融合多尺度特征Transformer的高分辨率遥感图像变化检测

李健慷[1]，张桂欣[2]，祝善友[1]，徐永明[1]，李湘雨[1]

1. 南京信息工程大学 遥感与测绘工程学院, 南京 210044;
2. 南京信息工程大学 地理科学学院, 南京 210044

摘 要：为了加强变化检测中深度学习网络的语义信息提取能力，捕获更多高阶多尺度特征细节以及突出影像差异信息，本文提出一种融合孪生结构和多尺度特征Transformer的高分辨率遥感影像变化检测模型MFTSNet（Multi-scale Feature Transformer Siamese Network）。该模型设计了语义特征Transformer模块ST（Semantic feature Transformer module）捕获不同层级特征图的语义信息，引入置入Transformer模块GT（Grounding Transformer module）和映射Transformer模块RT（Rendering Transformer module）加强低层和高层语义信息的获取，发掘高阶多尺度特征细节信息以及不同空间位置和通道间的全局上下文关系，进一步提升变化检测精度，增强地物检测结果的完整性、区域内部以及边缘细节。将MFTSNet与另外8种变化检测模型在4个公开数据集上的变化检测结果进行对比，并通过消融实验、参数分析等手段验证MFTSNet中各模块的有效性。对比实验结果表明MFTSNet网络模型在4个数据集上的F1和交并比IoU分别至少提高了0.465%、0.113%、0.369%、2.13%和0.723%、0.188%、0.304%、2.962%。消融实验表明GT、RT、ST 3个模块共同作用可有效提升网络模型性能。参数分析表明MFTSNet模型中的特征信息长度L与编码器—解码器个数是两个重要的网络结构参数，L在CDD、WHU-CD数据实验中取16，在SYSU-CD、LEVIR-CD数据实验中取8，4个数据集上设置（E_N，D_N）为（1，2）时，MFTSNet模型的检测结果最优。

关键词：高分辨率遥感，变化检测，深度学习，孪生网络，多尺度特征，Transformer，语义信息，消融实验

中图分类号：TP751.1/P2

引用格式：**李健慷,张桂欣,祝善友,徐永明,李湘雨.2025.融合多尺度特征Transformer的高分辨率遥感图像变化检测.遥感学报,29(1): 266-278**

Li J K, Zhang G X, Zhu S Y, Xu Y M and Li X Y. 2025. Change detection for high-resolution remote sensing images with multi-scale feature transformer. National Remote Sensing Bulletin, 29(1):266-278[DOI:10.11834/jrs.20243201]

1 引 言

遥感图像变化检测是指通过图像处理等手段对不同时期同一区域的遥感图像进行对比分析，以判断图像间的变化（佟国峰 等，2015）。根据研究对象不同，传统的变化检测方法可分为基于像素、基于特征和基于对象的变化检测方法（廖明生 等，2000；Hussain等，2013；佃袁勇 等，2016；赵敏和赵银娣，2018）。这些方法用于高空间分辨率遥感图像变化检测时，会受到传感器观测角度、地物阴影等因素的共同影响，导致光谱信息可分性减弱，在精度、自动化或普适性方面不能满足大量复杂应用场景的需求。基于深度学习的变化检测方法无需构建特征工程，检测精度和效率均有所提高，已发展为现阶段遥感变化检测领域的研究热点。基于深度学习的遥感变化检测方法通常利用神经网络模型提取遥感图像之间的深度差异特征，并在学习策略指导下训练变化检测模型，根据检测方案输出结果（杨彬 等，2023）。

收稿日期：2023-06-14；**预印本：**2024-01-11

基金项目：国家自然科学基金(编号:42171101, 42271351);高分辨率对地观测系统重大专项(编号:30-Y60B01-9003-22/23)

第一作者简介：李健慷,研究方向为遥感图像深度学习变化检测。E-mail：jkang_li@163.com

通信作者简介：祝善友,研究方向为热红外遥感基础理论与应用、生态遥感。E-mail：zsyzgx@163.com

在变化检测中，所使用的深度学习网络模型主要包括卷积神经网络、自动编码器、生成对抗网络、循环神经网络等。其中，孪生神经网络结构能够分别学习不同时相的影像特征，并对不同时相影像的层级特征进行融合，模型更加容易学习不同尺度的变化信息。Zhang等（2018）提出了一种基于孪生卷积神经网络的方法，利用不同时相影像训练孪生卷积神经网络以提取变化特征。Hughes等（2018）提出了一种基于伪孪生卷积神经网络的方法，将其应用于SAR影像和光学影像的变化检测。Daudt等（2018）以U-Net模型为特征提取网络，提出用于变化检测的孪生全卷积网络结构。Chen等（2021）提出了双注意力孪生网络结构，有效捕获长期依赖关系，进而获得更加具有区分度的特征表示。Chen等（2022）利用Transformer编码—解码器模型，结合语义特征构建变化检测网络。

在深度学习网络模型训练过程中，许多学者利用多尺度特征提取手段增强不同对象和区域间的特征及其差异。Liu等（2022）提出多尺度上下文聚合模块融合3个尺度信息，实验表明该方法能够有效提高网络模型语义表达能力。Guo等（2022）提出并行卷积结构集成时间序列图像的多尺度特征，突出变化区域特征，提高变化检测精度。Song等（2022a）等提出多尺度的Swin Transformer网络模型，从不同尺度特征图获取空间信息，提高网络模型性能。

综合分析国内外有关基于深度学习方法的遥感图像变化检测研究，可以发现变化检测模型在一定程度上存在着语义信息提取不足、多尺度细节特征丢失以及变化检测结果边界和内部细节不完整等问题，需要进一步开展深入研究。本文提出一种融合孪生结构和多尺度特征Transformer的高分辨率遥感影像变化检测模型MFTSNet（Multi-scale Feature Transformer Siamese Network），利用语义特征Transformer模块ST（Semantic Transformer）捕获双时相影像的语义信息，将浅层特征进行细化，实现语义信息跨空间交互，捕捉不同位置间的关联信息，增强模型特征表达能力；利用置入Transformer GT（Grounding Transformer）模块将浅层特征逐步细化，通过跨尺度信息交互，生成具有多尺度信息的高级语义特征；设计映射Transformer RT（Rendering Transformer）模块实现深层次特征高效表达，抑制背景无关特征；最后，利用浅层CNN网络结构生成变化检测结果。

2 原理与方法

2.1 孪生神经网络

孪生神经网络（Siamese Neural Network）由两个分支网络组成，两个分支网络分别对输入数据进行编码，编码特征前向传播时共享参数，生成深层次高阶目标特征，然后对高阶特征进行特征相似性比较。

2.2 Transformer基本结构

Transformer基本结构由4个部分组成，如图1所示（Vaswani，2017），分别是特征嵌入、多头注意力、位置前馈网络和位置编码。

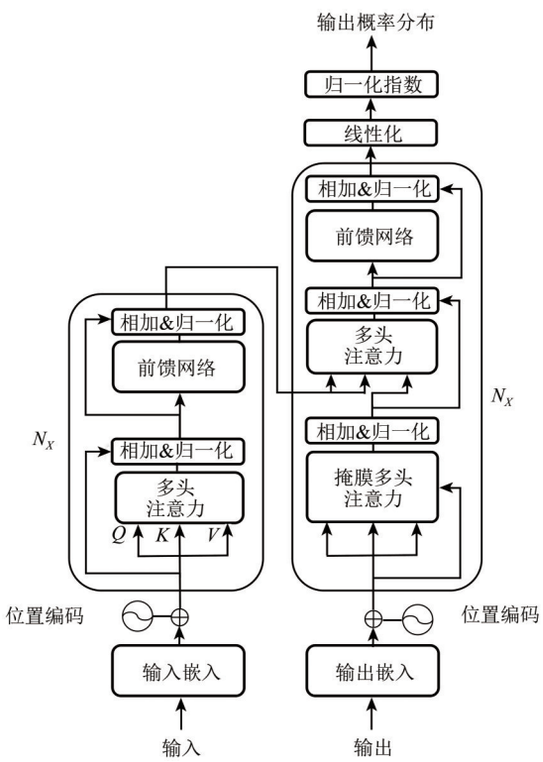


图1 Transformer结构图

Fig. 1 Illustration of Transformer architecture

自注意力将输入图像$\boldsymbol{X}\in\mathbb{R}^{n_x\times d_x}$，$\boldsymbol{Y}\in\mathbb{R}^{n_y\times d_y}$，映射为3个不同序列的向量（Query：$\boldsymbol{Q}$、Key：$\boldsymbol{K}$、Value：$\boldsymbol{V}$），该向量的值为RGB 3个通道的像素值，其中$n$和$d$分别为输入序列长度和维数。每个向量计算方式如下：

$$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}^{Q},\boldsymbol{K}=\boldsymbol{Y}\boldsymbol{W}^{K},\boldsymbol{V}=\boldsymbol{Y}\boldsymbol{W}^{V} \tag{1}$$

式中，$\boldsymbol{W}^{Q}\in\mathbb{R}^{d_x\times d_k}$，$\boldsymbol{W}^{K}\in\mathbb{R}^{d_y\times d_k}$，$\boldsymbol{W}^{V}\in\mathbb{R}^{d_y\times d_v}$分别是线

性权重矩阵。$\boldsymbol{Q}$由$\boldsymbol{X}$映射所得，$\boldsymbol{K}$、$\boldsymbol{V}$由$\boldsymbol{Y}$映射所得。

输入$\boldsymbol{X}$与$\boldsymbol{Y}$相同时被称为自注意力机制，$\boldsymbol{X}$与$\boldsymbol{Y}$不同的注意力被称为交叉注意力机制。自注意力机制被用于Transformer模块中的编码和解码过程，而交叉注意力机制仅在解码过程中起到连接作用。

利用向量$\boldsymbol{K}$、$\boldsymbol{Q}$和$\boldsymbol{V}$生成输出特征的过程如下式：

$$\mathrm{Atten}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\cdot\boldsymbol{V} \tag{2}$$

式中，$\boldsymbol{K}^{\mathrm{T}}$表示$\boldsymbol{K}$的转置，比例因子$\sqrt{d_k}$与Softmax函数将注意权重标准化分布。

为解决特征子空间大小限制、单个注意力全局建模能力差的问题，多头注意力机制MHA被应用于Transformer中。MHA将输入的线性特征映射到多个特征子空间中，并通过多个注意力头并行处理成向量，最后将结果向量拼接后输出。该过程可以表达为

$$\boldsymbol{O}_i = \mathrm{Atten}\left(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i\right), i = 1,\cdots,h \tag{3}$$

$$\mathrm{MHA}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right) = \mathrm{Concat}(\boldsymbol{O}_1,\boldsymbol{O}_2,\cdots,\boldsymbol{O}_h)\cdot\boldsymbol{W}^{O} \tag{4}$$

式中，h表示注意力头数量；$\boldsymbol{W}^{O}\in\mathbb{R}^{hd_v\times d_{\mathrm{model}}}$表示输出的投影矩阵，$\boldsymbol{Q}_i$表示每个注意力头的输出向量。多头注意力将输入特征分为h个，每个注意力头获取d_{model}/h维独立向量并行组成头部特征。

多头注意力机制MHA生成的特征图，经过两个连续的线性变换和ReLU激活函数，生成前馈网络的特征图，其表达式如下：

$$\mathrm{FFN}(\boldsymbol{x}) = \mathrm{ReLU}\left(\boldsymbol{W}_1\boldsymbol{x}+\boldsymbol{b}_1\right)\boldsymbol{W}_2+\boldsymbol{b}_2 \tag{5}$$

式中，$\mathrm{FFN}(\boldsymbol{x})$为前馈网络输出的特征图。

Transformer处理输入图像时不会保留图像块序列的位置信息，为了能够利用位置信息，常常在输入端附加一个额外的位置向量，称为位置编码，其表达式为

$$\mathbf{PE}_t^{(i)} = \begin{cases}\sin\left(w_i t\right), & \text{if } k = 2i\\ \cos\left(w_i t\right), & \text{if } k = 2i+1\end{cases} \tag{6}$$

$$w_i = \frac{1}{10000^{2i/d_{\mathrm{model}}}},\ i = 0,1,\cdots,\frac{d_{\mathrm{model}}}{2}-1 \tag{7}$$

式中，t代表元素的实际位置，$\mathbf{PE}_t\in\mathbb{R}^d$是该元素的位置向量，$\mathbf{PE}_t^{(i)}$是该位置向量中的第$i$个元素，$d_{\mathrm{model}}$是该元素的维度。

2.3 MFTSNet整体网络框架

本文提出的MFTSNet网络模型结构如图2所示。MFTSNet模型主要由基于孪生结构的特征提取网络、3种不同尺度与不同层级的多尺度特征Transformer模块（ST、GT和RT）以及浅层CNN结构的预测输出模块组成。

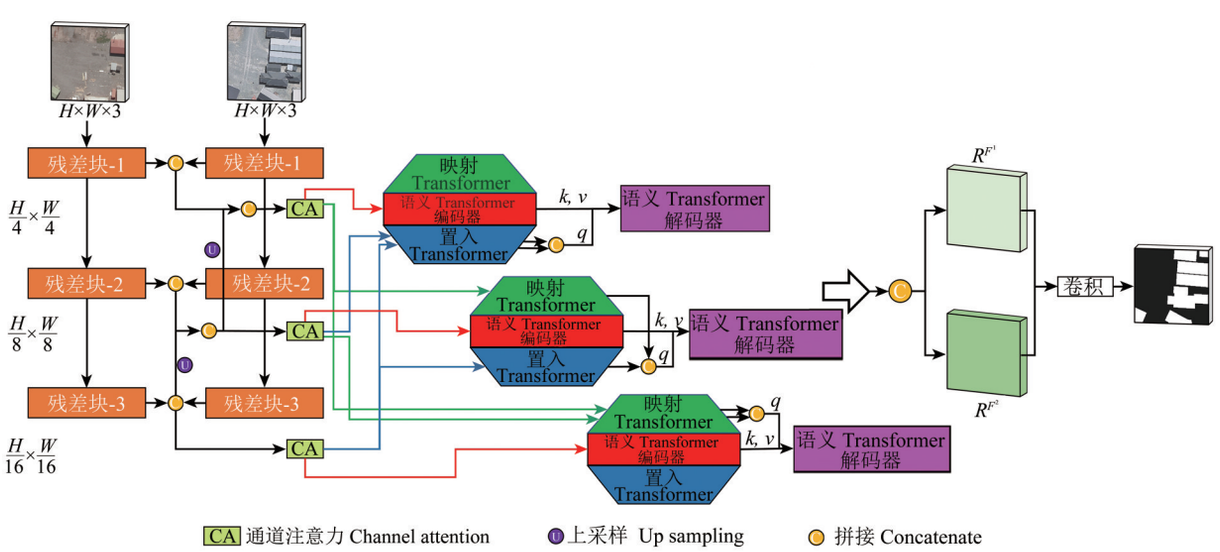


图2 MFTSNet模型结构图

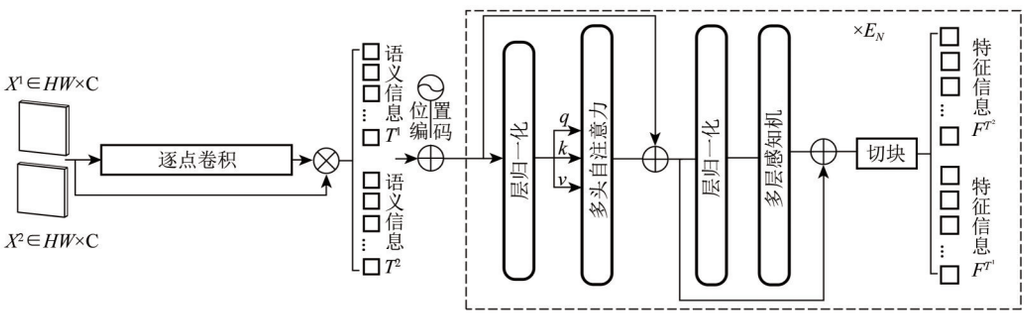
Fig. 2 Illustration of MFTSNet architecture

根据图2，在孪生结构的特征提取阶段，利用ResNet-18的前3层提取原始图像特征，输出特征图大小分别为输入图像大小（$H\times W$）的$\left\{\frac{1}{4},\frac{1}{8},\frac{1}{16}\right\}$。输出的特征图通过上采样与低层双时相特征图拼接，通道注意力给拼接后的特征

赋予通道权重。不同尺度的融合特征图被送入多尺度特征Transformer模块中，在GT模块和RT模块中进行跨尺度信息交互，在ST中模块中进行跨空间信息交互，经ST解码后得到精细化的双时相特征图。最后，双时相特征作差后，利用浅层CNN结构进行判别，生成最终变化检测结果。

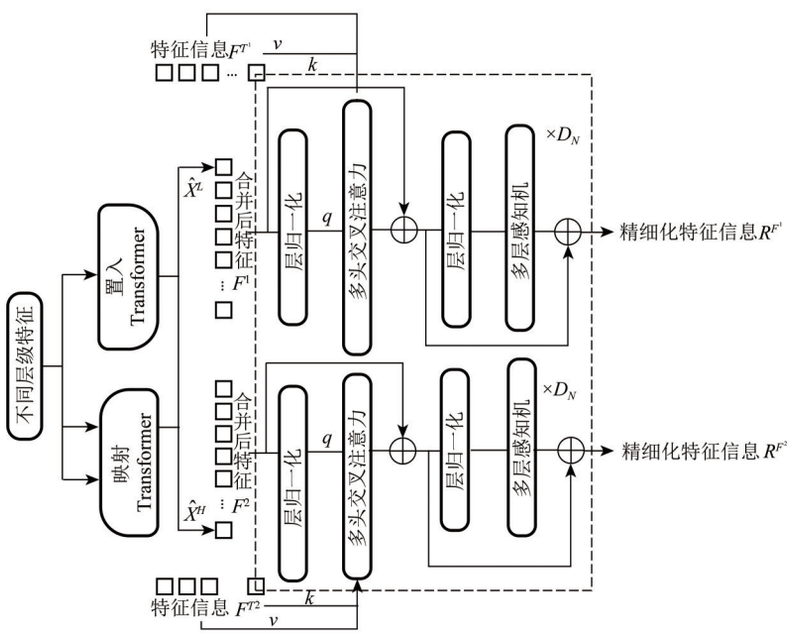
2.4 多尺度特征Transformer模块

目前，变化检测模型通常以上采样等方法将不同层级特征通过拼接的形式，链接图像深层信息和浅层信息，但在特征交互阶段，这种操作可能会导致特征信息丢失，同时会生成一些不真实的图像信息，从而在模型训练的特征交互过程中产生一些负面影响。本文利用Transformer交叉注意力和自注意力结构，根据Transformer基本原理和Feature Pyramid Transformer（Zhang等，2020）结构，设计了3种多尺度特征Transformer模块，分别是跨尺度交互的置入Transformer模块（GT）、映射Transformer模块（RT）和跨空间交互的语义特征Transformer模块（ST）。多尺度特征Transformer模型MFT（Multi-scale Feature Transformer）结构如图3所示。



（a）语义特征Transformer编码

（a）Semantic transformer encoder



（b）语义特征Transformer解码

（b）Semantic transformer decoder

图3 多尺度特征Transformer模型结构图

Fig. 3 Illustration of multi-scale feature Transformer architecture

2.4.1 置入Transformer（GT）模块

GT模块通过局部交互，将高级特征中的信息$\boldsymbol{k}_j$和$\boldsymbol{v}_j$通过欧几里得距离、混合归一化概率函数以及相乘的形式映射到低级特征$\boldsymbol{q}_i$中，通过度量相似

度信息，让模型更加关注变化信息，从而提高变化检测的准确性。图像特征分析过程中，不同尺度特征图包含了不同层次的语义信息，可利用欧几里得距离来定量表达两个特征图之间的相似度，其表达式为

$$\mathrm{F}_{\mathrm{eud}}\left(\boldsymbol{q}_i,\boldsymbol{k}_j\right)=-\left\|\boldsymbol{q}_i-\boldsymbol{k}_j\right\|^2 \tag{8}$$

式中，$\boldsymbol{q}_i=\boldsymbol{Q}(\boldsymbol{X}_i^L)$，$\boldsymbol{k}_j=\boldsymbol{K}(\boldsymbol{X}_j^H)$，$\boldsymbol{X}_i^L$是低级特征图$\boldsymbol{X}^L$第$i$个位置上的值，$\boldsymbol{X}_j^H$是高级特征图$\boldsymbol{X}^H$第$j$个位置的值。

Softmax函数通常用于网络模型的类别概率判别归一化。本文将Yang等（2018）基于Softmax提出的MoS函数引入GT模块中。基于MoS的归一化函数表示如下：

$$\mathrm{F}_{\mathrm{MoS}}\left(\boldsymbol{q}_{i,n},\boldsymbol{k}_{j,n}\right)=\sum_{n=1}^{N}\pi_n\frac{\exp\left(\boldsymbol{q}_{i,n},\boldsymbol{k}_{j,n}\right)}{\sum_j\exp\left(\boldsymbol{q}_{i,n},\boldsymbol{k}_{j,n}\right)} \tag{9}$$

式中，$\mathrm{F}_{\mathrm{MoS}}$可理解为将$\boldsymbol{q}_i$和$\boldsymbol{k}_j$分成$N$份，然后计算每一份中每一对$\boldsymbol{q}$，$\boldsymbol{k}$的相似度$\left(\boldsymbol{q}_{i,\ n}\cdot\boldsymbol{k}_{j,\ n}\right)$，$\pi_n$是第$n$个权值，可通过如下公式计算：

$$\pi_n=\mathrm{Softmax}\left(\boldsymbol{w}_n^{\mathrm{T}}\cdot\bar{k}\right) \tag{10}$$

式中，$\boldsymbol{w}_n$是一个可学习的线性归一化向量，$\bar{k}$是所有位置$\boldsymbol{k}_j$的平均值。综合式（9）和式（10）可得GT模块特征图的表达式：

$$\hat{\boldsymbol{X}}_i^L=\mathrm{mul}\left(\mathrm{F}_{\mathrm{MoS}}\left(\mathrm{F}_{\mathrm{eud}}\left(\boldsymbol{q}_{i,n}\cdot\boldsymbol{k}_{j,n}\right)\right),\boldsymbol{v}_j\right) \tag{11}$$

式中，$\boldsymbol{v}_j=\boldsymbol{V}(\boldsymbol{X}_j^H)$；$\hat{\boldsymbol{X}}_i^L$是$\hat{\boldsymbol{X}}^L$中第$i$个位置上转换后的特征。GT模块跨尺度交互时，低层特征图$\boldsymbol{q}_i$与高层特征图$\boldsymbol{k}_j$和$\boldsymbol{v}_j$的一部分相互作用。由于高层特征图比低层特征图维度更小，特征图边缘区域不参与交互过程。对于$\boldsymbol{k}_j$和$\boldsymbol{v}_j$超过索引的位置，用0值代替。

2.4.2 映射Transformer（RT）模块

RT模块采用由下向上的方式实现局部特征交互，输出维度大小与高级特征映射相同。利用低级特征中的信息来表征高级特征中的信息，被转换后的特征图调整到相应特征图大小，并与原始图像拼接。整个过程中，RT模块不是逐像素执行，而是按当前特征图的整个特征执行。高级特征图$\boldsymbol{X}^H$上的局部信息为$\boldsymbol{X}^Q$，低级特征图对应局部信息为$\boldsymbol{X}^K$和$\boldsymbol{X}^V$，$\boldsymbol{X}^K$和$\boldsymbol{X}^Q$之间以通道注意力进行交互以突出目标；$\boldsymbol{X}^K$通过全局平均池化（GAP）后和$\boldsymbol{X}^Q$外积得到$\boldsymbol{X}^{\mathrm{Q}_{\mathrm{att}}}$，将$\boldsymbol{X}^V$经过步长为3的卷积减小特征尺度后与经过3×3卷积后的精细化$\boldsymbol{X}^Q$相加；最后，对生成的特征进行3×3卷积细化处理。表达式为

$$\boldsymbol{X}^{\mathrm{Q}_{\mathrm{att}}}=\boldsymbol{X}^Q\times\mathrm{GAP}(\boldsymbol{X}^K) \tag{12}$$

$$\boldsymbol{X}^{\mathrm{V}_{\mathrm{down}}}=\mathrm{sconv}(\boldsymbol{X}^V) \tag{13}$$

$$\hat{\boldsymbol{X}}^H=\mathrm{conv}\left(\mathrm{Add}\left(\mathrm{conv}\left(\boldsymbol{X}^{\mathrm{Q}_{\mathrm{att}}}\right),\boldsymbol{X}^{\mathrm{V}_{\mathrm{down}}}\right)\right) \tag{14}$$

式中，“×”代表向量积，sconv代表步长为3的卷积，当步长为1时，$\boldsymbol{X}^V$和$\boldsymbol{X}^Q$的尺度相同，conv代表卷积核大小为3×3的卷积，Add代表矩阵元素相加，$\hat{\boldsymbol{X}}^H$表示经过RT模块后输出的特征图。

2.4.3 语义特征Transformer（ST）模块

ST模块由STE编码与STD解码2个部分组成。特征提取模块输出的双时相特征图通过逐点卷积操作，然后与双时相特征进行逐像素相乘，生成总长度为L的双时相语义信息标记$\boldsymbol{T}^i$（i=1，2），每个时相的语义信息标记长度相等，记为l。表达式为

$$\boldsymbol{T}^i=\mathrm{mul}\left(\left(\sigma\left(\rho(\boldsymbol{X}^i,\boldsymbol{W})\right)\right)^{\mathrm{T}}\right)\ i=1,2 \tag{15}$$

式中，$\sigma(\cdot)$表示Softmax函数，$\rho(\cdot)$表示逐点卷积，$\boldsymbol{W}$表示逐点卷积过程中的可学习卷积核，$\boldsymbol{X}^i$表示双时相特征图。

Transformer编码器可充分利用特征间的全局语义关系，在特征信息间进行跨空间的上下文语义信息建模。生成的双时相语义标记T^i经位置编码后输入STE中可捕获丰富的语义信息。STE由E_N个多头注意力模块和多层感知机组成，并利用前范数残差单元进行层归一化（Dosovitskiy等，2021）。综合式（1）—（4），STE中多头注意力的形式可表示为

$$\begin{cases}\boldsymbol{Q}=T^{(N_L-1)}\boldsymbol{W}^Q\\ \boldsymbol{K}=T^{(N_L-1)}\boldsymbol{W}^K\\ \boldsymbol{V}=T^{(N_L-1)}\boldsymbol{W}^V\end{cases} \tag{16}$$

$$\boldsymbol{O}_j=\mathrm{Atten}(T^{(N_L-1)}\boldsymbol{W}_j^Q,T^{(N_L-1)}\boldsymbol{W}_j^K,T^{(N_L-1)}\boldsymbol{W}_j^V) \tag{17}$$

$$\mathrm{MHSA}(T^{(N_L-1)})=\mathrm{Concat}(\boldsymbol{O}_1,\boldsymbol{O}_2,\cdots\boldsymbol{O}_h)\boldsymbol{W}^O \tag{18}$$

式中，在每层N_L计算语义标记$T^{(N_L-1)}\in\mathbb{R}^{2L\times C}$的$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$，$\boldsymbol{W}^Q$、$\boldsymbol{W}^K$、$\boldsymbol{W}^V$、$\boldsymbol{W}^O$是可学习线性参数，$h$是多头注意力的头部数量，$C$是通道维度。

多层感知机由两个线性变换层组成，输入维度为C，隐藏层采用GELU函数进行激活，输出维度为$2C$，具体表达式为

$$\mathrm{MLP}(T^{(N_L-1)})=\mathrm{GELU}(T^{(N_L-1)}\boldsymbol{W}_1)\boldsymbol{W}_2 \tag{19}$$

式中，$\boldsymbol{W}_1\in\mathbb{R}^{2C\times C}$，$\boldsymbol{W}_2\in\mathbb{R}^{2C\times C}$是多层感知机线性

可学习矩阵。经STE编码后生成特征标记F^{T^1}与F^{T^2}。特征标记作为新的键值与GT模块、RT模块生成的新查询值进行特征合并，在STD中进行跨空间的信息交互。

特征标记中具有丰富的上下文高级语义信息，有利于变化信息的表征。同时，解码器将高级语义信息映射回原始像素空间上获得像素级别特征。STD由D_N多头交叉注意力和多层感知机组成。多层感知机形式与STE中形式相同，STD中多头交叉注意力的形式可表示为

$$\boldsymbol{O}_j = \mathrm{Atten}\left(F^i\boldsymbol{W}_j^Q, F^{T^i}\boldsymbol{W}_j^K, F^{T^i}\boldsymbol{W}_j^V\right) \quad i = 1,2 \tag{20}$$

$$\mathrm{MHCA}\left(F^i, F^{T^i}\right) = \mathrm{Concat}(\boldsymbol{O}_1, \boldsymbol{O}_2, \cdots, \boldsymbol{O}_h)\boldsymbol{W}^O \tag{21}$$

经STD解码后生成像素级精细化特征$\boldsymbol{R}^{F^1}$与$\boldsymbol{R}^{F^2}$。最终，利用浅层CNN结构，对双时相精细化特征差异结果进行变化辨别，生成模型变化检测结果。具体形式可表示为

$$\mathrm{FDI} = \left|\boldsymbol{R}^{F^1} - \boldsymbol{R}^{F^2}\right| \tag{22}$$

$$\mathrm{PM} = \mathrm{Argmax}\left(\sigma\left(\phi\left(\mathrm{FDI}\right)\right)\right) \tag{23}$$

式中，FDI为双时相影像细化特征相应元素的差值结果，PM为预测变化检测结果图，$\phi(\cdot)$为2个3×3卷积组成的变化分类器。

3 变化检测数据集

为了评估本文模型MFTSNet的变化检测性能，在CDD、LEVIR-CD、SYSU-CD、WHU-CD这4个公开数据集上进行实验，4个数据集的采集传感器、空间分辨率、图像场景、时间跨度、主要变化类别等情况各不相同，能够全面、广泛地检验网络模型的变化检测性能。

3.1 CDD数据集

CDD（Change Detection Dataset）数据集包括11对忽略季节变化影响的遥感影像（Lebedev等，2018），样本中的主要变化对象有道路、建筑物、汽车等。影像均来源于Google Earth，空间分辨率为0.03—1 m，其中7对影像大小为4725×2200像素，其余影像对大小为1900×1000像素。按照256×256像素大小对影像进行裁剪和扩充，共得到15998对样本，其中10000对作为训练样本，3000对作为验证样本，2998对作为测试样本。

3.2 LEVIR-CD数据集

LEVIR-CD（Learning、Vision and Remote Sensing）数据集是由北京航空航天大学图像处理中心团队开源的建筑物变化数据集（Chen等，2020）。数据集原始影像从Google Earth获取，空间分辨率为0.5 m，涵盖了2002年—2018年间美国德克萨斯州20个不同地点，影像大小为1024×1024像素，其中训练集445对、验证集64对、测试集128对。本文按照重叠率为0、大小为256×256像素的滑动窗口裁剪原始影像，共得到训练样本7120张，验证样本1024张，测试样本2048张。

3.3 SYSU-CD数据集

SYSU-CD（Sun Yat-Sen University Dataset）是由中山大学团队制作的变化检测数据集（Shi等，2022b），包含20000对分辨率为0.5 m，大小为256×256像素的航空影像。影像涵盖时间从2007年至2014年，覆盖了香港部分地区。数据集包含城市扩张、更新以及自然变化等各种类型，整体变化情况较为复杂，检测难度较大。实验中将数据划分为12000个训练样本、4000个验证样本和4000个测试样本。

3.4 WHU-CD数据集

WHU-CD（Wu Han University Dataset）是由武汉大学团队提出的建筑物变化的数据集（Ji等，2019）。数据集影像大小为15354×32507像素，覆盖了新西兰—克赖斯特彻奇地区，数据获取时间分别为2012年和2016年，数据集中主要变化类型为建筑物灾后重建。本文按照重叠率为0、256×256像素大小的滑动窗口进行裁剪，共得到7620张样本，得到训练样本6096个、验证样本762个、测试样本762个。

4 模型检测结果与分析

4.1 实验设计及参数设置

本研究使用Pytorch框架实现模型构建及训练，硬件环境为Inter（R） Core（TM） i7-8700处理器及8G内存NVIDIA GeForce RTX 2060 SUPER显卡。模型训练过程中，根据实验硬件性能设置批量处理大小为8，训练迭代次数为200，初始学习率为

0.01，在迭代100次后采用学习率动量衰减策略（Darken等，1992）进行训练，使用交叉熵损失（Cross-entropy）作为损失函数，利用Adam优化器更新参数。

4.2 对比实验及结果

4.2.1 变化检测精度评价指标

本研究采用准确率P、召回率R、总体精度OA、F1和交并比IoU表征变化检测精度。各指标计算公式为

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{24}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{25}$$

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \tag{26}$$

$$\mathrm{F1} = \frac{2 \times P \times R}{P + R} \tag{27}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{28}$$

式中，TP表示检测为变化，实际也为变化（检测正确）的像素数，TN表示检测为未变化，实际也为未变化（检测正确）的像素数，FP表示检测为变化，实际为未变化（检测错误）的像素数，FN表示检测为未变化，实际为变化（检测错误）的像素数。

4.2.2 不同变化检测模型结果对比分析

选择了8种网络模型与本文提出的MFTSNet模型进行对比，网络结构参数与文献述及的最优参数一致。与本文提出模型相对比的其他几种模型简介如下：

（1）FC-EF：FC-EF基于U-net网络模型和早期融合策略，将双时相影像输入网络之前进行拼接，并使用跳跃连接结构来融合低级高级特征（Daudt等，2018）。

（2）FC-Siam-Conc：该模型采用孪生网络模型结构，将来自两个编码器分支和解码器相应层的3个特征图进行跳跃连接，用低层次空间细节信息来补充深层次抽象和更少局部化的信息，使输出图像中的边界预测更加准确（Daudt等，2018）。

（3）FC-Siam-Diff：与FC-Siam-Conc模型不同之处在于，双时相绝对差异特征通过跳跃连接与解码器输出特征相结合（Daudt等，2018）。

（4）DTCDSCN：基于孪生网络模型，DTCDSCN模型由两个共享权重的通道注意力-残差网络编码器和一个D-LinkNet解码器组成（Liu等，2021）。

（5）BIT：BIT网络模型是最早用于变化检测的Transformer网络模型，由特征提取网络、双时相Transformer层和预测层3个部分组成（Chen等，2022）。首先，使用改进的孪生ResNet-18网络提取高级语义特征并将其转换为语义标记；然后，通过Transformer层对语义标记进行上下文建模；最后，使用预测层将最终特征映射到像素空间上，经过卷积网络生成变化检测结果。

（6）ChangeFormer：该模型将Transformer编码器与多层感知机解码器结合到孪生网络模型（Bandara和Patel，2022），由提取双时相图像粗细特征的层次化Transformer编码器、计算不同尺度特征差异模块以及聚合多级特征差异并预测结果的浅层MLP解码器3部分组成。

（7）ICIF-Net：ICIF-Net模型利用CNN提取局部特征和Transformer全局语义建模来处理双时相图像对（Feng等，2022）。该方法构建了尺度内交叉交互模块和尺度间特征融合模块，用于时空上下文信息建模以生成更好的特征表示，并通过引入聚合和空间对齐模块实现不同分辨率信息集成，最后聚合特征图并生成变化检测结果。

（8）DMINet：利用跨层级联合注意力模块生成每个输入的全局特征分布，激发不同层级表示间的信息耦合，并利用差值和级联差异以及增量特征对齐的多级差异聚合，突出差异特征，生成变化检测结果（Feng等，2023）。

这9种网络模型在CDD、LEVIR-CD、SYSU-CD和WHU-CD共4种数据集上的变化检测精度对比结果如表1—4所示。根据表1—4，本文提出的变化检测模型MFTSNet在4个数据集上的F1指标和交并比IoU均为最优，分别为95.784%、90.368%、77.897%、89.696%和91.78%、82.429%、63.796%、80.845%，相较于次优模型分别提高了0.465%、0.113%、0.369%、2.13%和0.723%、0.188%、0.304%、2.962%。

图4为不同模型在部分数据集上的变化检测结果。

表1 CDD数据集上不同模型指标对比

Table 1 Comparison of precision indices for different models on CDD dataset

/%

模型	*P*	*R*	F1	OA	IoU
FC-EF	83.751	56.735	67.645	93.597	51.109
FC-Siam-Diff	91.339	51.745	66.064	93.728	49.325
FC-Siam-Conc	90.734	65.944	76.378	95.187	61.783
DTCDSCN	94.029	92.452	93.234	98.417	87.325
ChangerFormer	89.298	81.444	85.19	96.659	74.201
BIT	95.466	92.51	93.965	98.63	88.616
ICIF-Net	94.181	87.569	90.755	97.895	83.074
DMINet	**96.042**	94.607	95.319	98.910	91.057
MFTSNet	95.962	**95.607**	**95.784**	**98.945**	**91.78**

注：加粗数值表示最高精度。

表2 LEVIR-CD数据集上不同模型指标对比

Table 2 Comparison of precision indices for different models on LEVIR-CD dataset

/%

模型	*P*	*R*	F1	OA	IoU
FC-EF	88.609	81.283	84.788	98.514	73.593
FC-Siam-Diff	92.086	81.442	86.437	98.698	76.114
FC-Siam-Conc	90.456	84.013	87.116	98.734	77.172
DTCDSCN	91.016	**89.507**	90.255	99.015	82.241
ChangerFormer	89.811	81.059	85.210	98.567	74.232
BIT	**92.582**	87.639	90.043	99.013	81.889
ICIF-Net	90.866	88.076	89.449	98.942	80.912
DMINet	92.121	88.296	90.168	99.019	82.096
MFTSNet	91.331	89.426	**90.368**	**99.029**	**82.429**

注：加粗数值表示最高精度。

表3 SYSU-CD数据集上不同模型指标对比

Table 3 Comparison of precision indices for different models on SYSU-CD dataset

/%

模型	*P*	*R*	F1	OA	IoU
FC-EF	81.171	67.246	73.555	88.597	58.172
FC-Siam-Diff	81.207	65.885	72.748	85.791	53.233
FC-Siam-Conc	**83.423**	71.453	76.975	89.919	62.569
DTCDSCN	81.624	73.668	77.442	89.879	63.189
ChangerFormer	78.909	75.112	76.963	89.396	62.553
BIT	76.681	78.393	77.528	89.283	63.302
ICIF-Net	78.194	76.333	77.252	89.398	62.935
DMINet	81.479	73.321	77.185	**91.454**	63.492
MFTSNet	76.900	**78.920**	**77.897**	89.438	**63.796**

注：加粗数值表示最高精度。

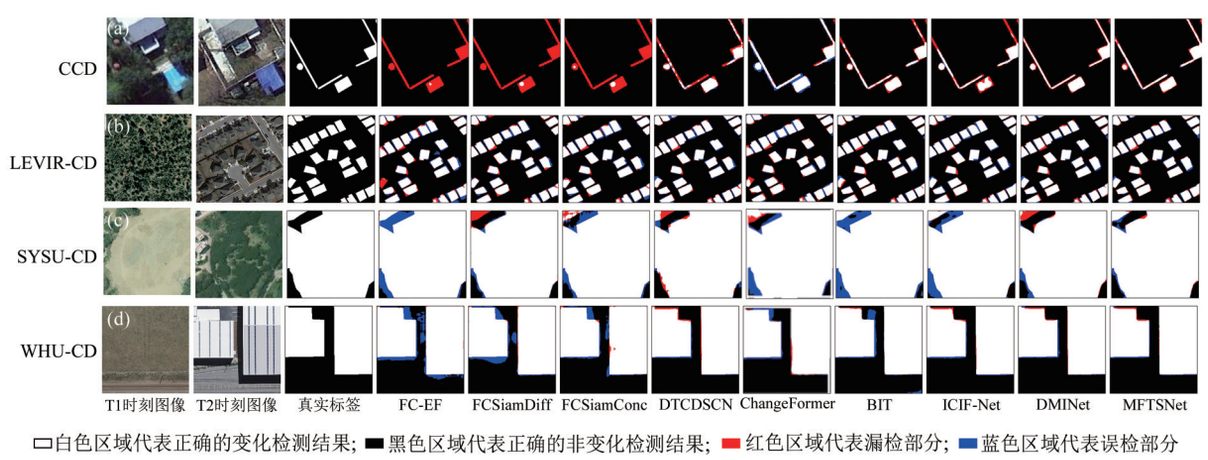
表4 WHU-CD数据集上不同模型指标对比

Table 4 Comparison of precision indices for different models on WHU-CD dataset

/%

模型	*P*	*R*	F1	OA	IoU
FC-EF	85.852	67.401	75.516	98.262	60.663
FC-Siam-Diff	85.808	74.616	79.822	98.500	66.420
FC-Siam-Conc	87.127	74.385	80.253	98.544	67.019
DTCDSCN	83.333	87.331	85.285	98.801	74.345
ChangerFormer	82.044	86.612	84.266	98.714	75.739
BIT	77.823	**87.558**	82.404	98.513	70.074
ICIF-Net	86.037	84.932	85.480	98.852	74.643
DMINet	88.784	86.382	87.566	99.024	77.883
MFTSNet	**94.840**	85.081	**89.696**	**99.177**	**80.845**

注：加粗数值表示最高精度。



□白色区域代表正确的变化检测结果；■黑色区域代表正确的非变化检测结果；■红色区域代表漏检部分；■蓝色区域代表误检部分

图4 9种不同模型在4个数据集上的变化检测结果

Fig. 4 Comparison of change detection results for various dataset using nine different models

由图4可见MFTSNet整体精度最高，证明了本文模型提出的有效性。其中，全卷积变化检测网络模型（包括FC-EF、FC-Siam-Diff、FC-Siam-Conc）检测结果较差，原因在于其特征编码、解码过程单一，网络结构简单，难以检测出复杂场景下的变化像素，误检、漏检现象较为严重；DTCDSCN、BIT、ICIF-Net、DMINet共4种模型的变化检测结果较好。由图4（a）可见，MFTSNet的检测结果明显优于对比模型，说明MFTSNet模型在季节变化较为明显、色彩偏差较大、有一定遮挡的情况下仍能较好识别出变化区域，并且MFTSNet在小目标检测效果上表现更好。由图4（b）可见，DMINet检测效果相比于MFTSNet更为准确，而MFTSNet在图4（c）样本上的检测效果更好。由图4（c）可见，在发生大面积变化时，MFTSNet能够准确地检测出变化区域边界。由图4（b）、（d）的建筑物变化检测结果，在只针对建筑物变化检测任务中，MFTSNet模型检测结果能够很好的保持建筑物边缘和建筑物内部完整性，说明MFTSNet模型鲁棒性较强。

4.3 消融实验及结果

为进一步评估MFTSNet方法以及多尺度特征Transformer中各个模块性能的有效性，考虑到时间成本，在SYSU-CD与WHU-CD 2个数据集上设计消融实验。将ST、GT、RT这3个模块混合搭配进行实验，共设计8个消融对进行比较分析，定量评估结果如表5所示。

表5 ST、GT、RT消融实验评估指标对比

Table 5 Comparison of ablation experiments for including different module of ST, GT and RT

/%

消融实验	ST	GT	RT	SYSU-CD		WHU-CD	
				F1	IoU	F1	IoU
1	×	×	×	71.153	56.763	73.355	61.460
2	×	×	√	72.026	57.636	80.493	70.663
3	×	√	×	71.983	57.912	81.076	71.526
4	√	×	×	76.725	62.239	87.745	78.165
5	×	√	√	73.329	59.112	85.696	74.972
6	√	×	√	77.187	62.534	88.814	79.399
7	√	√	×	77.304	62.004	89.023	80.218
8	√	√	√	**77.897**	**63.796**	**89.696**	**80.845**

注：×代表去掉该模块，√表示加入该模块。

由表5可见，多尺度特征Transformer中各个模块共同改善了网络模型性能。其中，比较消融实验1和4，可发现ST模块对模型性能的影响最大，加入ST模块前后，模型在两个数据集上的F1与IoU指标分别提高5.572%、14.39%与5.476%、16.705%；对比消融实验1—3可知，只利用GT或RT模块很难大幅提升网络模型的检测性能，这说明在网络中加入ST模块能够为后续模块提供丰富的变化信息；从消融实验4与6—8的精度指标来看，GT、RT模块能共同提升模型性能，相较于未添加GT、RT模块前，模型在两个数据集上的F1与IoU指标分别提高1.172%、1.951%与1.557%、2.68%。

图5为消融实验的部分变化检测结果。可见，前3个消融实验的检测结果均出现较大面积的误检现象，说明缺少多尺度特征Transformer模块时，伪变化信息容易对模型结果产生影响，模型鲁棒性低。从实验4、6和7的检测结果中可看出，GT和RT模块可明显减弱网络模型对小尺度区域的误检，体现了GT与RT模块跨尺度交互功能的必要性。比较实验1—3、5与8的检测结果，可发现去掉ST模块，模型检测结果中图5（b）、（d）与（f）部分区域存在孔洞现象，变化区域内部不完整，说明ST模块跨空间交互能够捕获不同位置之间的关联信息，进而增强特征图的表达能力，提高变化检测结果区域完整性。

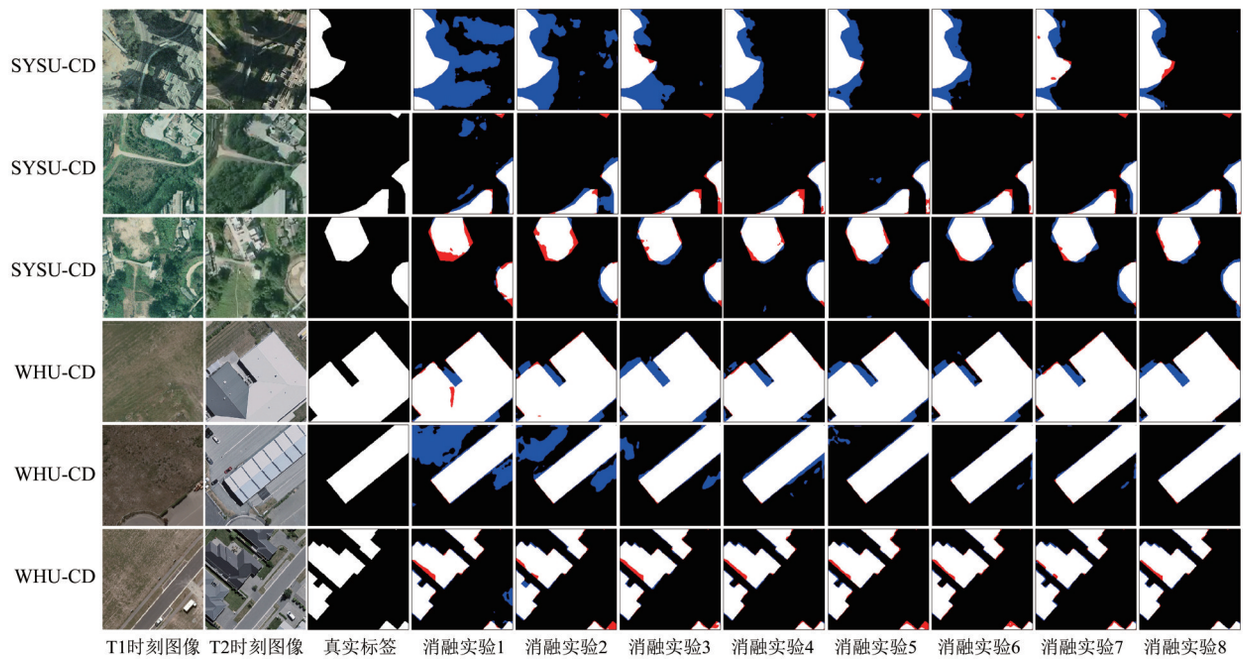


图5 不同消融实验的变化检测结果对比

Fig. 5 Comparison of change detection results for different ablation experiments

4.4 影响参数分析

在MFTSNet网络模型中，语义信息标记总长度参数L和编码器-解码器的个数（E_N，D_N）对变化检测效果的影响如何，需要进一步分析。对于语义信息标记总长度L，分别在4个数据集上取$L\in\{2, 4, 8, 16, 32, 64\}$，将$E_N$和$D_N$设定为1，分析$L$取值对模型结果指标的影响（Chen等，2022；Shi等，2022a；Song等，2022b）。选择不同L取值时，MFTSNet在4个数据集上的F1和IoU结果比较见表6。

表6 不同语义信息长度L取值对MFTSNet结果的影响

Table 6 Influence of different semantic information L values on MFTSNet results

/%

信息长度 L	CDD		LEVIR-CD		SYSU-CD		WHU-CD	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU
2	95.343	91.101	89.972	81.772	77.139	62.749	88.480	80.016
4	95.427	91.254	90.095	81.975	77.172	62.835	88.472	80.003
8	95.487	91.363	**90.257**	**82.245**	**77.774**	**63.631**	89.023	80.218
16	**95.526**	**91.434**	89.635	81.247	77.593	63.076	**89.583**	**80.657**
32	95.521	91.427	89.984	81.792	77.492	63.258	89.539	80.146
64	95.314	91.048	90.106	81.994	77.328	63.041	88.734	79.75

注：加粗数值表示最高精度。

从表6可看出，在数据集CDD和WHU-CD中，MFTSNet模型参数L=16时效果最优，F1和IoU分别为95.526%、89.583%和91.434%、80.657%。在数据集LEVIR-CD和SYSU-CD上L=8时效果最优，F1和IoU分别为90.257%、77.774%和82.245%、63.631%。实验结果表明，合适的L取值能够有效捕获丰富的语义信息。当L取值较小时，模型会造成部分语义信息丢失；当L取值较大时，模型中获得的冗余语义信息，会降低模型变化检测性能。

对于编码器—解码器的个数E_N和D_N，在CDD与LEVIR-CD两个数据集上进行实验。根据表6的分析结果，在数据集CDD与LEVIR-CD上进行训练时，分别设置L为16和8，分别选择（E_N，D_N）=（1，1），（1，2），（1，4），（2，1），（4，1），比

较MFTSNet在两个数据集上的F1与IoU指标，实验结果如表7所示。根据表7，当（E_N，D_N）=（1，2）时，MFTSNet在两个数据集上的F1与IoU指标最优，分别为95.784%、90.368%与91.78%、82.429%。E_N=1时的结果精度最优，说明单个编码器能够提供网络所必需的双时相影像变化信息。相对于E_N，D_N变化对检测结果精度影响更大，这表明恰当的解码器个数可为网络模型提供更加精细化的变化特征信息。

表7 E_N、D_N不同组合时MFTSNet在2个数据集上的精度指标比较

Table 7 Comparison of precision indicators of MFTSNet on two Datasets under different combinations of E_N and D_N

/%

E_N	D_N	CDD		LEVIR-CD	
		F1	IoU	F1	IoU
1	1	95.526	91.434	90.257	82.245
1	2	**95.784**	**91.78**	**90.368**	**82.429**
1	4	95.316	91.314	90.061	81.920
2	1	95.490	91.382	90.252	82.194
4	1	95.513	91.412	90.210	82.166

注：加粗数值表示最高精度。

5 结 论

针对变化检测模型语义信息提取不充分、多尺度细节特征丢失以及变化检测结果边界和内部细节不完整等现象，本研究提出MFTSNet模型，并通过在4个数据集上进行对比实验、消融实验与参数分析，验证了MFTSNet模型。主要结论如下：

（1）通过ST模块的跨空间交互，可以捕捉图像中不同位置之间的关联信息，促进不同位置之间的信息共享，从而提高模型的特征表达能力；（2）通过GT、RT模块进行特征间的跨尺度交互，能够获取不同尺度之间的关联信息，尤其是能够明显改善小尺度区域误检现象，对于提高模型的泛化能力具有重要作用；（3）与8个模型的对比实验结果表明，MFTSNet模型整体性能较优，变化检测结果最好，在4个公开数据集上F1指标和交并比IoU上相比于其他变化检测模型分别至少提高0.465%、0.113%、0.369%、2.13%和0.723%、0.188%、0.304%、2.962%；（4）特征信息长度L在CDD、WHU-CD数据集上取16，在SYSU-CD、LEVIR-CD数据集上取8，同时（E_N，D_N）=（1，2），MFTSNet模型检测结果最优。

本文主要通过变化检测精度评价指标和消融实验对比分析了MFTSNet模型的精度，而模型机理分析相对不足。后续需要针对多尺度特征Transformer模块用于变化检测的可解释性分析方法深入开展研究工作。

参考文献（References）

Bandara W G C and Patel V M. 2022. A transformer-based Siamese network for change detection//2022 IEEE International Geoscience and Remote Sensing Symposium. Kuala Lumpur: IEEE: 207-210 [DOI: 10.1109/IGARSS46834.2022.9883686]

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]

Darken C, Chang J and Moody J. 1992. Learning rate schedules for faster stochastic gradient search//Neural Networks for Signal Processing II Proceedings of the 1992 IEEE Workshop. Helsingoer: IEEE: 3-12 [DOI: 10.1109/NNSP.1992.253713]

Daudt R C, Le Saux B and Boulch A. 2018. Fully convolutional Siamese networks for change detection//25th IEEE International Conference on Image Processing. Athens: IEEE: 4063-4067 [DOI: 10.1109/ICIP.2018.8451652]

Dian Y Y, Fang S H and Yao C H. 2016. Change detection for high-resolution images using multilevel segment method. Journal of Remote Sensing (in Chinese), 20(1): 129-137 (佃袁勇, 方圣辉, 姚崇怀. 2016. 多尺度分割的高分辨率遥感影像变化检测. 遥感学报, 20(1): 129-137) [DOI: 10.11834/jrs.20165074]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16x16 words: transformers for image recognition at scale//9th International Conference on Learning Representations. [s.l.]: ICLR

Feng Y C, Jiang J W, Xu H H and Zheng J W. 2023. Change detection on remote sensing images using dual-branch multilevel intertemporal network. IEEE Transactions on Geoscience and Remote Sensing, 61: 4401015 [DOI: 10.1109/TGRS.2023.3241257]

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]

Guo Q L, Zhang J P, Zhu S Y, Zhong C X and Zhang Y. 2022. Deep multiscale siamese network with parallel convolutional structure and self-attention for change detection. IEEE Transactions on Geoscience and Remote Sensing, 60: 5406512 [DOI: 10.1109/TGRS.2021.3131993]

Hughes L H, Schmitt M, Mou L C, Wang Y Y and Zhu X X. 2018. Identifying corresponding patches in SAR and optical images with a pseudo-Siamese CNN. IEEE Geoscience and Remote Sensing Letters, 15(5): 784-788 [DOI: 10.1109/LGRS. 2018. 2799232]

Hussain M, Chen D M, Cheng A, Wei H and Stanley D. 2013. Change detection from remotely sensed images: from pixel-based to object-based approaches. ISPRS Journal of Photogrammetry and Remote Sensing, 80: 91-106 [DOI: 10.1016/j.isprsjprs.2013.03.006]

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]

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: 565-571 [DOI: 10.5194/isprs-archives-XLII-2-565-2018]

Liao M S, Zhu P and Gong J Y. 2000. Multivariate change detection based on canonical transformation. Journal of Remote Sensing (in Chinese), 4(3): 197-201 (廖明生, 朱攀, 龚健雅 . 2000. 基于典型相关分析的多元变化检测 . 遥感学报, 4(3): 197-201) [DOI: 10.11834/jrs.20000307]

Liu M X, Chai Z Q, Deng H J and Liu R. 2022. A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15: 4297-4306 [DOI: 10.1109/JSTARS.2022.3177235]

Liu Y, Pang C, Zhan Z Q, Zhang X M and Yang X. 2021. Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model. IEEE Geoscience and Remote Sensing Letters, 18(5): 811-815 [DOI: 10.1109/LGRS.2020.2988032]

Shi N, Chen K M and Zhou G Y. 2022a. A divided spatial and temporal context network for remote sensing change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15: 4897-4908 [DOI: 10.1109/JSTARS.2022.3176858]

Shi Q, Liu M X, Li S C, Liu X P, Wang F and Zhang L P. 2022b. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing, 60: 5604816 [DOI: 10.1109/TGRS.2021.3085870]

Song F, Zhang S X, Lei T, Song Y X and Peng Z M. 2022a. MSTDSNet-CD: multiscale Swin transformer and deeply supervised network for change detection of the fast-growing urban regions. IEEE Geoscience and Remote Sensing Letters, 19: 6508505 [DOI: 10.1109/LGRS.2022.3165885]

Song X Y, Hua Z and Li J J. 2022b. Remote sensing image change detection transformer network based on dual-feature mixed attention. IEEE Transactions on Geoscience and Remote Sensing, 60: 5920416 [DOI: 10.1109/TGRS.2022.3209972]

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]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc.: 6000-6010

Yang B, Mao Y, Chen J, Liu J Q, Chen J and Yan K. 2023. Review of remote sensing change detection in deep learning: bibliometric and analysis. National Remote Sensing Bulletin, 27(9): 1988-2005 (杨彬, 毛银, 陈晋, 刘建强, 陈杰, 闫凯 . 2023. 深度学习的遥感变化检测综述: 文献计量与分析 . 遥感学报, 27(9): 1988-2005) [DOI: 10.11834/jrs.20222156]

Yang Z L, Hu Z H, Salakhutdinov R and Cohen W W. 2018. Breaking the softmax bottleneck: a high-rank RNN language model//6th International Conference on Learning Representations. Vancouver: ICLR

Zhang C S, Li G J and Cui W H. 2018. High-resolution remote sensing image change detection by statistical-object-based method. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(7): 2440-2447 [DOI: 10.1109/JSTARS. 2018. 2817121]

Zhang D, Zhang H W, Tang J H, Wang M., Hua X S and Sun Q R. 2020. Feature pyramid transformer//16th European Conference on Computer Vision. Glasgow: Springer: 323-339 [DOI: 10.1007/978-3-030-58604-1_20]

Zhao M and Zhao Y D. 2018. Object-oriented and multi-feature hierarchical change detection based on CVA for high-resolution remote sensing imagery. Journal of Remote Sensing (in Chinese), 22(1): 119-131 (赵敏, 赵银娣 . 2018. 面向对象的多特征分级CVA遥感影像变化检测 . 遥感学报, 22(1): 119-131) [DOI: 10.11834/jrs.20186293]

Change detection for high-resolution remote sensing images with multi-scale feature transformer

LI Jiankang[1], ZHANG Guixin[2], ZHU Shanyou[1], XU Yongming[1], LI Xiangyu[1]

1.School of Remote Sensing & Geomatics Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China;
2.School of Geographical Sciences, Nanjing University of Information Science & Technology, Nanjing 210044, China

Abstract: Wetland is an important ecosystem and plays a vital role in maintaining regional ecological security. Wetland structure changes respond sensitively to natural and human activities, and flood wetlands experience drastic seasonal water and vegetation changes due to intermittent flood inundations. Mapping high-accuracy wetland structures is challenging because of frequent water and vegetation alternations, which cause spectral confusion and misclassification in optical satellite images. Several wetland extraction methods are available today, including object-oriented methods, whose parameters need to be decided subjectively, and machine learning methods, which have relatively low accuracy. With the continuous development of deep learning in image semantic segmentation, a precise and automatic remote sensing image binary classification becomes possible. Recent studies have suggested that deep learning semantic segmentation methods show great potential for mapping wetland changes in high-resolution images. However, the extraction of wetland structures in complex floodplain scenarios places high demands on models in terms of mining deep spatial information. The deformable U-Net (D-UNet) semantic segmentation model is improved to enhance the accuracy of the extraction of floodplain wetland structure.

In this study, the Taitema Lake in Xinjiang, China was selected as the study area because it is a typical floodplain wetland in the arid zone. A multiscene and multitemporal wetland sample dataset was collected using Sentinel-2 remote sensing images in the study area. The D-UNet for wetland structure extraction used VGG16 to build the encoding/decoding network and focused on improving the convolutional layer in the network. D-UNet was improved by replacing the convolution block before dimensionality reduction with multiscale dilated convolutions to enhance the network's receptive field, fuse features of different scales, and avoid loss of detailed information in high-resolution remote sensing images. After pretraining D-UNet, we determined that a multiscale convolution module consisting of three scales with dilation rates of 1, 2, and 3 would maximize the network's receptive field. We eventually input multiple remote sensing images from multiple scenes to fine-tune our model.

The applicability of the improved D-UNet model, traditional index-thresholding methods, and four classical semantic segmentation networks for extracting wetland structural information in floodplains was compared. Results showed that the improved D-UNet had an overall accuracy of 96.3% in single-temporal image wetland structure extraction, with a kappa coefficient of 0.839. Moreover, it demonstrated better transferability on time-series images, with an overall multitemporal accuracy of 92.3%. Compared with five models and the index-thresholding method, the improved D-UNet model showed better application potential in the extraction of floodplain wetland structure. It reduced misclassification and omission of wetland water bodies and vegetation by 7.2% and 48.9% compared with the index-thresholding method and by 0.6% and 5.4% compared with D-UNet, respectively.

This study proposes an effective classification method for the identification of fine structures in floodplain wetlands. It verifies the excellent performance of the semantic segmentation model in the extraction of complex feature information from remote sensing images. The improved D-UNet model can be used for information extraction for floodplain wetlands in similar environments. It provides a reference for rapid automated mapping of large-scale wetlands.

Key words: high resolution remote sensing, change detection, deep learning, siamese network, multi-scale feature, transformer, semantic information, ablation experiment

Supported by National Natural Science Foundation of China (No. 42171101, 41871028); Major Project of High Resolution Earth Observation System (No. 30-Y60B01-9003-22/23)