DOI:10.19651/j.cnki.emt.2313583

融合 Swin Transformer 的遥感影像建筑物变化检测*

于政尧[1,2] 黄建华[1,2] 孙希延[1,2] 罗明明[1,2] 万逸轩[1,2]
(1.桂林电子科技大学广西精密导航技术与应用重点实验室 桂林 541004;
2.桂林电子科技大学信息与通信学院 桂林 541004)

摘 要:针对不同时序遥感影像中多种地物类型的变化信息杂乱、背景复杂导致难以清晰地提取关键特征的问题,本文提出了一种将 Swin Transformer 与孪生网络融合实现建筑物变化检测的新方法。该方法通过 4 个 Swin Transformer Block 的结构来获取不同层次的特征,针对不同尺度的特征图进行差异计算,以获取变化特征图。此外,在本文算法的基础上还引入了差异特征融合模块和边缘感知注意力模块。差异特征融合模块能更好地表达不同感受野下的特征,提高对细节特征和全局特征的融合效果;边缘感知注意力模块细化特征提取时特征图中建筑物的边缘特征,扩大模型的局部感受野,增强模型对于细节信息的检测能力,从而提高网络结构对建筑物边缘特征的提取能力。实验结果表明,本文方法与现有经典变化检测网络全卷积早期融合 FC-EF 相比,在两个公开数据集上的 F1 值分别提高了 7.36%和 19.67%。

关键词:遥感影像;孪生网络;Swin Transformer;变化检测
中图分类号:TP751.1 **文献标识码**:A **国家标准学科分类代码**:510.40

Remote sensing image building change detection by incorporating Swin Transformer

Yu Zhengyao[1,2] Huang Jianhua[1,2] Sun Xiyan[1,2] Luo Mingming[1,2] Wan Yixuan[1,2]
(1. Guangxi Key Laboratory of Precision Navigation Technology and Application, Guilin University of Electronic Technology, Guilin 541004, China; 2. Information and Communicaiton School, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: Aiming at the problem that it is difficult to extract key features clearly due to the cluttered change information of multiple feature types in different time-series remote sensing images and complex backgrounds, this paper proposes a new method of fusing Swin Transformer with twin networks to achieve building change detection. The method obtains features at different levels through the structure of four Swin Transformer blocks, and performs difference calculations for feature maps at different scales to obtain change feature maps. In addition, a difference feature fusion module and an edge-aware attention module are introduced based on the algorithm in this paper. The difference feature fusion module can better express the features under different perceptual fields and improve the fusion effect on detailed features and global features; the edge-aware attention module refines the edge features of buildings in the feature map during feature extraction, expands the local perceptual field of the model, enhances the detection ability of the model for detailed information, and thus improves the extraction ability of the network structure for building edge features. The experimental results show that the F1 values of this paper's method are improved by 7.36% and 19.67% on two public datasets, respectively, compared with the existing classical change detection network FC-EF.

Keywords: remote sensing images; twin networks; Swin Transformer; change detection

0 引 言

遥感影像变化检测是一项重要的研究领域,其主要针对同一位置下不同时期的遥感影像之间的变化进行检测,以掌握土地类型的变化信息[1]。遥感影像的变化检测在国土资源监管、城市土地资源规划、非法建筑群的监管和灾害

收稿日期:2023-05-09
*基金项目:国家自然科学基金(61861008,62061010,62161007)、广西自然科学基金(2018GXNSFAA294054,2019GXNSFBA245072)、桂林市科技局"桂林市国家可持续发展议程创新示范区建设"重点项目(20190219-1)资助

评估等方面有广泛的应用[2]，而针对建筑物类别的变化检测是其重要分支之一。建筑物变化检测包括对建筑物拆除、新建和因灾害而导致毁坏等情况的检测，对于让自然资源管理相关部门快速掌握地物变化信息，更好地开展城市建筑规划、防违建与违拆等工作具有重要意义。

随着我国卫星遥感技术的发展，高分辨率遥感影像已成为城市建筑物变化检测的主要数据来源。目前，越来越多的研究者利用不同的方法在高分影像上进行变化检测的研究。在过去，常见的建筑物变化检测方法主要基于像素级别的分析，通过对像素层面上的变化进行识别来确定建筑物的变化区域。然而，近年来也有研究者采用机器学习方法来进行变化检测。例如，高桂荣等[3]基于半监督支持向量机算法提出了结合空间信息和光谱信息的渐进直推式支持向量机（progressive transductive support vector machine，PTSVM）算法用于遥感图像的变化检测。张永梅等[4]提出先基于比值法进行像素级变化检测得到候选变化区域，然后根据候选变化区域的纹理分布和色调来进行特征级变化检测。施文灶等[5]利用像元法来构造图边，将两期影像的建筑物分割转化为图的分割，最后根据分割结果来进行识别对比。张志强等[6]提出先利用随机森林获取像元级变化检测结果，然后对后时相进行图像分割来获得影像对象，最后将二者融合来得到变化检测的结果。潘伟豪等[7]提出了一种基于 D-S（dempster-shafer）证据理论的变化检测模型，该模型通过多尺度分割得到建筑物变化的证据集合，融合了不同尺度下建筑物的变化证据。由于城市发展速度的加快，传统的变化检测方法需要大量人力物力进行操作，这些方法因其繁琐程度渐渐难以满足现代遥感应用变化检测的需求，更精准高效且自动化程度高的深度学习变成了遥感变化检测的主要方法。

传统方法对深层次特征缺少挖掘，对背景复杂以及伪变化区域难以精确检测，与传统方法不同，深度学习不仅能够挖掘遥感影像的深层特征，而且能通过捕捉上下文信息实现对背景复杂和伪变化区域精确检测，从而提高建筑物变化检测的精度。因此，将深度学习引入到遥感影像建筑物变化的检测中受到了广泛关注。例如，Daudt 等[8]提出了基于 UNet 改进的孪生结构变化检测网络，成为卷积神经网络应用在遥感变化检测领域的重要里程碑。张翠军等[9]在 UNet 网络上进行改进，通过将图像中的像素分为变化类与非变化类来进行变化检测。朱节中等[10]提出在 UNet＋＋中加入对不同语义层的加权融合，从而提取更精确的建筑特征。Chen 等[11]基于孪生卷积神经网络加入了双生注意力机制，捕获长距离的依赖关系来获取更多建筑特征。Liu 等[12]利用通道与空间位置的相互依赖关系来改善特征显示，解决一些深度学习方法提取边缘不规则，区域不完整的问题。Chen 等[13]将双时态图像表达为几个 token，并用 Transformer 编码器在基于 token 的时空中进行上下文信息建模，最后将学习到的信息用 Transformer 解码器细化原始特征。这些方法都在一定程度上取得了较好的效果，利用卷积神经网络进行特征提取和分类，使得变化检测的精度和效率都得到了提高。但是上述经典卷积神经网络在遥感图像背景复杂，干扰较多的情况下，由于缺少全局信息的关注可能会导致建筑物特征提取效果较差，容易丢失图像中不同区域相互关联的有效信息，而使用 Transformer 自注意力结构保证了全局信息的获取，但是容易丢失特征提取过程中多层级的局部细节，因此上述方法无法兼顾全局信息与局部信息，从而导致检测结果中的变化区域出现缺陷、分布不均以及边缘粗糙不规则的问题。

针对上述问题，本文提出了一种基于 Swin Transformer 和孪生网络结构的建筑物变化检测方法。该方法采用 Swin Transformer 多层次特征提取层来获取不同层次的建筑物特征，并使用孪生编码器逐步提取图像中的建筑物特征，实现不同时期建筑物特征的对比。为了能够在提取全局信息的过程中尽可能减少局部信息的丢失，本文设计在每一个 Swin Transformer Block 前加入一个边缘感知模块来优化网络对于建筑物边缘的感知，拓宽模型局部感受野的同时提高模型对于局部信息的关注度。在解码器中引入融合多尺度特征图的差异特征融合模块，通过融合不同层级的特征来获取图像中建筑物的最相关特征，最终实现多尺度建筑物变化特征的融合。

本文设计了 3 个实验来验证模型的有效性，首先在两个公开数据集上进行了对比实验，即将本文改进模型与经典的变化检测网络在公开数据集上分别进行训练和预测。实验结果显示，本文模型所提取的特征边缘形状均匀且规则，相比其他变化检测网络更接近真实标签。此外，通过消融实验来验证改进模块的有效性，实验结果证明本文所提出的改进模块对于模型性能的提升起到了积极的作用。

与以往研究相比，本文提出的方法在捕捉建筑物变化的特征信息方面更具优势，有望在土地资源监管、城市规划和灾害评估等领域得到应用。

1　Swin Transformer 的网络结构

2017 年，Transformer 网络模型被提出后在自然语言处理领域得到广泛应用[14]。2020 年，Google 团队将 Transformer 改进成 Vision Transformer（ViT），并应用于图像处理领域[15]。ViT 将输入图片分为 16 个 patch 以减少 Transformer 自注意力计算的参数量，但该方法缺少卷积神经网络针对多尺度特征的卷积计算。因此，亚洲微软研究院在 2021 年提出了 Swin Transformer[16]，该模型基于 Transformer 中的自注意力机制，设计了一个多层级多阶段的网络结构，使其能够在不同尺度上提取特征，并作为主干网络应用于各种图像处理任务。Swin Transformer 的出现使得 Transformer 在图像处理领域的应用更加广泛。Swin Transformer 相比于 ViT 有着参数量更少、提取多层

次特征更强的优点,Swin Transformer 一共包含 4 个 Stage,该网络结构的基本流程如图 1 所示。

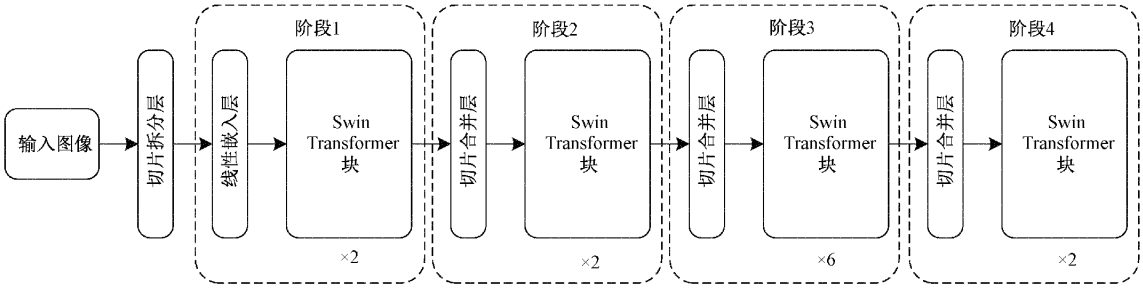


图 1　Swin Transformer 网络结构图

高度为 H 宽度为 W 的 3 通道图片输入到网络后经过切片拆分层(patch partition)后,每个像素都被切分为 16 个 patch,然后各个像素都沿通道的方向进行展平,因此展平后图像的通道数变为 $4\times4\times3=48$,图像的大小变为 $H/4$,$W/4$,输出后经过线性嵌入层(linear embedding)得到 $H/4\times W/4\times C$ 作为第一个 Swin Transformer Block 的输入。与 ViT 不同的地方主要是将多头自注意力(multi-head self-attention,MSA)更换为基于移位窗口的自注意力(shifted windows multi-head self-attention,SW-MSA)和基于窗口的自注意力(windows multi-head self-attention,W-MSA)。在每一个归一化层(layer normalization,LN)后都连接一个 W-MSA,加入残差连接后将输出接入到另一个归一化层中,最后通过多层感知机(multilayer perceptron,MLP)来完成非线性变换。第二层结构与上一层不同的地方仅在于将 W-MSA 换成了 SW-MSA。每一个 Swin Transformer Block 都由数个这样两层的结构组合在一起。如下图 2 所示是一个 Swin Transformer Block 的结构图:

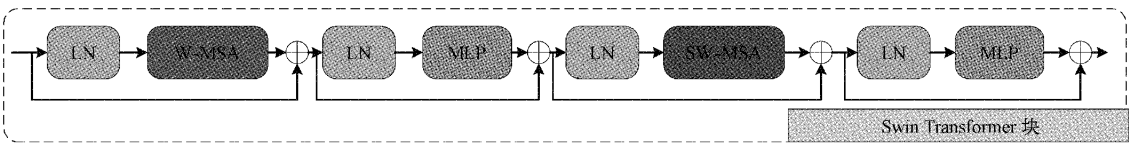


图 2　Swin Transformer 块网络结构

Swin Transformer Block 的计算公式如下[16]:

$$\hat{z}l = \mathrm{W-MSA}(\mathrm{LN}(z^{l-1})) + z^{l-1} \tag{1}$$

$$z^{l} = \mathrm{MLP}(\mathrm{LN}(\hat{z}l)) + \hat{z}l \tag{2}$$

$$\hat{z}^{l+1} = \mathrm{SW-MSA}(\mathrm{LN}(z^{l})) + z^{l} \tag{3}$$

$$z^{l+1} = \mathrm{MLP}(\mathrm{LN}(\hat{z}^{l+1})) + \hat{z}l + 1 \tag{4}$$

其中,$\hat{z}^{l}$ 和 z^{l} 分别代表(S)W-MSA 和 MLP 模块的输出,而 z^{l-1} 和 z^{l+1} 则代表 Swin Transformer Block 的输入和输出。

用以下输入图像为例,如图 3 所示,第 L 层将图片平均划分为 4 个窗口,并对每个窗口做自注意力的操作。第 L+1 层在原来划分规则的基础上进行平移划分,窗口的数量由原本的 4 个规则窗口增加到 9 个大小不一的窗口。然后通过循环移位使得移位窗口自注意力的计算量仍与原来一样。以上操作使得不同层里的不同窗口都得以信息交互,将此过程归纳为 Swin Transformer 第一个 Stage。

在实际中多尺度信息能更好地提取图像中的特征,卷积神经网络中通常使用下采样的操作来提取不同尺度的特征信息,而 Swin Transformer 采用了图块拼接(Patch Merging)的方法来进行类似于下采样的操作,减少图像的大小同时增加图像的通道数。在每个 Stage 之间插入一个图块拼接层来进一步挖掘深层次信息,这样的结构使得

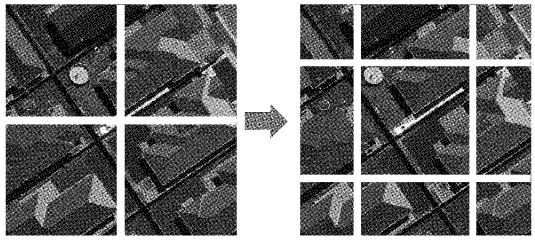


图 3　输入图像样例

Swin Transformer 拥有类似于卷积神经网络的多尺度特征提取的优点,弥补了 ViT 中只能对某一深度的特征图进行特征提取的缺点。

Swin Transformer 根据模型大小以及参数量的不同设计了 4 种应对不同场景的网络结构,本文采用了 Swin-T 的结构,其中 4 个 Stage 分别包含 2、2、6、2 个 Swin Transformer Block。

2　基于改进 Swin Transformer 的网络模型

2.1　算法框架

本文所提方法以 Swin Transformer 为主干网络,采用孪生网络的结构来构建变化检测算法网络。该网络主要由编码器、差异特征融合模块和解码器组成。编码器部分

采用 Swin Transformer 的四层网络结构，用于从输入特征图多层级、多尺度地提取信息。Swin Transformer 编码器中一共包含 4 个阶段，在每个阶段中，都在 Swin Transformer Block 前嵌入了一个边缘感知注意力模块用于细化建筑物边缘，引导网络关注局部信息，增强有效特征和抑制伪变化区域特征，改善自注意力结构对局部信息缺少关注的问题。解码器部分则是利用上采样来恢复图像的维度和大小，最终提取出来的结果是一个二分类结果图，白色部分表示发生变化区域，黑色部分则是未发生变化区域。本文的算法网络结构图如图 4 所示。

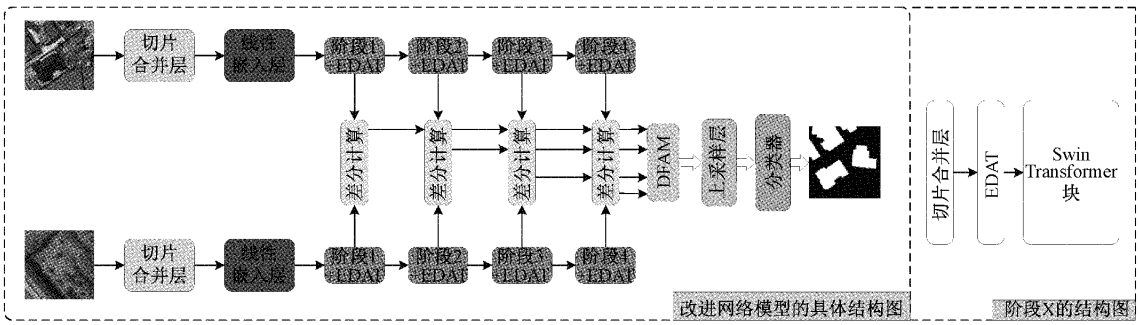


图 4　所提方法网络结构图

2.2　差异特征融合模块

差异特征融合注意力模块(differential feature fusion attention module，DFAM)是一种基于注意力机制的多尺度特征融合模块，用于不同层次和尺度的特征图进行特征融合。与传统的多尺度特征融合方法不同，DFAM 引入了注意力机制来调整输入特征图的融合权重，以确保最终特征图的质量和稳定性。

DFAM 主要基于迭代式注意力特征融合(iterative attentional feature fusion，iAFF)组成而来[17]，iAFF 模块的核心思想是通过对输入的两个特征图进行比较，计算出特征之间的差异，并将差异度量作为注意力权重，进而控制两个特征图在融合过程中的贡献程度。DFAM 不仅能减少因特征图质量差异引起的性能下降的同时，还有效平衡融合特征图之间语义和尺度信息的不一致。

图 5(a)为 iAFF 模块结构，图 5(b)为 iAFF 所包含的注意力机制(multi-scale channel attention module，MS-CAM)结构：

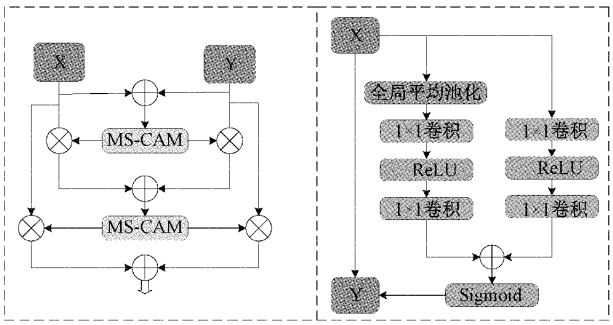


图 5　所提各模块的结构图

其中，图 5(a)表示 iAFF 将两个不同尺度的变化特征图进行特征融合。输入 X 和输入 Y 分别为不同尺度上经过差异模块计算后得到的变化特征图，将 X 和 Y 元素相加并通过 MS-CAM 计算得出不同特征图的融合权重，然后将各自的权重乘上对应的输入特征图，最后相加可得到第一层经过注意力机制所调整的初始特征。为了使得模型对输入特征图拥有更为完整的感知，可以继续将第一层的初始融合特征输入到第二个 MS-CAM 中，进行同样的加权平均融合从而得到最终的输出结果，这种通过两个注意力来融合特征的方法就是迭代式注意力特征融合(iAFF)。iAFF 中一次的特征融合可表示为[17]：

$$Z = M(X \oplus Y) \otimes X + (1 - M(X \oplus Y)) \otimes Y \quad (5)$$

其中，Z 表示第一次融合操作的输出，X 和 Y 则代表了输入的两张特征图，$\oplus$ 表示广播操作，$\otimes$ 表示对应元素相乘，M 代表了一次 MS-CAM 操作，iAFF 实际上进行了两次特征融合操作。

如图 5(b)所示，MS-CAM 注意力模块主要由全局注意力与本地注意力组成，MS-CAM 模块首先将输入的 X、Y 特征图进行相加，左侧经过全局平均池化减少特征图像维度，并将图片大小缩小至 1×1 大小。之后经过两个逐点卷积之间穿插激活函数 ReLU 的操作，通过逐点卷积来关注通道的尺度问题，同时保证该模块的轻量化。右侧相比左侧减少一个全局平均池化的操作，最后分别接入一个归一化层(batch normalization，BN)来加快模型的训练和收敛速度，最后将两侧输出通过广播操作进行相加。此外，MS-CAM 还使用残差连接防止梯度消失，减少过拟合。MS-CAM 可以用以下公式表示[17]：

$$X' = X \otimes M(X) = X \otimes \sigma(L(X) \oplus g(X)) \quad (6)$$

$$L(X) = B(PWConv_2(\delta(B(PWConv_1(X))))) \quad (7)$$

$$g(X) = bB(PWConv_2(\delta(B(PWConv_1(\mathbf{GAP}(X)))))) \quad (8)$$

其中，X 代表输入图像，GAP 代表全局平均池化层，PWConv(pointwise convolutional)代表 1×1 卷积层，σ 代表 Sigmoid 激活函数，δ 代表 ReLU 激活函数，B 代表归一

化层 BN。

一般传统的特征融合仅仅是将不同尺度的特征进行相加,这样会使得图像中的特征细节模糊化然后丢失,在特征融合中加入注意力机制会使得神经网络对输入特征图有完整的感知。相比于传统的线性操作,这种将多尺度的上下文特征信息进行融合并聚集到注意力模块中的方法特征提取能力更强。在本文所提方法中采用了类金字塔的方式进行 DFAM 的搭建,4 个输入特征对应 Swin Transformer 的四个阶段的输出特征进行差异计算的结果,以下图的方式使 4 个差异特征图最终输出为一张经过多尺度融合的特征图,DFAM 整体模块结构如图 6 所示。

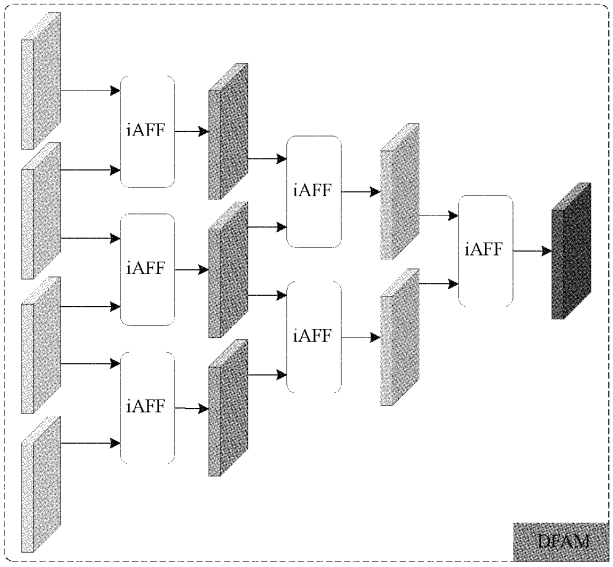


图 6　差异特征融合模块结构图

Swin Transformer 输出的不同尺度特征变化图反应了不同的变化信息,通过融合注意力机制的多尺度特征融合模块来加强建筑物变化部分的特征,减少干扰信息,使得网络在建筑物变化检测上更具优势。

2.3　边缘感知注意力模块

Transformer 模块在实际应用中可能会忽略特征图中的部分局部信息,尽管 Swin Transformer 的窗口分割已经有效缓解了这种现象,但背景复杂、树木、阴影错综交横以及噪点比较多的情况下还是会出现边缘粗糙不连续、模型难以识别建筑物的情况。针对上述问题,本文提出一个融合注意力的边缘感知模块(edge-aware attention module,EDAT),将 Triplet 注意力融入到建筑物变化检测的特征提取中,通过加强建筑物边缘信息的提取能力来提高变化检测的精度。这个注意力模块的设计相对于传统的注意力模块更加注重不同维度之间的交互,通过 Z-pool 操作和轴的旋转来实现空间维度和通道维度之间的交互[18]。这样可以在不增加太多模型参数的前提下加强特征图的多维度交互,从而更好地捕捉输入特征中的关键信息。EDAT 通过融入注意力机制来改进引文的边缘感知模块[19],受空洞空间卷积池化金字塔 ASPP(atrous spatial pyramid pooling)中用空洞卷积扩大局部感受野的启发[20],通过扩大特征提取网络的局部感受野来提取更充分的细节信息,丰富特征提取后的建筑物细节。上述操作可以丰富局部信息,使得本文模型在实际应用中更加可靠和有效。该模块的结构如图 7 所示。

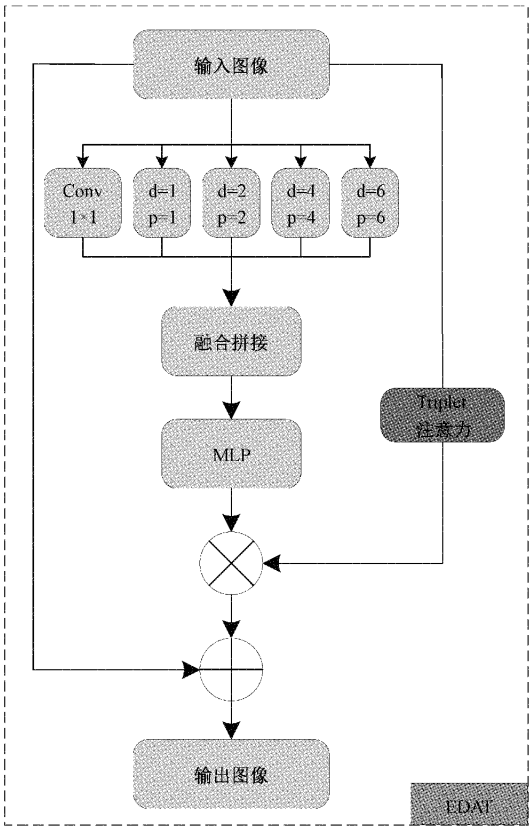


图 7　边缘提取模块结构

首先将特征图输入到一个类 ASPP 模块,将 ASPP 原来的池化层去掉并设置一层扩张率和填充率更大的空洞卷积,该类 ASPP 共包含 4 个空洞卷积层和一个 1×1 卷积层,空洞卷积层的扩张率和填充率分别为 1、2、4、6,卷积核设置为 3,表示在不同尺度上对特征图进行空洞卷积计算,以达到扩大局部感受野的效果。将运算结果进行通道串联以聚合特征图,对聚合后的特征接入批归一化 BN 层、激活函数 ReLU 层以及随机失活 Dropout 层处理,然后通过多层感知机(MLP)调整通道维度并获取深层次语义信息后,可以得到一个特征权重。用公式表示以上操作如下:

$$x_i = MLP(ReLU(BN(Conv(k,d,p)(X)))) \quad (9)$$

其中,Conv 代表空洞卷积,k、d、p 分别代表卷积核大小、扩张率与填充率,BN 则代表归一化层,ReLU 代表激活函数 ReLU,MLP 代表多层感知机。

将该权重与经过 Triplet 注意力引导的原特征图相乘,加入残差连接以防止梯度消失,最后输出融合注意力的边缘提取特征,模块可以帮助模型识别显著变化的建筑物区

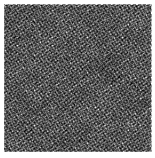
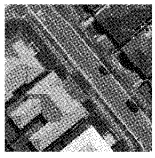
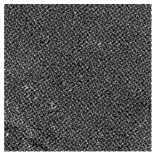
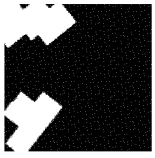
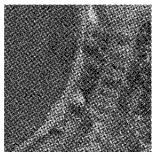
域并且抑制非建筑物变化的伪变化区域。

3　实验过程分析与讨论

3.1　实验数据集

由于公开可用的建筑物变化检测数据集较少，本文采用两种常用的遥感影像数据集去验证网络的改进效果，分别是 WHU-CD 数据集[21]和 LEVIR-CD 数据集[11]，其中 LEVIR-CD 包含更多季节气候、昼夜光照变化的影像。WHU-CD 来自于武汉大学的季顺平教授团队所制作的遥感建筑物数据集，数据来源于新西兰 Christchurch 市的部分地区建筑，其包含两张不同时相的高分辨率遥感影像和一张实际产生变化建筑的标签。数据的预处理采用了滑窗裁剪的方式将原影像裁剪为 256×256 大小，其中包含训练集/验证集/测试集的数量为 3 442/700/2 000。LEVIR-CD 影像数据来源于美国德克萨斯州的部分城市，该数据集包含 637 对大小为 1 024×1 024 的遥感影像，拍摄时间从 2002～2018 年都有涵盖，数据集中发生变化的区域对象主要为该城市的各类型建筑物。考虑到 GPU 显存上限的问题，将该数据集裁剪为 256×256 大小并切分为 3 组数据，训练集/验证集/测试集的数量分别为 7 120/1 024/2 048。其中 WHU-CD 数据集裁剪后如表 1 所示。

表 1　WHU-CD 数据集示例图

序号	T1	T2	Label
1			
2			
3			

3.2　实验环境

本文采用 Pytorch 框架进行模型训练，硬件环境 CPU 为 AMD R5 5600，GPU 为 NVIDIA GeForce 3090，显存 24 G，操作系统为 Windows11，采用 CUDA11.1 和 CuDNN8.2.0 进行加速 GPU 运算。

一般情况下，学习率设置较大可以加快训练，且能避免局部最优解，但是可能会导致模型无法收敛。学习率设置较小可以提高识别精度，但是会导致训练时间过长，容易得到局部最优解。因此，本文采用了余弦退火衰减算法[22]，设置学习率初始值为 0.000 06 并且动态调整，同时优化器设置为 AdamW 来加快模型训练。Epoch 设置为 200 次，batch_size 设置为 8，损失函数采用交叉熵函数（cross-entropy，CE）。

3.3　评估指标

为了有效评定模型性能，本文采用混淆矩阵的方法定量评定精度，通过选取精确率（Precision）、召回率（Recall）、F1、总体精度（Acc）、交并比 IoU 作为评价指标。精确率主要针对预测结果，表示预测样本中预测为正样本且预测正确的像素点占总体预测样本像素点的比值；召回率表示预测正确的像素在真实的正样本中所占的比值；F1 是调和精确率和召回率的平均值，在某些场景下精确率和召回率会出现相互矛盾的情况，因此需要综合两者分数的 F1 来衡量模型的性能；总体精度则是预测正确的像素占总像素的比值；交并比是预测的分类二值图和真实标签作对比；以下分别是五个评价指标的计算公式：

$$\mathrm{Acc}=\frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

$$\mathrm{Precision}=\frac{TP}{TP+FP} \tag{11}$$

$$\mathrm{Recall}=\frac{TP}{TP+FN} \tag{12}$$

$$\mathrm{F1}=2\cdot\frac{\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{13}$$

$$\mathrm{IoU}=\frac{TP}{TP+FP+FN} \tag{14}$$

其中，TP 为预测正确的正样本，FN 为实际标签是负样本预测为正样本，FP 为实际标签是正样本预测为负样本，TN 为预测正确的负样本。

3.4　对比其他网络实验分析

将本文算法与 FC-EF[8]、FC-Siam-conc[8]、FC-Siam-diff[8]、STANet[11]、DTCDSCN[12]、BIT[13]等变化检测经典网络进行比较，其中 FC-EF、FC-Siam-conc 与 FC-Siam-diff 分别为文献[8]的 3 个改进 UNet 网络，将以上 6 个网络与本文算法分别在 WHU 和 LEVIR-CD 数据集上进行训练，各个网络在两个数据集上的具体评价指标分别如表 2 和 3。以 F1 值作为主要评价指标，可以看到本文的改进网络优于其他变化检测网络，在两个数据集上的 F1 值均为最高。在 LEVIR 数据集上，与 FC-EF、FC-Siam-conc 与 FC-Siam-diff 相比 F1 值分别提高约 7.36%、5.15% 和 5.75%，相比 DTCDSCN（dual task constrained deep siamese convolutional network）也提高约 2.11%，说明融入 Swin Transformer 的网络设计能够显著提升变化检测模型的精度。STANet（spatial-temporal attention network）的召回率高而精度低可能是对于图像中的细节过多误检，将部分负样本判断为正样本。而在 WHU 数据集上，本文的网络设计对比其他网络仍然具有一定优势，

在 F1、召回率以及精确度等指标都有提升。在 F1 指标上,相比 FC-EF 提升大约 20%,相比 DTCDSCN 也有约为 7.93%的提升。本文设计的网络与 FC-EF 等经典变化检测网络相比都有较大幅度性能提升,有效缓解了小目标漏检、建筑物形状不规则以及边缘粗糙等现象。表 2 和 3 分别为实验模型在 LEVIR 数据集和 WHU 数据集上的预测结果对比。

表 2 不同网络模型在 LEVIR 数据集上的预测结果对比

模型	Accuracy/%	F1/%	precision/%	Recall/%	IoU/%
FC-EF	98.46	84.18	88.52	82.36	72.67
DTCDSCN	98.94	89.43	91.08	87.84	88.74
FC-Siam-conc	98.64	86.39	86.84	84.54	76.04
FC-Siam-diff	98.60	85.79	88.73	83.10	75.12
STANet	98.30	84.40	78.80	**90.08**	72.90
BIT	98.99	89.94	91.57	88.38	82.25
Ours	**99.16**	**91.54**	**93.60**	89.56	**84.40**

表 3 不同网络模型在 WHU 数据集上的预测结果对比

模型	Accuracy/%	F1/%	precision/%	Recall/%	IoU/%
FC-EF	97.33	73.21	64.48	84.68	57.75
DTCDSCN	98.63	84.95	85.22	88.23	73.83
FC-Siam-conc	97.87	78.30	71.93	88.52	64.34
FC-Siam-diff	98.11	80.35	73.65	87.64	66.72
STANet	99.00	73.30	63.60	86.40	57.80
BIT	98.98	88.11	88.87	87.37	78.75
Ours	**99.39**	**92.88**	**94.21**	**91.59**	**86.71**

图 8 展现了本文算法与其他经典变化检测网络结果比较的视觉效果。从图中实验结果来看,文献[5]的 3 个网络与真实值相差较大,变化区域中出现空洞和不规则的地方较多,也有误判为变化建筑物的部分。而本文算法的结果边缘更为平滑规则,基本上没有出现空洞区域和误检区域,相比其他网络更接近真实标签所呈现的二值图。

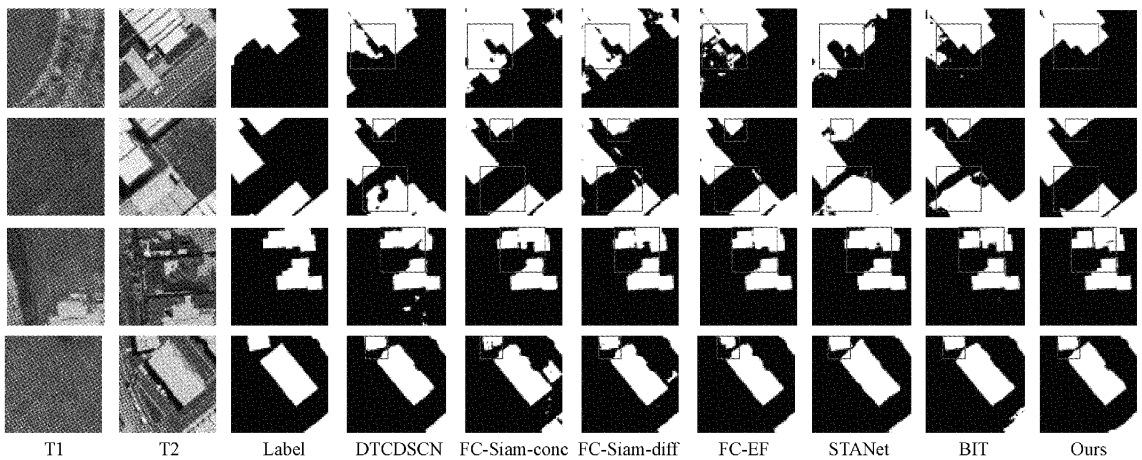


图 8 WHU-CD 数据集上不同方法变化检测结果比较

图 9 展现了 LEVIR-CD 数据集上所训练的 7 个模型预测结果。从图中结果来看,虽然其他经典变化检测网络的预测结果已经比较接近真实标签,但是仍会出现建筑物变化区域之间相互粘连,每个独立变化区域没有清晰地分离,本文算法所预测结果展示了变化建筑的每一个独栋个体,基本没有出现粘连的情况。从图中观察得到,变化区域的分割结果平整规则且边界清晰,是最接近真实标签的预测结果。

对比图 8 与 9 上不同数据集的结果比较,在某些样例中本文的方法对于边缘的提取更为准确、更细节化,对于检测发生变化的建筑物边缘更平滑。这是由于加入的 EDAT 对边缘提取进行了优化,扩大局部感受野来增强不同尺度上的建筑物边缘细节。本文基于 DFAM 的多尺度融合更能捕捉到变化检测过程中建筑物的语义信息,在两个数据集上都取得了更为优秀的表现,更接近真实标签所呈现的二值图。

3.5 消融实验

将本文所提方法中的 DFAM 与 EDAT 去除后的网络称为基线网络(baseline),为了验证两个模块对网络改进的有效性,本文在 WHU 数据集上针对两个模块进行消融实验。本文经过实验发现加入上述的两个模块都可以有效提升该网络的检测精度。表 4 为本文所提算法网络的消融实验结果对比。

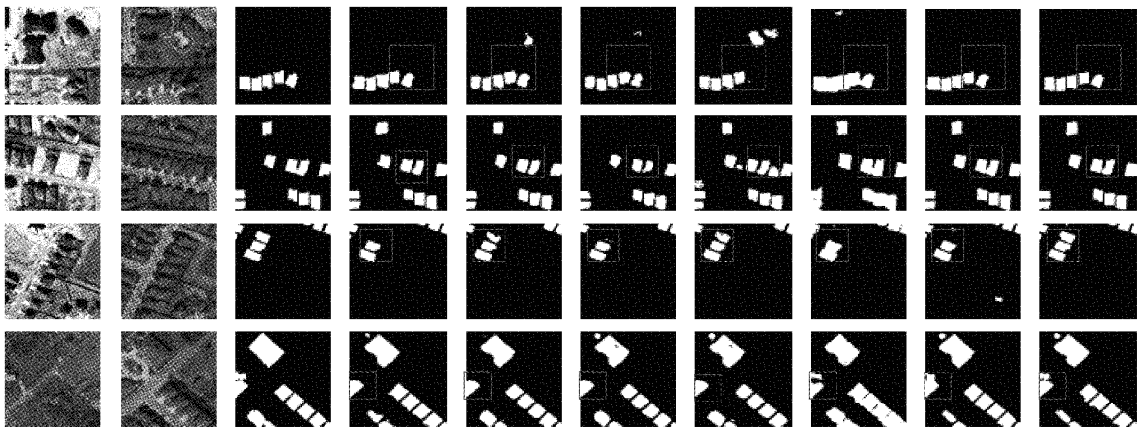


图 9　LEVIR-CD 数据集上不同方法变化检测结果比较

表 4　消融实验结果对比

DFAM	EDAT	Accuracy/%	F1/%	precision/%	Recall/%	IoU/%
×	×	98.84	86.68	86.68	87.75	76.49
√	×	99.22	87.43	92.72	88.85	83.06
×	√	99.29	91.71	92.77	90.77	84.69
√	√	**99.39**	**92.88**	**94.21**	**91.59**	**86.71**

4　结　　论

本文基于孪生网络和 Swin Transformer 设计了一个新的变化检测网络。该方法通过将 Swin Transformer 作为主干网络来实现卷积神经网络对于图像的多尺度特征提取，同时弥补了传统变化检测网络中对全局上下文信息缺少关注的问题。通过将 EDAT 嵌入到 Swin Transformer 来提升局部的感受野，使得网络对变化检测过程中的建筑物边缘特征更为敏感，从而提高网络的精度。此外还基于 iAFF 结构设计了一个多尺度、多层级的特征融合模块，可以将不同的 Swin Transformer Block 所输出的特征图经过注意力机制分配特征权重后融合其中的语义信息。经过两个模块对网络的优化，该网络兼顾了全局特征和局部特征的信息，针对建筑物所完成的变化检测分割精度也得以进一步提高。

实验结果表明，本文所设计的网络各方面性能要优于其余网络，意味着算法智能在城市遥感、土地资源规划等方面的进一步应用。实际应用中建筑物变化检测还面临着数据集难以收集、处理、标注等问题，后续本文将开展建筑物变化检测领域的弱监督或无监督工作。

参考文献

[1]　眭海刚，冯文卿，李文卓，等. 多时相遥感影像变化检测方法综述[J]. 武汉大学学报(信息科学版)，2018，43(12)：1885-98.

[2]　任秋如，杨文忠，汪传建，等. 遥感影像变化检测综述[J]. 计算机应用，2021，41(8)：2294-305.

[3]　高桂荣，严威，夏晨阳等. 结合空间信息的 PTSVM 的遥感图像变化检测[J]. 电子测量技术，2016，39(4)：45-48，52.

[4]　张永梅，李立鹏，姜明，等. 综合像素级和特征级的建筑物变化检测方法[J]. 计算机科学，2013，40(1)：286-93.

[5]　施文灶，毛政元. 基于图分割的高分辨率遥感影像建筑物变化检测研究[J]. 地球信息科学学报，2016，18(3)：423-32.

[6]　张志强，张新长，辛秦川，等. 结合像元级和目标级的高分辨率遥感影像建筑物变化检测[J]. 测绘学报，2018，47(1)：102-12.

[7]　潘伟豪，徐赛博，郭弘扬，等. 基于 D-S 证据理论的高分遥感影像建筑物变化检测[J]. 电子测量与仪器学报，2022，36(8)：194-203.

[8]　DAUDT R C，LE SAUX B，BOULCH A，et al. Fully convolutional siamese networks for change detection [C]. 2018 25th IEEE International Conference on Image Processing (ICIP)，Athens，Greece，2018：4063-4067.

[9]　张翠军，安冉，马丽. 改进 U-Net 的遥感图像中建筑物变化检测[J]. 计算机工程与应用，2021，57(3)：239-46.

[10]　朱节中，陈永，柯福阳，等. 基于 Siam-UNet++的高分辨率遥感影像建筑物变化检测[J]. 计算机应用研究，2021，38(11)：3460-5.

[11]　CHEN H，SHI Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection[J]. Remote Sensing，2020，12(10)：1662.

[12]　LIU Y，PANG C，ZHAN Z，et al. Building change detection for remote sensing images using a dual-task

constrained deep siamese convolutional network model[J]. IEEE Geosci Remote Sens Lett, 2020, 18(5): 811-5.

[13] CHEN H, QI Z, SHI Z. Remote sensing image change detection with transformers [J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-14.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30, DOI: 10.48550/arXiv.1706.03762.

[15] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. ArXiv Preprin, 2020, ArXiv: 2010.11929.

[16] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. 2021 IEEE/CVF International Conference on Computer Vision(ICCV), 2021: 9992-10002.

[17] DAI Y, GIESEKE F, OEHMCKE S, et al. Attentional feature fusion[C]. 2021 IEEE Winter Conference on Applications of Computer Vision(WACV), 2021: 3559-3568, DOI: 10.1109/WACV48630.2021.00360.

[18] MISRA D, NALAMADA T, ARASANIPALAI A U, et al. Rotate to attend: Convolutional triplet attention module [C]. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021: 3138-3147, DOI: 10.1109/WACV48630.2021.00318.

[19] CHEN Z, ZHOU Y, WANG B, et al. EGDE-Net: A building change detection method for high-resolution remote sensing imagery based on edge guidance and differential enhancement [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 191: 203-22.

[20] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]. European Conference on Computer Vision(ECCV), 2018: 801-818.

[21] JI S, WEI S, LU M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57: 574-86.

[22] LOSHCHILOV I, HUTTER F. Sgdr: Stochastic gradient descent with warm restarts [J]. ArXiv Preprint, 2016, ArXiv: 1608.03983.

作者简介

于政尧,硕士研究生,主要研究方向为深度学习、图像处理等。

E-mail: 1316003943@qq.com

黄建华(通信作者),博士,高级工程师,主要研究方向为地理信息系统、卫星导航、卫星遥感理论、技术研究,主要致力于带研究生团队开展地理空间智能、数字孪生、自然资源视频图像数据监测、北斗应用等。

E-mail: 810175926@qq.com