学习能够提高相关任务的性能 [64].

定性比较. 在 xBD 数据集上的可视化结果如图 4 所示. 其中分析了来自不同灾害的 7 个示例, 包括飓风 – 佛罗伦萨、飓风 – 哈维、海啸、洪水、地震、火灾和野火. 为了更加清晰地比较可视化结果, 本文将示例飓风 – 哈维和洪水的部分区域进行了放大展示, 图 4 中有青色外框的图片即为相应的放大图. 从图中可以看出, 本方法的建筑分割结果更准确, 变化等级的分类也与真值更加一致. 例如, 在图 4 的第 5 行结果以及其第 6 行的放大图中, 有几座建筑物被洪水损毁. 其他方法无法识别完整的建筑物 (例如, Siamese-UNet (ResNext50) [34], Siamese-UNet (DPN92) [34] 和 Dual-HRNet [59]) 或无法准确识别损毁等级 (例如, Baseline [13] 和 Dual-Temporal Fusion [39]), 而 CHTR 实现了更好的可视化结果, 非常接近地面真值. 这些在 xBD 数据集上的可视化结果, 进一步验证了本方法在建筑物分割和多级变化检测两个任务中的有效性.

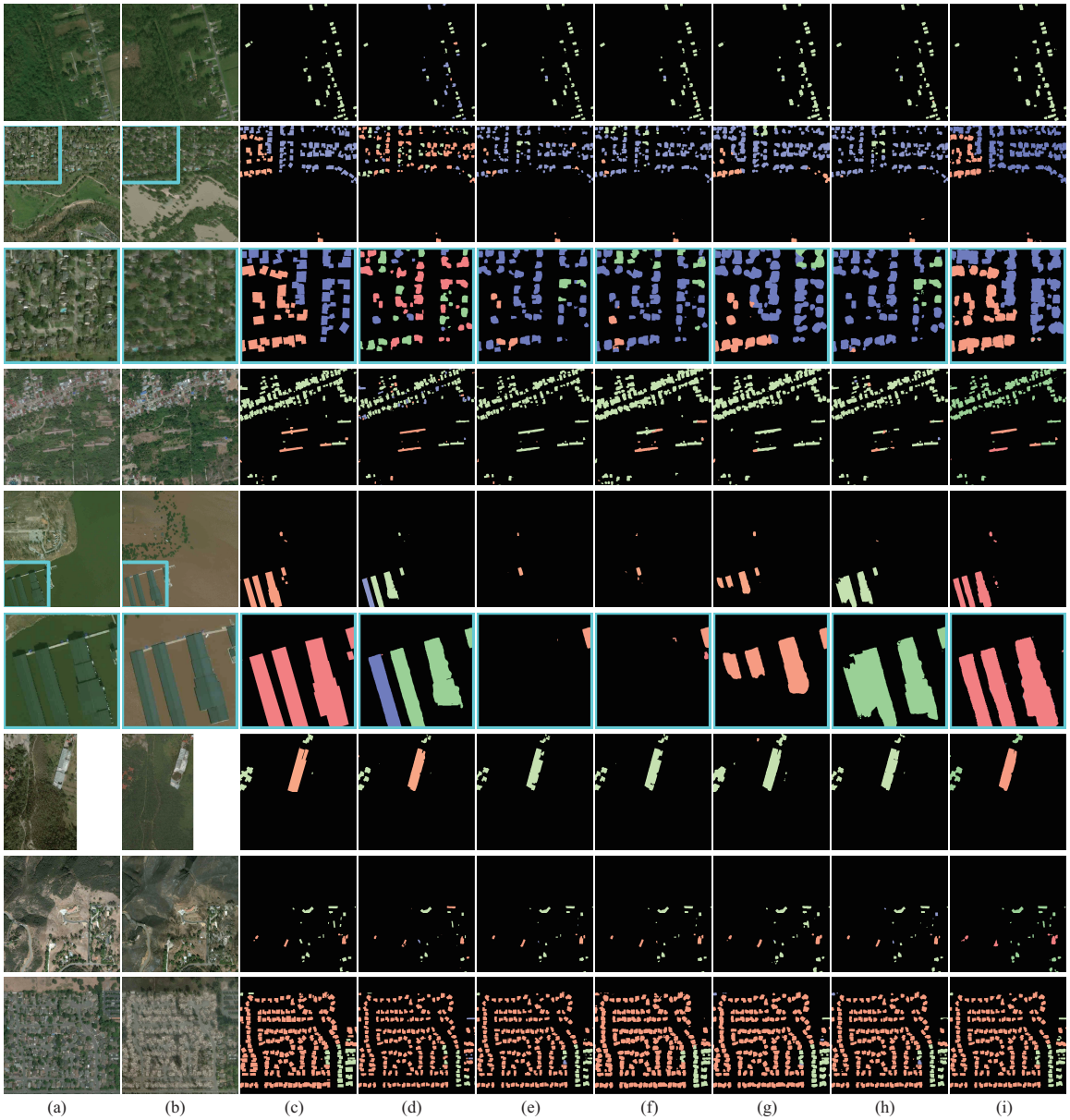

图 4 (网络版彩图) 本文方法 CHTR 与其他方法的可视化比较. 第 1~9 行分别展示了不同灾害的结果, 包括飓风 – 佛罗伦萨、飓风 – 哈维、飓风 – 哈维的放大图、海啸、洪水、洪水的放大图、地震、火灾和野火. 有青色外框的图片为放大图. (a) 灾前遥感图像; (b) 灾后遥感图像; (c) 多级别损毁的真值, 其中的 4 种颜色代表 4 种损毁等级: 绿色、蓝色、橙红色和红色分别表示无损毁、轻微损毁、严重损毁和完全损毁; (d)~(i) Baseline [13], Siamese-UNet (ResNext50) [34], Siamese-UNet (DPN92) [34], Dual-HRNet [59], Dual-Temporal Fusion [39] 和 CHTR

Figure 4 (Color online) Qualitative comparison of our CHTR and other methods. The first to the ninth rows show the results with different disasters, including hurricane-Florence, hurricane-Harvey, zoomed-in images of hurricane-Harvey, tsunami, flooding, zoomed-in images of flooding, earthquake, fire, and wildfire. The image with the cyan frame is an enlarged image. (a) Pre-disaster satellite imagery; (b) post-disaster satellite imagery; (c) ground truth of multi-level damages, where four colors represent four damage scales. The green, blue, orange, and red denote no damage, minor damage, major damage, and destroyed. (d)~(i) Results of Baseline [13], Siamese-UNet (ResNext50) [34], Siamese-UNet (DPN92) [34], Dual-HRNet [59], Dual-Temporal Fusion [39], and CHTR

表 3 全局差异模块 (GD) 和局部门控注意力模块 (LGA) 的消融实验 (%)

Table 3 Ablation study for the proposed global difference (GD) module and local gated attention (LGA) module (%)[a)]

No.	GD	LGA	Time (s)	$F1$ score	Localization $F1$	Damage $F1$	Undamaged $F1$	Minor $F1$	Major $F1$	Destroyed $F1$
1			0.015	69.38	79.89	64.88	84.86	46.11	68.49	73.65
2	✔		0.019	71.08	82.42	66.22	86.37	48.83	67.13	74.32
3		✔	0.018	71.67	**84.19**	66.31	86.97	47.93	68.53	74.79
4	✔	✔	0.022	**73.65**	84.14	**69.16**	**88.42**	**51.21**	**71.50**	**76.84**

a) No.1 is the network with ResNet-50 as the encoder and progressive upsampling blocks as the decoder. We add the GD and LGA modules to show their effectiveness (No.2 and No.3). No.4 is the full version of our proposed approach CHTR. 'Time' represents the inference time for one image.

5.3 消融实验

所设计模块的有效性. 在 CHTR 中, 本文设计了一个全局差异 (GD) 模块来获取全局变化信息, 提出了一个局部门控注意力 (LGA) 模块以学习多级别变化之间的局部依赖关系. 如表 3 中所列, 本文进行了消融实验来验证所提出两个模块的有效性. No.1 是以 ResNet-50 作为编码器, 渐进式上采样块作为解码器的基准网络. 将设计的全局差异模块和局部门控注意力模块添加到 No.1 后, Damage $F1$ 从 64.88% 分别提高到 66.22% 和 66.31%, Localization $F1$ 提高了 2.53% 和 4.30%. 结合这两个模块后, 相比于 No.1, Damage $F1$ 提高了 4.28%, $F1$ score 提高了 4.27%. 这些实验验证了本文所设计的两个模块可以提高对灾害前后图像中不同变化的判别能力. 同时, 本文也分析了在不同模块配置下, 模型对于一张 1024×1024 图像的推理时间. 可以看到, 不同配置下模型的推理时间较为接近, No.1 的模型较为简单, 推理时间更快.

本文方法 CHTR 在不同模型配置下的可视化结果如图 5 所示. 从图中可以看出, 基准网络对于建筑分割和不同变化等级的识别都相对较差. 添加了全局差异模块后, 能够更好地定位变化的位置, 但是其在一些局部位置不同变化等级的识别上会出错. 局部门控注意力模块能够帮助局部范围内变化等级的识别, 在一些小区域变化等级的识别上结果较好, 但是对变化位置的定位能力较差, 会出现错误识别大块区域变化等级的情况. 结合这两个模块后, 可以有效缓解使用单个模块出现的问题, 获得更好的变化检测结果.

Transformer 中不同配置的影响. 本文开展了多个实验来分析 Transformer 层中不同参数设置对结果的影响, 包括 Transformer 的层数 l 和多头自注意力的头数 h. 对 Transformer 层数的分析如表 4 所示, 当 Transformer 层数 $l = 2$ 时, $F1$ score 为 73.65%, 比其他两个设置 $l = 1$ 和 $l = 3$ 分别高了 4.18% 和 2.62%. 表 5 是对多头自注意力的头数 h 的分析结果. 可以看出, 当头数 $h = 4$ 时性能是最好的, $F1$ score 相比于头数 $h = 2$ 和 $h = 8$ 分别提升了 2.39% 和 1.45%. 此外, 本文也分析了局部门控注意力模块中不同尺寸特征块的消融实验, 结果如表 6 所示. 可以看出, 4×4 是 LGA 模块中特征块尺寸的最佳设置, 而特征块越大或者越小都会限制性能的提升. 在 Transformer 中, 序列的长度与特征块尺寸的平方成反比. 减少特征块尺寸可以提高性能, 是因为 Transformer 层使用更长的输入序列来编码复杂的依赖关系. 但是, 太小的特征块可能会阻碍模型学习多级变化之间的相似性, 进而影响变化检测的结果.

根据上述实验, 本文设置 Transformer 层数 $l = 2$, 设置多头自注意力中的头数 $h = 4$, 并使用 4×4 作为本方法中的默认特征块尺寸.

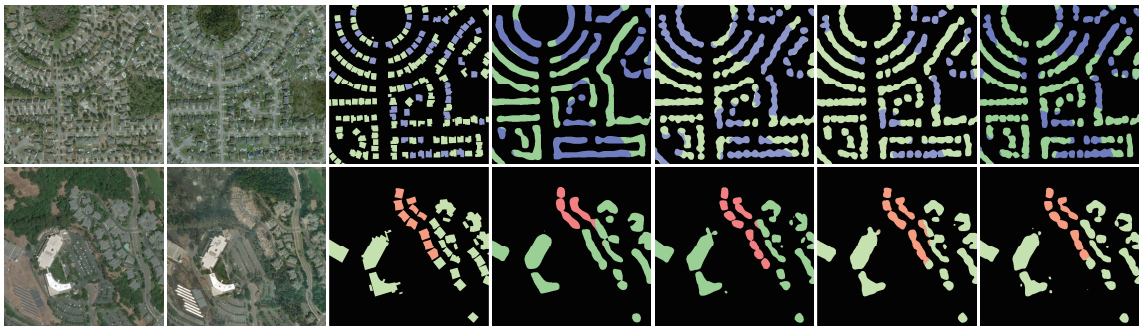

图 5 (网络版彩图) 本文方法 CHTR 在不同模型配置下的可视化结果. 第 1 和 2 行分别是来自飓风和野火灾害的示例. (a) 灾前遥感图像; (b) 灾后遥感图像; (c) 多级别损毁的真值; (d) 表 3 中 No.1 基准网络的可视化结果; (e) 表 3 中 No.2 的可视化结果; (f) 表 3 中 No.3 的可视化结果; (g) 表 3 中 No.4, 即本文方法 CHTR 的可视化结果.

Figure 5 (Color online) The visualization results of our CHTR under different model configurations. The first and second rows show the examples from hurricane and wildfire disasters, respectively. (a) Pre-disaster satellite imagery; (b) post-disaster satellite imagery; (c) ground truth of multi-level damages; visualization results of (d) No.1 in Table 3, i.e., benchmark network, (e) No.2 in Table 3, (f) No.3 in Table 3, and (g) No.4 in Table 3, i.e., our CHTR

表 4 Transformer 层数 l 的消融实验 (%), 其中全局差异模块和局部门控注意力模块中的 Transformer 层数设置为相同

Table 4 Ablation study for the number of transformer layers l (%), where the numbers of transformer layers in the GD and LGA modules are set the same

No.	l	$F1$ score	Localization $F1$	Damage $F1$	Undamaged $F1$	Minor $F1$	Major $F1$	Destroyed $F1$
1	1	69.47	79.89	65.00	84.82	46.07	67.82	75.21
2	2	**73.65**	**84.14**	**69.16**	88.42	**51.21**	**71.50**	**76.84**
3	3	71.03	79.22	67.52	**90.31**	47.84	70.45	76.51

表 5 Transformer 中多头自注意力的头数 h 的消融实验 (%), 其中全局差异模块和局部门控注意力模块中多头自注意力的头数设置为相同

Table 5 Ablation study for the number of heads h of the multi-head self-attention in the transformer layer (%), where the head numbers in the GD and LGA modules are set the same

No.	h	$F1$ score	Localization $F1$	Damage $F1$	Undamaged $F1$	Minor $F1$	Major $F1$	Destroyed $F1$
1	2	71.26	81.27	66.97	87.64	48.58	70.01	74.37
2	4	**73.65**	**84.14**	**69.16**	88.42	**51.21**	**71.50**	76.84
3	8	72.20	84.12	67.09	**88.63**	48.70	66.53	**78.26**

两个任务不同的损失函数组合的影响. 对于建筑物分割和变化检测两个任务损失函数的选择, 本文开展了消融实验来分析, 结果如表 7 所示. 本文选择了被广泛使用的交叉熵损失和能够较好地处理类别不平衡问题的 Combo Loss [54], 其中 Combo Loss 是 Dice Loss [55] 和 Focal Loss [56] 的加权和. 通过分析这两个损失函数对于两个任务的几种典型组合, 可以看出, 对建筑物分割任务采用 Combo Loss 以及对变化检测任务采用交叉熵损失时, 本模型可以达到最优的性能. 因此本文针对这两个任务选择了目前所用的损失函数.

表 6 对局部门控注意力模块中不同尺寸特征块的消融实验 (%)

Table 6 Ablation study for the proposed LGA module with different patch sizes (%)

No.	Patches	$F1$ score	Localization $F1$	Damage $F1$	Undamaged $F1$	Minor $F1$	Major $F1$	Destroyed $F1$
1	2×2	71.11	79.22	67.64	**89.27**	48.37	70.09	**76.94**
2	4×4	**73.65**	**84.14**	**69.16**	88.42	**51.21**	**71.50**	76.84
3	8×8	72.47	81.85	68.44	88.08	50.41	71.34	75.54

表 7 对建筑分割和变化检测两个任务损失函数组合的消融实验 (%)

Table 7 Ablation study on the combination of two loss functions for the building segmentation and change detection tasks (%)[a)]

No.	Building segmentation	Change detection	$F1$ score	Localization $F1$	Damage $F1$
1	CE Loss	CE Loss	71.46	81.51	67.15
2	Combo Loss	Combo Loss	72.57	82.78	68.19
3	CE Loss	Combo Loss	71.99	82.11	67.66
4	Combo Loss	CE Loss	**73.65**	**84.14**	**69.16**

a) 'Combo Loss' [54] is the weighted sum of Dice Loss [55] and Focal Loss [56]; 'CE Loss' is the cross-entropy loss.

6 总结与讨论

本文提出了一种新颖的变化检测模型 CHTR, 基于双时序遥感图像同时进行建筑物分割和多级变化检测. 本文结合 CNN 擅长学习局部细节特征和 Transformer 可以建模长程依赖关系的优势, 采用混合 CNN 和 Transformer 的架构作为编码器, 同时引入渐进式上采样策略作为解码器. 一张高分辨率遥感图像通常覆盖较大的面积, 其中包含的建筑物在自然灾害发生后会受到不同程度的损毁. 本文提出了一个全局差异 (GD) 模块来学习全局变化模式. 同时设计了一个局部门控注意力 (LGA) 模块, 以获取多级别变化之间的局部依赖性并增强对双时序图像之间不同变化的区分. 在 xBD 数据集上的大量实验证明了本方法的优越性, 开展的消融实验也验证了所提出的两个模块的有效性. 未来, 我们计划调整 Transformer 层中的自注意力机制, 以进一步提升其性能.

参考文献

1 Chen C F, Son N T, Chang N B, et al. Multi-decadal mangrove forest change detection and prediction in Honduras, Central America, with Landsat imagery and a Markov chain model. Remote Sens, 2013, 5: 6408–6426

2 Song C, Huang B, Ke L, et al. Remote sensing of alpine lake water environment changes on the Tibetan Plateau and surroundings: a review. ISPRS J Photogrammetry Remote Sens, 2014, 92: 26–37

3 Marin C, Bovolo F, Bruzzone L. Building change detection in multitemporal very high resolution SAR images. IEEE Trans Geosci Remote Sens, 2014, 53: 2664–2682

4 Mahdavi S, Salehi B, Huang W, et al. A PolSAR change detection index based on neighborhood information for flood mapping. Remote Sens, 2019, 11: 1854

5 Singh A. Review article digital change detection techniques using remotely-sensed data. Int J Remote Sens, 1989, 10: 989–1003

6 Im J, Jensen J R. A change detection model based on neighborhood correlation image analysis and decision tree classification. Remote Sens Environ, 2005, 99: 326–340

7 Gapper J J, El-Askary H, Linstead E, et al. Coral reef change detection in remote pacific islands using support vector machine classifiers. Remote Sens, 2019, 11: 1525

8 Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image seg-

mentation. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 2481–2495
9 Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 801–818
10 Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell, 2017, 40: 834–848
11 Ji M, Liu L, Buchroithner M. Identifying collapsed buildings using post-earthquake satellite imagery and convolutional neural networks: a case study of the 2010 Haiti earthquake. Remote Sens, 2018, 10: 1689
12 Rudner T G J, Rußwurm M, Fil J, et al. Multi^{3}Net: segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2019. 33: 702–709
13 Gupta R, Goodman B, Patel N, et al. Creating xBD: a dataset for assessing building damage from satellite imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019. 10–17
14 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778
15 Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015. 234–241
16 Huang Z, Wang X, Huang L, et al. CCNet: criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 603–612
17 Mei J, Cheng M M, Xu G, et al. SANet: a slice-aware network for pulmonary nodule detection. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 4374–4387
18 Wang X, Girshick R, Gupta A, et al. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 7794–7803
19 Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers. In: Proceedings of European Conference on Computer Vision. Cham: Springer, 2020. 213–229
20 Zheng S, Lu J, Zhao H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 6881–6890
21 Roy M, Routaray D, Ghosh S, et al. Ensemble of multilayer perceptrons for change detection in remotely sensed images. IEEE Geosci Remote Sens Lett, 2013, 11: 49–53
22 Jia L, Li M, Zhang P, et al. SAR image change detection based on correlation kernel and multistage extreme learning machine. IEEE Trans Geosci Remote Sens, 2016, 54: 5993–6006
23 Volpi M, Tuia D, Bovolo F, et al. Supervised change detection in VHR images using contextual information and support vector machines. Int J Appl Earth Observation Geoinf, 2013, 20: 77–85
24 Kasetkasem T, Varshney P K. An image change detection algorithm based on Markov random field models. IEEE Trans Geosci Remote Sens, 2002, 40: 1815–1823
25 Huo C, Zhou Z, Lu H, et al. Fast object-level change detection for VHR images. IEEE Geosci Remote Sens Lett, 2009, 7: 118–122
26 Tan K, Zhang Y, Wang X, et al. Object-based change detection using multiple classifiers and multi-scale uncertainty analysis. Remote Sens, 2019, 11: 359
27 Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell, 2016, 39: 1137–1149
28 Wang Y, Mei J, Zhang L, et al. Self-supervised feature learning with CRF embedding for hyperspectral image classification. IEEE Trans Geosci Remote Sens, 2018, 57: 2628–2642
29 Hou Q B, Han L-H, Liu J-J, et al. Autonomous learning of semantic segmentation from Internet images. Sci Sin Inform, 2021, 51: 1084–1099 [侯淇彬, 韩凌昊, 刘姜江, 等. 互联网图像驱动的语义分割自主学习. 中国科学: 信息科学, 2021, 51: 1084–1099]
30 Gao S H, Cheng M M, Zhao K, et al. Res2Net: a new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell, 2021, 43: 652–662
31 Daudt R C, Saux B L, Boulch A, et al. Guided anisotropic diffusion and iterative learning for weakly supervised change

detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019

32 Papadomanolaki M, Verma S, Vakalopoulou M, et al. Detecting urban changes with recurrent neural networks from multitemporal Sentinel-2 data. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, 2019. 214–217

33 Liu J, Gong M, Qin K, et al. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. IEEE Trans Neural Netw Learn Syst, 2016, 29: 545–559

34 Daudt R C, Saux B L, Boulch A. Fully convolutional siamese networks for change detection. In: Proceedings of IEEE International Conference on Image Processing (ICIP), 2018. 4063–4067

35 Chen H, Wu C, Du B, et al. Change detection in multisource VHR images via deep siamese convolutional multiple-layers recurrent neural network. IEEE Trans Geosci Remote Sens, 2019, 58: 2848–2864

36 Du B, Ru L, Wu C, et al. Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images. IEEE Trans Geosci Remote Sens, 2019, 57: 9976–9992

37 Wu C, Chen H, Du B, et al. Unsupervised change detection in multitemporal VHR images based on deep kernel PCA convolutional mapping network. IEEE Trans Cybern, 2022, 52: 12084–12098

38 Duarte D, Nex F, Kerle N, et al. Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach. ISPRS Ann Photogramm Remote Sens Spatial Inf Sci, 2018, 4: 89–96

39 Weber E, Kané H. Building disaster damage assessment in satellite imagery with multi-temporal fusion. In: Proceedings of International Conference on Learning Representations Workshop, 2020

40 He K, Gkioxari G, Dollár P, et al. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 2961–2969

41 Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 2117–2125

42 Bai Y, Hu J, Su J, et al. Pyramid pooling module-based semi-siamese network: a benchmark model for assessing building damage from xBD satellite imagery datasets. Remote Sens, 2020, 12: 4055

43 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 5998–6008

44 Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 4171–4186

45 Wang Y, Xu Z, Wang X, et al. End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 8741–8750

46 Wang H, Zhu Y, Green B, et al. Axial-DeepLab: stand-alone axial-attention for panoptic segmentation. In: Proceedings of European Conference on Computer Vision. Cham: Springer, 2020. 108–126

47 Zhao H, Jia J, Koltun V. Exploring self-attention for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 10076–10085

48 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations, 2020

49 Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers distillation through attention. In: Proceedings of International Conference on Machine Learning, 2021. 10347–10357

50 Zhu X, Su W, Lu L, et al. Deformable DETR: deformable transformers for end-to-end object detection. In: Proceedings of International Conference on Learning Representations, 2020

51 Zhang C, Yue P, Tapete D, et al. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS J Photogrammetry Remote Sens, 2020, 166: 183–200

52 Chen J, Yuan Z, Peng J, et al. DASNet: dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images. IEEE J Sel Top Appl Earth Observ Remote Sens, 2020, 14: 1194–1206

53 Chen H, Qi Z, Shi Z. Remote sensing image change detection with transformers. IEEE Trans Geosci Remote Sens, 2022, 60: 1–14

54 Taghanaki S A, Zheng Y, Zhou S K, et al. Combo loss: handling input and output imbalance in multi-organ segmentation. Computized Med Imag Graphics, 2019, 75: 24–33

55 Milletari F, Navab N, Ahmadi S A. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the 4th International Conference on 3D Vision (3DV), 2016. 565–571
56 Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 2980–2988
57 Valanarasu J M J, Oza P, Hacihaliloglu I, et al. Medical transformer: gated axial-attention for medical image segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2021. 36–46
58 Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019. 32: 8026–8037
59 Wang J, Sun K, Cheng T, et al. Deep high-resolution representation learning for visual recognition. IEEE Trans Pattern Anal Mach Intell, 2021, 43: 3349–3364
60 Gupta R, Shah M. RescueNet: joint building segmentation and damage assessment from satellite imagery. In: Proceedings of the 25th International Conference on Pattern Recognition (ICPR), 2021. 4405–4411
61 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009. 248–255
62 Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1492–1500
63 Chen Y, Li J, Xiao H, et al. Dual path networks. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017. 4470–4478
64 Zhang Y, Yang Q. A survey on multi-task learning. IEEE Trans Knowl Data Eng, 2022, 34: 5586–5609

Damage assessment with global differences and local attention

Jie MEI & Ming-Ming CHENG*

Tianjin Media Computing Center (*TMCC*)*, College of Computer Science, Nankai University, Tianjin* 300350*, China*
* Corresponding author. E-mail: cmm@nankai.edu.cn

Abstract Detecting the different changes caused by a natural disaster is critical for effectively directing humanitarian assistance and disaster response operations. However, it is challenging due to the large-scale disaster areas and complex ground environments. Existing assessment methods are usually labor-intensive and unsuitable for multiple disasters. In this paper, we propose a change transformer (CHTR) model for simultaneous building localization and multi-level change detection from dual-temporal satellite imagery. Based on the advantages that convolutional neural networks (CNNs) are good at learning detailed local features and the transformer can model long-range dependencies, we adopt a hybrid CNN-transformer architecture as the encoder. A natural disaster usually causes varying degrees of damage to buildings in a complex environment; thus, we propose a global difference module on the original features obtained by the CNN to capture the global change pattern and improve the overall awareness of the variations between dual-temporal images. Furthermore, a local gated attention module on the patches of features after the CNN is further developed to learn the local dependencies among the multi-level changes, which augments the discrimination of different changes. Extensive experiments on the largest building damage assessment dataset, xBD, demonstrate that the proposed CHTR model establishes new state-of-the-art results.

Keywords building segmentation, change detection, satellite imagery, global-local architecture, transformer