SWSACNet:面向多源影像的震后倒塌建筑物变化检测网络模型

龙颖[1]，窦爱霞[1*]，王斐斐[2]，王书民[1]

1. 中国地震局地震预测研究所,北京,100036;
2. 河南省地震局,郑州,450016

摘　要：针对不同时相的多源遥感影像存在的空间异质性问题，本文对全变网络模型（Fully Transformer Network，FTN）进行改进，提出一种端到端的、基于滑窗式特征增强和卷积注意力混合机制的倒塌建筑物变化检测网络模型（Sliding-Window-Shift Attention Convolution mix Network，SWSACNet）。SWSACNet基于FTN的模型框架，使用ACmix（Attention Convolution mix）高效识别多源影像对中的倒塌建筑物特征，并通过滑窗相似度特征匹配减弱多源影像中位置偏差的影响。以2023年2月6日土耳其M_w7.8级地震为例，通过获取震前高分二号、Google影像和震后北京三号影像构建倒塌建筑物变化检测数据集，对SWSACNet、FTN等五种变化检测模型进行训练和震区倒塌建筑物提取测试。实验结果表明，SWSACNet识别精度F1 score达80.8%，mIoU为67.8%，均优于其他四类模型；SWSACNet在应用于Fevaipasa、Nurdagi和Islahiye三个测试场景中，模型平均识别精度F1 score为60.84%，表明模型在泛化性能上有待提升。

关键词：多源影像，深度学习，变化检测，倒塌建筑物提取

中图分类号：P2

引用格式：龙颖,窦爱霞,王斐斐,王书民.XXXX.SWSACNet:面向多源影像的震后倒塌建筑物变化检测网络模型.遥感学报,XX(XX):1-18

Long Ying, Dou Aixia, Wang Feifei, Wang Shumin. XXXX. SWSACNet: A change detection network for collapsed buildings extraction using multi-source images. National Remote Sensing Bulletin, DOI: 10.11834/jrs.20244057]

1　引言

遥感技术已成为震后应急救援、损失评估和恢复重建的重要技术手段，在汶川、玉树、土耳其等国内外多次地震应急响应工作中发挥了重要作用。随着遥感技术和传感器的发展，星载光学影像、SAR影像以及无人机数据越来越多地应用在震后倒塌建筑物检测任务中，并在时效性和准确性上得到了大幅提升（Dong，Shan，2013）。地震中的倒塌建筑物是导致人员伤亡和经济损失的主要因素。应用遥感技术获取建筑物的受损信息，起初主要采用人工目视解译。然而，该方式的人力和时间成本较高。伴随计算机视觉技术的发展，自动化提取倒塌建筑物的研究不断涌现。早期主要依赖传统的手工特征提取和机器学习方法检测倒塌建筑物（Rasika等，2006）。随着深度学习方法的日益成熟，研究人员也开始尝试将不同的网络模型应用至倒塌建筑物提取这一任务中。目前，基于遥感影像的倒塌建筑物自动化识别方法可分为基于震后单时相影像的震害识别和震前震后影像变化检测两类方法。前者大多依托深度学习方法，从震后影像上挖掘丰富的光谱、纹理、几何、上下文信息以提取倒塌建筑物（陈梦，王晓青，2019；伍燚垚，2020）。

研究发现，相较于仅使用震后单时相的遥感影像，利用地震前后多时相遥感影像进行联合分

收稿日期：XXXX-XX-XX；**预印本：**XXXX-XX-XX

基金项目：国家自然科学基金面上项目(编号:42271090);中国地震局地震预测研究所基本科研业务经费地震业务科技支撑项目(编号:CEAIEF2022050504);中国地震局地震预测研究所基本科研业务经费面上项目(编号:CEAIEF20230202).

第一作者简介：龙颖,研究方向为遥感与地震风险评估. E-mail: longying_98@163.com.

通信作者简介：窦爱霞. E-mail: axdothy@163.com.

析时的倒塌建筑物检测精度更高（Adriano 等，2020）。因为多时相遥感影像具备目标在时间维度和空间维度的信息，呈现的地物特征更全面，因此相较单时相影像的检测精度更高。针对震前震后多时相遥感影像进行倒塌建筑物变化检测的研究，主要依赖手工特征提取和深度学习两类方法。使用手工特征提取方法检测倒塌建筑物的相关研究有：王天临，金亚秋（2012）借助震前光学和震后SAR影像，对震后建筑物形态进行仿真，并计算仿真图和SAR影像的多类互信息量对建筑物破坏状态进行评估。这类方法在某些特定场景和小规模数据上表现良好，但在面向复杂任务时，特征表达能力有限，扩展性差。深度学习方法无需人工参与特征工程，而是从原始数据中学习有效的特征表示，在复杂任务上往往可以取得更好的性能（Girshick 等，2014；Krizhevsky 等，2012）。变化检测任务中的深度学习网络模型，如全卷积早期融合网络FC-EF（Daudt 等，2018）；时空注意力网络STANet（Chen，Shi，2020）以及双注意力全卷积神经网络的DASNet（Chen 等，2021）等，也取得了较大突破。在提取倒塌建筑物这一应用中，变化检测常作为在提取震前震后完整建筑物后进行比对的后处理方法，如Ge 等（2023）采用域增量式的学习策略，提高预训练模型对震后影像中完整建筑物的识别性能，再通过比对前后两个时相的完整建筑物得到倒塌建筑物。近几年，陆续有研究将变化检测网络模型直接应用于倒塌建筑物提取任务中。如，Zheng 等（2021）提出ChangeOS，通过提取震前影像中的完整建筑物轮廓，再对震后影像进行对象级的建筑物状态分类。Miyamoto，Yamamoto（2021）将建筑物年代、结构类型等GIS数据与使用3D卷积获取震前震后影像的特征图进行融合，通过多个模态信息的输入，提高倒塌建筑物的识别精度。Shen 等（2021）则结合多尺度卷积神经网络和交叉注意力机制对建筑物的受损状态进行分类。

然而，震后的倒塌建筑物形态各异，规则建筑物变化信息在提取震害建筑物中应用效果不佳。本文认为变化检测网络模型在应用于倒塌建筑物识别任务中存在的难点有：①与常规变化检测任务不同，检测地震后倒塌建筑物，不仅仅是检测位置发生变化的地物，其重点识别形态、光谱、纹理等均发生变化的建筑物。②因实际应急工作中多使用多源影像，而多源影像间在成像角度、空间分辨率上的差异会导致目标地物在两个时相的影像上出现形态不一致以及空间位置错位等问题。针对上述问题，本文以2023年2月6日发生的土耳其7.8级地震前后多种卫星平台拍摄的遥感影像为例，对变化检测网络模型全变网络（Fully Transformer Network，FTN）模型（Yan 等，2022）进行改进，通过加入滑窗式特征增强模块削弱空间偏移对检测目标的影响，并借助卷积注意力混合机制驱使模型关注倒塌建筑物在地震前后的遥感影像上的变化特征，减弱非目标地物（如新增的应急帐篷等）对模型的干扰，以达到精确提取震后倒塌建筑物的目的。

2 SWSACNet

2.1 FTN模型框架

全变网络（Fully Transformer Network，FTN）是一种全新的深度构建、深度还原的变化检测框架，其主要由SFE（Siamese Feature Extraction）、DFE（Deep Feature Enhancement）、PCP（Progressive Change Prediction）、DS（Deep Supervision）四个模块组成，其结构如图1所示。SFE模块使用Swin Transformer提取原始图像的特征，并通过孪生网络结构共享两个时相在特征提取阶段的权重。Swin Transformer具有层次化逐步构建特征图的特点，在浅层得到局部细节，在深层获取全局信息，不仅可以减少计算量还能获得与Transformer相同甚至更优的性能，在该模块最后引入额外的Swin模块来扩大特征映射的接受域。DFE模块则对SFE提取的两时相不同深度的特征进行加减强化操作，得到两时相在不同层次或深度下的特征强化图。PCP模块将特征强化图输入至堆叠的注意力金字塔PAM（Pyramid Attention Module）结构，进行渐进式推理，输出得到4种不同分辨率的预测图。最后DS模块分别计算四种分辨率预测图与标签图的结构相似性指数（Wang 等，2004）损失函数、交叉熵损失和IoU损失等进行深度监督学习。通过对强化特征进行不同尺度的还原，驱使模型关注地物变化的多重特征，抑制非目标地物的识别。

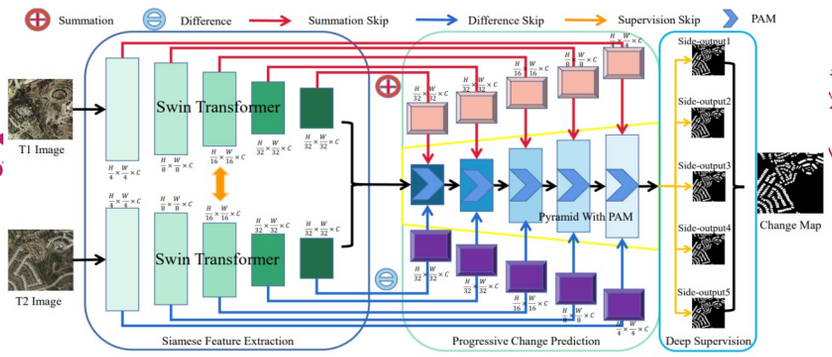


图1 FTN结构（来源：Yan等，2022）

Fig.1 The structure of FTN (source: Yan et al., 2022)

2.2 SWSACNet

针对倒塌建筑物复杂的破坏形态，和震前震后影像中建筑物位置匹配问题，本文对FTN结构的SFE模块的特征提取方法、DFE模块的特征增强方法进行了改进，构建了Sliding-Window-Shift Attention Convolution mix Network（SWSACNet）倒塌建筑物变化检测网络模型，模型主体框架仍然由FTN的四部分构成，网络修改部分如图2黄框所示。特征提取阶段（SFE），Attention Convolution mix（ACmix）由于同时具备卷积和注意力的优点，能够有效提高模型识别性能（Pan等，2023）。本文将使用ACmix中性能更优的基于分层注意力机制的Swin Transformer与Convolution的组合，即Swin_ACmix，替换孪生网络结构中用于编码的Swin Transformer。在特征增强阶段（DFE），为降低模型误识别，采取一种基于滑窗式的特征增强（SWSFE，Sliding-Window Feature Enhancement）方法，在模型训练过程中，不再按照FTN中对特征图相应位置的元素进行相加或相减（Feature Enhancement，FE），而是基于相似度和位置约束对不同时相的特征图元素进行重匹配，之后再进行不同时相特征的加减强化操作。此外，由于数据集中倒塌建筑物样本数量有限，可能存在各类别样本量不平衡问题，因此在训练部分用Focal Loss（Lin等，2017）替换原有交叉熵损失函数，降低样本类别均衡性影响。

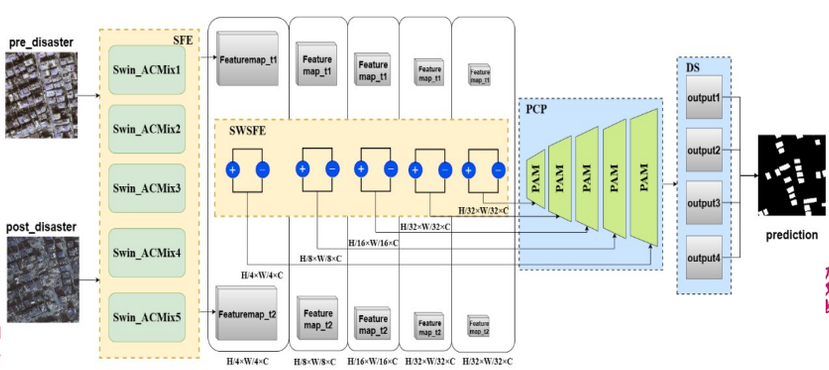


图2 SWSACNet结构

Fig.2 The structure of SWSACNet

2.2.1 ACmix

目前，主流的网络模型用于图像特征提取的底层结构主要包括卷积和注意力机制两种。卷积的原理如式（1）所示，$x'(i,j)$表示输入x在(i,j)位置处经过$K\times K\times C$的卷积核的加权求和后的输出值，其中$W(m,n,c)$表示卷积核在(m,n,c)位置处的值。借助卷积操作可以有效捕捉图像的局部特征并通过权重共享减少模型参数。

$$x'(i,j)=\sum_{c=1}^{C}\sum_{m=1}^{k}\sum_{n=1}^{k}x(i+m-1,j+n-1,c)\cdot W(m,n,c)$$

（1）

然而，卷积的感受野有限，难以捕获目标在图像中的长程关系。相反，注意力机制则通过建立图像的长距离依赖，使模型关注全图中具有关键信息的区域。其过程如式（2）所示。其中，X为输入，W_q、W_k、W_v分别为获取query（Q）、key（K）、value（V）的三个权重矩阵，y即为对X进行注意力打分后得到的注意力值矩阵。但是，对全局进行关联性计算也会增加模型的运算负担。

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{W}_q \cdot \boldsymbol{X} \\ \boldsymbol{K} &= \boldsymbol{W}_k \cdot \boldsymbol{X} \\ \boldsymbol{V} &= \boldsymbol{W}_v \cdot \boldsymbol{X} \\ \boldsymbol{y} &= \boldsymbol{V} \cdot softmax(\boldsymbol{K}^T \cdot \boldsymbol{Q}) \end{aligned} \tag{2}$$

ACmix则通过对卷积与自注意力进行特征分解和整合，对图像全局和局部建模的同时减少了自注意力层的计算量。卷积分解是通过3个1×1的标准卷积核将输入图像分解为3×N的特征块，自注意力分解是通过N个head得到N块注意力值，其过程类似于1×1的卷积映射，将输入特征投影为查询、键和值。将卷积和自注意力得到的特征块输入轻型全连接网络，将得到的特征图进行移位和聚合，最后将两条路径得到的特征图进行加权和操作得到$Attn_Conv$，其数学表达如式（3）所示。式中，$K_{p,q}$表示标准卷积核在(p, q)位置处的核权重，f表示输入特征图，则$Convf$表示经标准卷积、平移、聚合后输出的特征图，W_q、W_k、W_v分别表示查询、键、值的权重矩阵，d为$W_q f_{i,j}$的特征维数，则$Attnf$为经N个自注意力计算、聚合后的输出特征图，w_1、w_2对应卷积和自注意力输出特征的权重。

$$\begin{aligned} \boldsymbol{Convf} &= \sum_{p,q} \boldsymbol{K}_{p,q} \boldsymbol{f}_{i+p-\lfloor k/2 \rfloor, j+q-\lfloor k/2 \rfloor} \\ \boldsymbol{Attnf} &= \Big\|_{l=1}^{N} \sum_{a,b} softmax\left(\frac{\left(\boldsymbol{W}_q^l f_{i,j}, \boldsymbol{W}_k^l f_{a,b}\right)}{\sqrt{d}}\right) \boldsymbol{W}_v \boldsymbol{f}_{ab} \\ \boldsymbol{Attn\_Conv} &= w_1 * \boldsymbol{Convf} + w_2 * \boldsymbol{Attnf} \end{aligned} \tag{3}$$

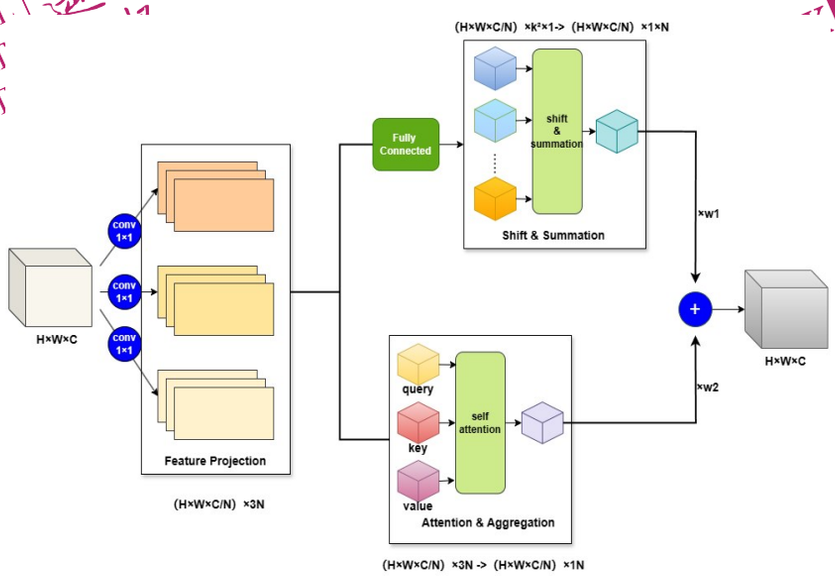


图3　ACmix结构

Fig.3　The structure of ACmix

2.2.2　滑窗平移特征增强方法（SWSFE）

计算两个时相特征图像上对应位置的特征差异时，会由于位置偏差导致增强部位信息不准确，Sliding-Window-Shift Feature Enhancement（SWSFE）借助滑窗进行相似度匹配，限制位置偏差的影响，从而减少模型由于位置不匹配而产生的误识别。SWSFE先计算两幅图像的相似度并以此作为索引依据，使用滑窗覆盖特征图，在滑窗内的区域搜寻与(i, j)匹配的像元位置，记录位置索引(idx_i, idx_j)，根据索引按照相似度重新匹配像元，最后对两幅图像进行加或减操作。相似度计算方法如式（4）所示，式中，τ为像素相似度权重，$v1_{(i,j)}$表示t_1时刻(i, j)位置的像素值，$v2_{(k,l)}$表示t_2时刻(i, j)对应滑窗区域内(k, l)位置处的像素值，(idx_i, idx_j)表示与$v1_{(i,j)}$最匹配的特征位置，$add_{shift(i,j)}$、$diff_{shift(i,j)}$分别表示变化前后两

时相特征的增量和差值。

$$sim_{t1,t2} = \tau*\left(v1_{(i,j)} - v2_{(k,l)}\right)^2 + (1-\tau)*\sqrt[2]{(i-k)^2+(j-l)^2}$$
$$idx_i, idx_j = min(sim_{t1,t2})$$
$$diff_{shift(i,j)} = Attn\_Conv_{t1(i,j)} - Attn\_Conv_{t2(idx_i,idx_j)}$$
$$add_{shift(i,j)} = Attn\_Conv_{t1(i,j)} + Attn\_Conv_{t2(idx_i,idx_j)} \quad (4)$$

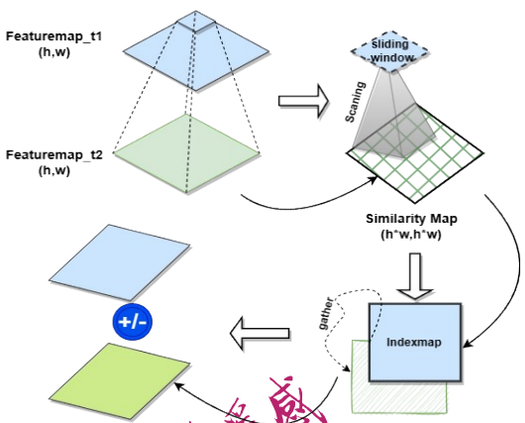


图4 SWSFE原理示意图

Fig.4 The process of SWSFE

2.2.3 Pyramid Attention Module

为进行深度监督，Pyramid Attention模块主要使用ppattention计算四个不同分辨率的特征增强图，再经过LayerNorm得到四种不同分辨率的预测图，并计算四个不同尺度下预测图和标签图的损失，最后通过误差反向传播调整模型中的权重参数。ppattetion过程如图5所示，通过交叉卷积进行通道增强，得到输入特征图不同通道之间的交互信息，再使用全连接层和sigmoid激活函数，计算经平均池化后的注意力权重。计算交叉卷积各通道的加权和，并将其与交叉卷积的输出一起输入到张量中。最后，对四种特征进行加权求和、批量归一化处理得到不同分割尺度下的预测结果。图中，CrossConv、avgpool、fc分别表示交叉卷积、平均池化、全连接层等过程。

2.2.4 损失函数

本文采用3类混合损失函数训练SWSACNet，分别是IoU损失、Focal Loss、结构相似性损失。

1）IoU损失：IoU用于计算预测结果*pred*和真实值*gt*之间的交并比，通过模型的不断迭代，增大预测框和真实框的几何重合度，计算公式如下。

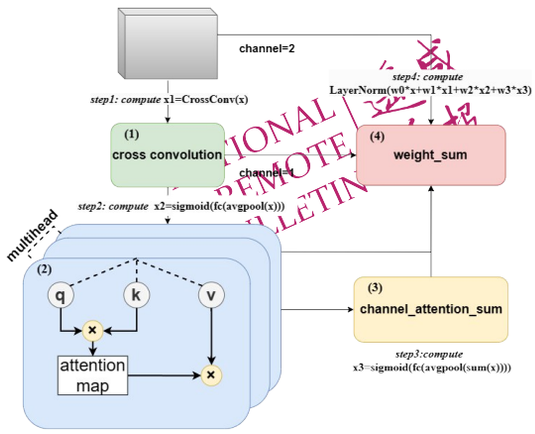


图5 ppattention结构

Fig.5 The structure of ppattention

$$IoU_{Loss} = 1 - \frac{Intersection(\boldsymbol{pred} \cap \boldsymbol{gt})}{Union(\boldsymbol{pred} \cup \boldsymbol{gt})} \quad (5)$$

式中，*Intersection*、*Union*分别表示预测和真实值的交集和并集的面积。

2）Focal Loss：针对样本中存在的不平衡问题，Focal Loss对交叉熵进行加权，以使模型更关注于正样本中易被判错的类别。

$$FL = -(1-pt)^{\gamma} log(pt) \quad (6)$$

式中，γ>0，*pt*表示正确分类的样本数和总样本数的比值，*pt*越大说明越接近真值。

3）结构相似性损失（Structural Similarity Loss）：用于度量两幅图像在亮度、对比度、结构上的相似性，和人类的视觉系统（HVS）类似，主要衡量对局部结构变化的敏感度。通过计算两个图像的均值、方差、协方差，再经高斯加权平均得到整幅图的结构性指数，计算公式如下所示。

$$SSIM_{Loss} = 1 - \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2)(\sigma_x^2 + \sigma_y^2)} \quad (7)$$

式中，x、y分别表示预测值和真实值，μ表示均值，σ表示协方差，c_1、c_2、c_3为常数。

3 建筑物震害数据集构建

3.1 震例概况

2023年2月6日，土耳其境内发生两次M_w7.8级地震，震源深度20公里，两次地震震中相距96公里。地震波及到土耳其Gaziantep、Osmaniye、Hatay、Nurdagi、Kahramanmaras等多个省份，造成约50500人员死亡，房屋倒塌逾2818栋，1040亿美元经济损失（CGTN，2023；Reuters，2023；Bianet，2023）。地震破裂带及两侧数十公里内的

大中城市和广大乡镇农村地震烈度达到8度以上，最高烈度达到10度（高孟潭，2023）。本文选取距离两次地震震中平均约50公里的Hatay、Kahramanmaras等建筑物破坏严重且具备地震前后影像的地区作为研究区。

3.2 研究数据及其预处理

本文收集了震前2022年9月–2022年10月高分二号（GF–2）、Google影像，震后2023年2月10日–15日北京三号（BJ–3）影像，如图6所示。GF–2、BJ–3均为光学遥感卫星，搭载的传感器覆盖红、绿、蓝、近红外以及全色五个波段。其中，GF–2多光谱的分辨率为3.2米，全色波段分辨率为0.8米，重返周期为5天。BJ–3多光谱分辨率为1.2–2.0米，全色波段分辨率为0.3–0.5m米，重返周期为3–5天。文中使用的Google影像空间分辨率为0.5m，包括红绿蓝三个波段。

对收集的GF–2、BJ–3影像进行了全色影像和多光谱影像匹配、融合、震前震后影像空间配准等处理，为使输入网络中的震前震后图像范围与大小保持一致，将GF–2、BJ–3、Google影像重采样至相同分辨率，得到用于制作数据集的初始影像对。

3.3 样本制作

样本数据制作采用人工目视解译方式，对比震前震后影像中的建筑物轮廓规则度、纹理一致性、光谱差异等特征变化，标注倒塌建筑物单体（如图7），共1891个。按224×224将影像裁剪后，得到2450个样本，包括震前震后影像对和标签图像，随机抽取1909个样本对作为训练图像，541对作为测试图像。由于上述训练样本集中包含倒塌建筑物的图像对和不包含倒塌建筑物的背景图像对数量不平衡，冗余信息过多，因此需要从训练样本集中剔除部分场景相似、重复出现或纯影像黑边的背景图，最终得到1000个训练图像对。

3.4 数据增强

数据增强是一项提高模型泛化性的技术，目的是可以提高模型识别不同场景中同一目标的能力，即通过阻止网络学习不相关的特征来提升网络的识别能力。通常采用旋转、平移或翻转等几何变换，使模型减少对目标空间位置或目标朝向的关注，而颜色变换、亮度增强、锐化、白平衡等光谱增强方法使模型减少对于光照或不同颜色的组合引起的同一幅影像的色彩差别。这些增强方法比较适合不同传感器平台拍摄时所形成的影像差异。本文使用了翻转、旋转60°、平移、颜色扰动、对比度增强、自适应亮度增强、颜色增强、白平衡、锐化共9种增强方法（Shorten，Khoshgoftaar，2019），将训练数据集扩充至10000对。

（a） Kahramanmaras局部区域震前GF-2影像（左）、震后BJ-3影像（右）

（a） Pre-earthquake GF-2 image (left) and post-earthquake BJ-3 image (right) of Kahramanmaras local area

（b） Hatay局部区域震前Google影像（左）、震后BJ-3影像（右）

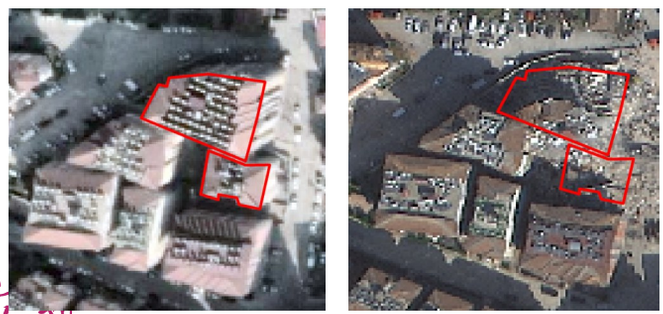
（b） Pre-earthquake Google image (left) and post-earthquake BJ-3 image (right) of Hatay local area

图6 土耳其7.8级地震前后遥感影像

Fig.6 Remote sensing images of Turkey before and after Turkey M_w7.8 earthquake

（a） 震前GF-2影像
（a） Pre-earthquake GF-2 image

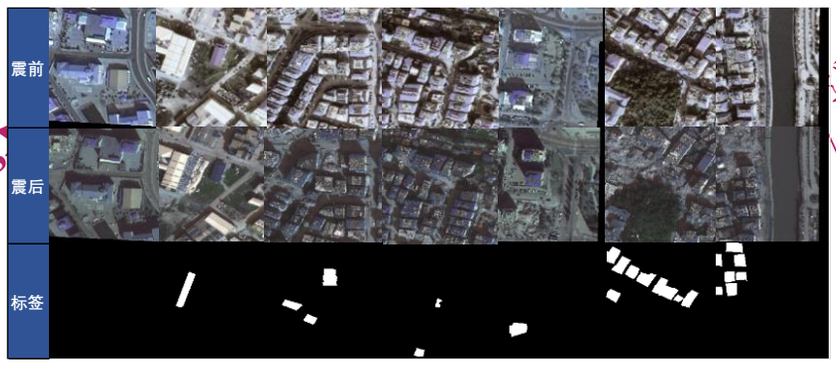
（b） 震后BJ-3影像
（b） Post-earthquake GF-2 image



（c） 震前Google影像
（c） Pre-earthquake Google image

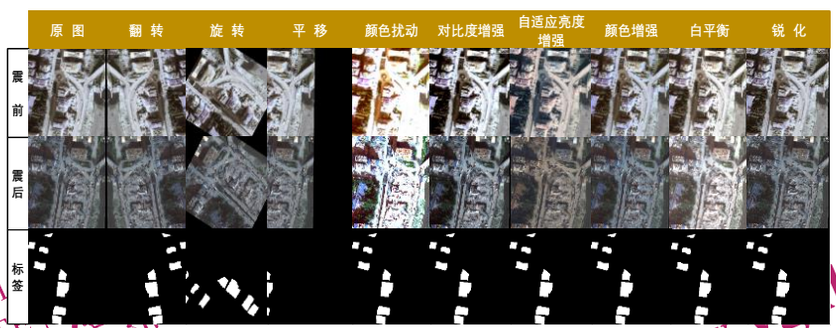
（d） 震后BJ-3影像
（d） Post-earthquake BJ-3 image

图7 震前震后影像中的倒塌建筑物示例
Fig.7 Example of pre- and post-earthquake images and collapse buildings



（a） 裁剪后样本集

（a） Cropped sample set（from top to bottom：pre-earthquake image，post-earthquake image，collapsed building label）



（b） 数据增强后样本示例

（b） Example after data enhancement（from top to bottom：pre-earthquake，post-earthquake，collapsed building label，from left to right：original image，flip，rotation 60°，translation，color jitter，contrast enhancement，adaptive brightness enhancement，color enhancement，white balance，sharpening）

图8 倒塌建筑物变化检测数据集

Fig.8 Collapsed building change Detection Dataset

4 实验

4.1 实验条件

1）硬件配置。实验基于RTXA30显卡、使用python3.8开发语言，深度学习框架选用pytorch1.11.0。

2）超参数设置。超参数设置为训练轮次epoch=150；训练一次的数据量batc_size=16；迭代时更新参数的步长learning_rate=0.01；drop_path_rate设为0.3，通过随机丢弃网络层减少模型过拟合。

4.2 评价指标

实验选用Precision、Recall、F1 score、OA（Overall Accuracy）、mIoU作为评价指标。

1）Precision：分类精度，计算被预测为正例的样本中真正为正例的比例，多用于图像分类、实例分割等任务。分类精度越高，模型对正例预测的准确程度越高，反映模型的查准率。

$$precision = \frac{TP}{TP + FP} \tag{8}$$

2）Recall：召回率，计算所有实际为正例的样本中被正确预测为正例的比例。召回率越高，表明模型漏检的类别数量越少，反映模型的查全率。

$$recall = \frac{TP}{TP + FN} \tag{9}$$

3）F1 score：分类精度*precision*和召回率*recall*的调和平均值。F1 score用于衡量模型在查准率和查全率上的平衡能力。相较于单一使用Precision或Recall，F1 score是一个更全面的评估指标。

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall} \tag{10}$$

4）OA：总体精度，计算预测正确类别与总体类别的比值。图像中倒塌建筑物数量较少，背景类别占比较高，OA值会偏高。

$$OA = \frac{TP + TN}{TP + TN + FN + FP} \tag{11}$$

5）mIoU：平均交并比，计算不同预测类别与实际类别在图像上的交集面积与并集面积之比的平均值，反映模型在各个类别上的分割效果。下式中，$k+1$表示目标类别与背景类别的数量总和。

$$mIoU = \frac{1}{k+1} \cdot \sum_{i=0}^{k} \frac{TP}{TP + FP + FN} \tag{12}$$

4.3 模型对比

为评估本文提出的SWSACNet的倒塌建筑物识别效果，实验选取了FTN、DASNet、STANet和FC-EF等四类变化检测模型与SWSACNet进行对照实验。使用土耳其地震倒塌建筑物变化检测数据集进行训练，并计算五个模型在测试数据集上的精度指标，如表1所示。由表可知，五个网络模型中SWSACNet精度最优：Precision达81.83%，F1 score达80.8%，Recall为79.8%，OA达90.64%，mIoU达67.8%。为描述模型之间识别结果的差异，本文定义正变化为前一时刻存在，后一时刻消失的地物，反之则为负变化。其中，倒塌建筑物属于正变化一类。根据典型区域的各模型测试结果（图9）可知：1）FC-EF模型检测能力最弱，未检测出大部分受损建筑物，且正负变化信息识别均受到抑制；2）STANet、DASNet的识别准确度较为接近，对于负变化信息在一定程度上有所抑制，对于测试集中小部分的震前建筑物轮廓明显，震后轮廓全部模糊的损坏建筑物识别较准确，但这两个模型并未区分正变化中的非目标地物和倒塌建筑物（如图9中红框标注部分）；3）FTN、SWSACNet对于正变化地物中倒塌建筑物与非倒塌建筑物的区分度更高。结合表1可知SWSACNet的误分率更低，且各精度指标均略高于FTN，综合表明SWSACNet在测试数据集检测倒塌建筑物实验效果最佳。

表1 倒塌建筑物模型检测精度对比

Table 1 Accuracy comparison of damaged building model detection result

Methods	Precision	F1 score	Recall	OA	mIoU
FC-EF	37.1%	36.6%	36.11%	81.15%	22.37%
STANet	47.36%	47.11%	46.86%	75.22%	30.82%
DASNet	56.12%	60.3%	65.15%	80.4%	43.18%
FTN	79.7%	79.24%	78.79%	89.39%	65.66%
SWSACNet	81.83%	80.8%	79.8%	90.64%	67.8%

4.4 消融实验

图9 模型预测结果对比

Fig.9 Prediction comparison of different models

本文提出的SWSACNet沿用FTN框架，在Encoder部分将Swin Transformer替换为ACmix，使用SWSFE滑窗式特征增强结构替换DFE模块中直接进行特征增强的FE结构。为检验模型改进效果，除对比FTN外，本文设置了ACmix+SWSFE、ACmix+FE、Swin Transformer+SWSFE三组模型结构对照实验，实验结果精度见表2。据表可知，ACmix和SWSFE的组合（SWSACNet）在

Precision、F1 score、Recall、OA、mIoU上表现最优，相较Swin Transformer+SWSFE和ACmix+FE平均高出1.91%、2.82%、3.66%、0.815%、3.81%。结合表1可知，Swin Transformer+SWSFE与ACmix+FE在检测精度上均优于FTN，但在Recall上略小于FTN。然而，由于滑窗操作中大量的匹配计算增加了模型的运算量，使得模型的训练时间较FTN延长约10小时，平均每张图像的推理速度慢0.5s。因此，滑窗算法仍需提高计算效率以缩短模型训练和推理时长。综上可得，本文提出的SWSACNet结构中的ACmix和SWSFE组合对模型性能的提升起到一定作用。然而，在图10中黄框标记区域三种模型组合结构都出现误分，针对该问题的分析详见4.5。

表2 SWSACNet模型消融实验结果对比

Table 2 Accuracy comparison of ablation experiments with different modules

	Precision	F1 score	Recall	OA	mIoU	Training Time	Inference Speed
Swin Transformer+ SWSFE	79.97%	78.81%	77.68%	89.55%	65.14%	27.7h	0.41(s/picture)
ACmix+ FE	79.88%	77.15%	74.6%	90.1%	62.84%	22.6h	0.32(s/picture)
ACmix+ SWSFE	81.83%	80.8%	79.8%	90.64%	67.8%	26.2h	0.67(s/picture)
FTN	79.7%	79.24%	78.79%	89.39%	65.66%	16.5h	0.22(s/picture)

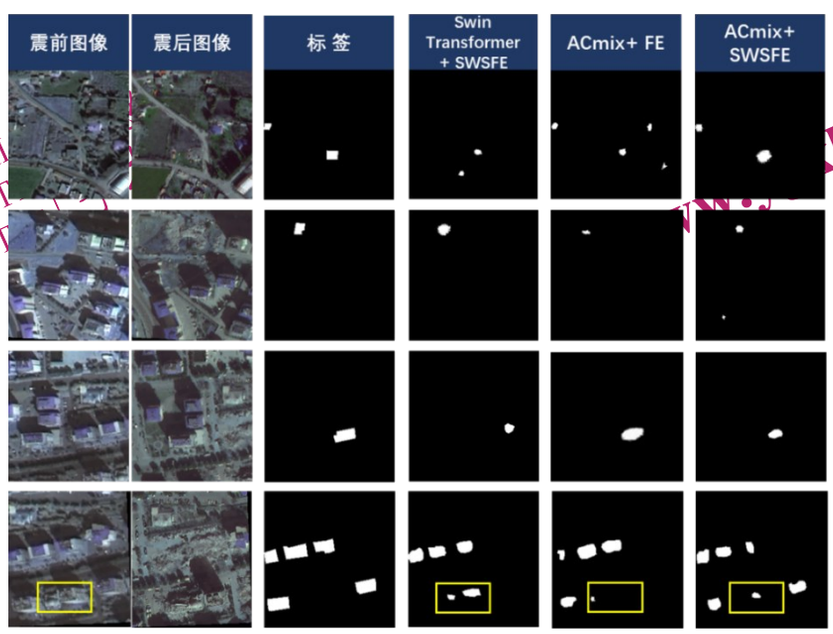


图10 消融实验预测结果对比

Fig.10 Ablation comparison of predictions with different modules

4.5 多源影像对梯度变化率影响分析

针对上述三种结构均出现误识别的情况，对该图像对的中间特征图进行可视化，如图11所示)。据图可知，模型将震后影像中未倒塌但具有高噪点的建筑物误判为倒塌建筑物。经分析，模型出现此类误识别的原因可能为初始训练数据集中多源影像对的梯度变化率不统一所致。因此，本节将探讨Google、GF-2、BJ-3三类影像间的梯度变化率差异对模型识别精度的影响。虽然三种数据经预处理后具有相同分辨率，但由于原始空间分辨率不同，导致地物边界处的梯度变化率存在差异。BJ-3空间分辨率最高，在不同地物的分界处梯度变化最剧烈，在图像上表现为地物间界限更分明，GF-2、Google原始影像空间分辨率较BJ-3低，地物边界过渡平缓，在图像上表现为地物间的边界较模糊。因此，本文对原始数据集中的BJ-3影像进行高斯平滑处理，统一影像间的梯度变化率，如图12所示。使用平滑后的变化检测数据集，重新训练和测试SWSACNet、FTN、DASNet、STANet、FC-EF等五个模型。结果表明

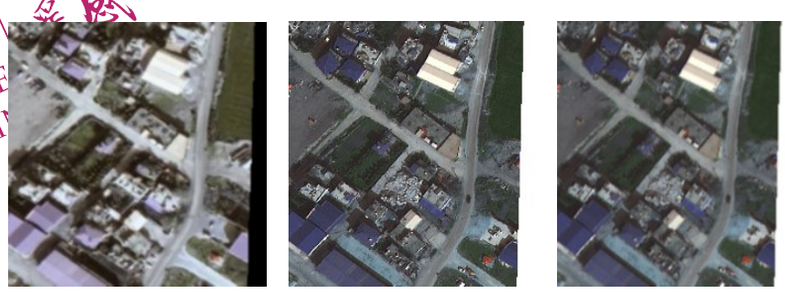
（表3），平滑数据集得到的五个模型精度有显著提升，其中，SWSACNet、FTN、FC-EF均出现1~3%的增幅。在未平滑前，震前震后样本对存在较大的空间异质性，经平滑处理后的样本对在空间分辨率一致的情况下被进一步同质化，模型在比较两幅图像时更易区分倒塌建筑物与未倒塌建筑物间的差异。一般情况下，影像梯度变化率越大，地物边界越清晰，越有助于模型对目标物和非目标物进行区分。然而，在利用多源影像对目标进行变化检测时，适当降低图像梯度变化率，保持图像对梯度变化率的一致有助于模型对目标地物的判别。

（a）震前 GF-2　（b）震后 BJ-3　（c）特征图

(a)Pre-earthquake GF-2 image　(b)Post-earthquake BJ-3 image　(c)Feature map

图11　SWSACNet特征图可视化

Fig.11　Feature map visualization of SWSACNet



（a）震前Google　（b）震后BJ-3　（c）高斯平滑处理后BJ-3

（a）Pre-earthquake Google image　（b）Post- earthquake BJ-3 image　（c）BJ-3 image after Gaussian Smoothing

图12　高斯平滑处理前后图像对比

Fig.12　Image before and after Gaussian Smoothing

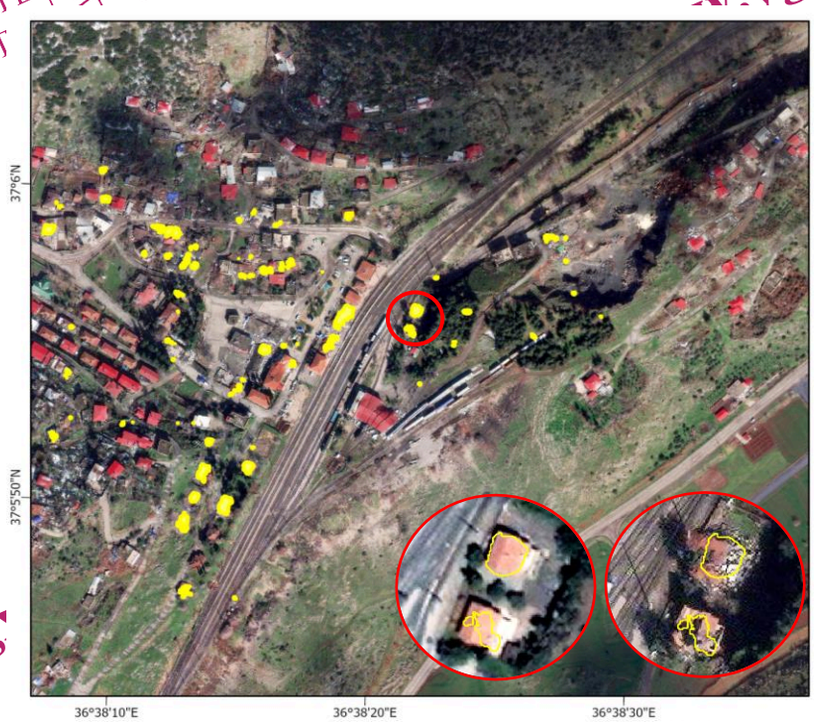
表3　五种变化检测模型在平滑后数据集上的测试精度

Table 3　Testing performance of five change detection networks in smoothing datasets

Metrics (rise)	SWSACNet	FTN	DASNet	STANet	FC-EF
Precision	86.42%(+4.59%)	84.08%(+4.38%)	57.67%(+1.55%)	47.95%(+0.59%)	39.01%(+2.51%)
Recall	81.17%(+1.37%)	80.42%(+1.63%)	66.28%(+1.13%)	47.87%(+1.01%)	37.51%(+1.4%)
F1 score	83.72%(+2.92%)	82.21%(+2.97%)	61.68%(+1.38%)	47.91%(+0.8%)	38.53%(+1.93%)
OA	97.52%(+6.88%)	95.19%(+5.8%)	85.24%(+4.84%)	77.31%(+2.09%)	65.6%(+4.45%)
mIoU	71.99%(+4.19%)	69.8%(+4.14%)	43.58%(+0.4%)	31.5%(+0.68%)	23.87%(+1.5%)

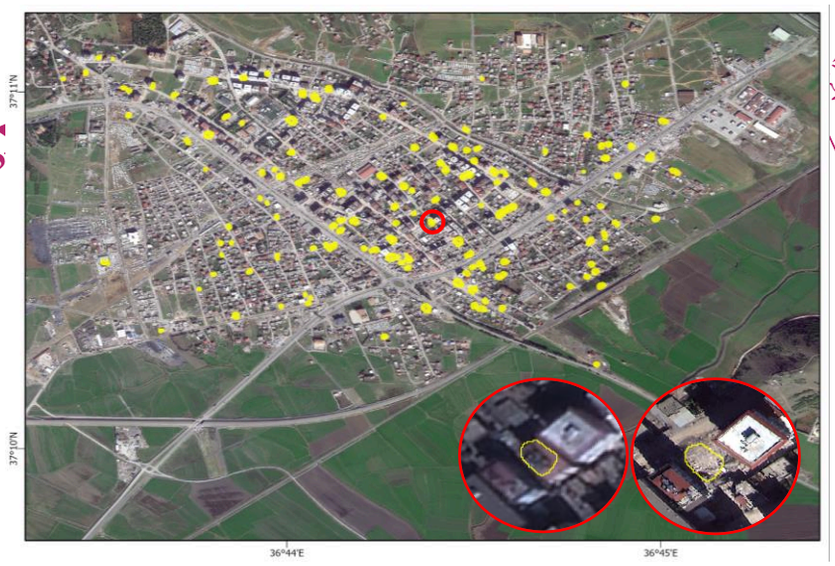
5 模型应用

为评估本文搭建的SWSACNet模型的泛化性能，本文选取了Fevaipasa镇局部区域的2022年7月和2023年2月7日空间分辨率为0.3m的WorldView无偏影像对，评估模型在成像角度接近的影像对中对倒塌建筑物的识别性能。同时，选取Nurdagi镇中心区域的2022年10月13日的GF-2和2023年2月13日的BJ-3影像，Islahiye镇中心区域的2022年12月27日的WorldView和2023年2月12日BJ-2（空间分辨率为0.8m）影像，评估模型在成像角度不一致的影像对中对倒塌建筑物的识别性能。表4为采用平滑前后数据集训练得到的SWSACNet模型在Fevaipasa、Nurdagi和Islahiye三个场景中的精度指标。图13显示了模型在三个场景中的推理结果的空间分布情况，子图的红色圆框中显示了倒塌建筑物在震前震后影像上的变化。整体而言，模型对WorldView震前震后无偏影像对（Fevaipasa地区）中位移变化型倒塌建筑物的识别效果良好，在该场景中，模型主要存在对裸地的误分情况；在有偏影像对中，模型对GF-2与BJ-3的组合（Nurdagi地区）存在少数漏分情况，对WorldView与BJ-2的组合（Islahiye地区）模型虽有效抑制部分未倒塌建筑物的识别，但仍存在误分的情况。针对上述情形，分析模型识别性能受限的可能原因有：1）训练数据集涵盖的震前震后场景较单一，缺少目标地物的多元变化特征；2）影像间存在空间分辨率与数据质量上的差异，使得模型易混淆完好建筑物与倒塌建筑物；3）在不同影像组成的场景中，模型对于地物偏移程度的适应性较弱。综上，模型在不同场景的泛化性能仍有待提升。由表4可知，使用平滑数据集训练后，模型在Fevaipasa、Nurdagi、Islahiye区域各精度指标分别平均提升约0.864%、1.304%、0.75%。但需要注意，当图像间空间分辨率相差较大时，一味进行平滑操作以追求梯度变化率的统一，会导致图像细节信息的缺失。因此，平滑操作更加适合多源影像间空间分辨率相差较小的情况。



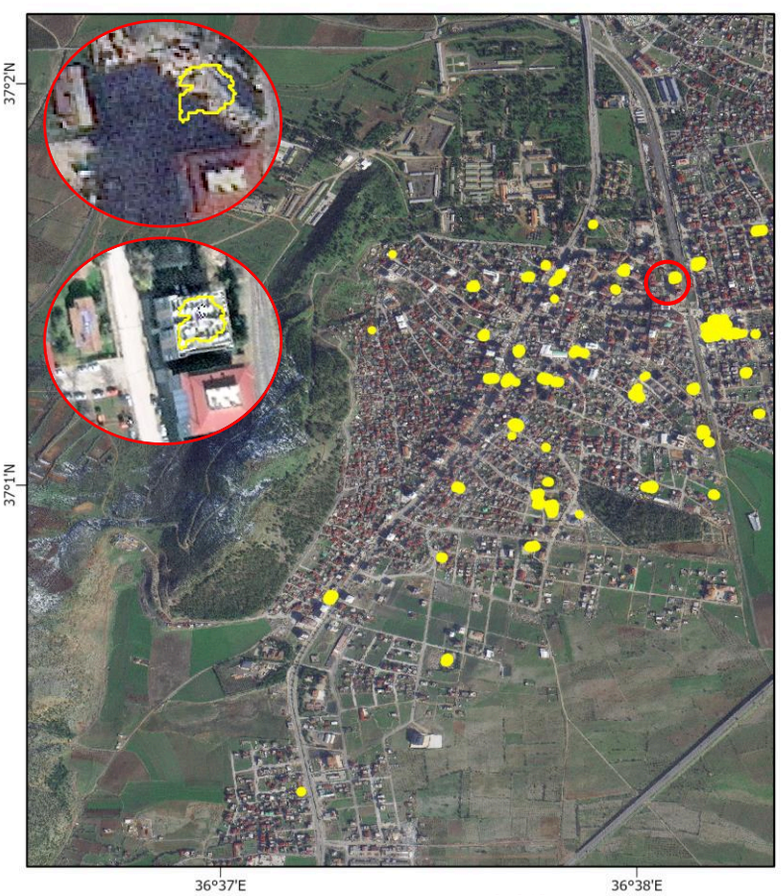
（a） Fevaipasa（底图为震后WorldView影像，黄框标注为SWSACNet识别的倒塌建筑物）

（a） Model inference results of Fevaipasa (base map provided by WorldView, yellow marker represents collapsed buildings extracted by SWSACNet)



（b） Nurdagi（底图为震后BJ-3影像，黄框标注为SWSACNet识别的倒塌建筑物）

（b） Model inference results of Nurdagi（base map provided by BJ-3，yellow marker represents collapsed buildings extracted by SWSACNet）



（c） Islahiye（底图为震后BJ-2影像，黄框标注为SWSACNet识别的倒塌建筑物）

（c） Model inference results of Islahiye（base map provided by BJ-2，yellow marker represents collapsed buildings extracted by SWSACNet）

图13 SWSACNet模型应用

Fig.13 Application of SWSACNet

表4　SWSACNet模型不同场景测试

Table 4　Testing of SWSACNet in different scenarios

	Precision	F1 score	Recall	OA	mIoU
Fevaipasa （平滑后模型精度涨幅）	87.24% （+0.57%）	61.21% （+0.84%）	47.14% （+0.83%）	97.9% （+1.2%）	44.1% （+0.88%）
Nurdagi （平滑后模型精度涨幅）	78.72%（+1.58%）	68.69% （+1.26%）	60.93%（+1.04%）	93.71%（+1.17%）	52.32% （+1.47%）
Islahiye （平滑后模型精度涨幅）	59.85%（+0.56%）	52.61% （+0.54%）	46.94%（+0.5%）	79.3%（+1.66%）	35.7% （+0.49%）

6　结论

针对震前震后遥感影像倒塌建筑物信息提取需求，本文提出端到端的、融合滑窗式特征增强和卷积注意力混合的变化检测网络模型（SWSACNet）变化检测深度学习网络模型。以土耳其地震为例，对比该模型与FTN、STANet、DASNet、FC-EF四类变化检测模型的倒塌建筑物提取精度。结果表明SWSACNet识别精度优于其他四类模型。同时对SWSACNet模型结构进行了三组纵向对照实验：ACmix+SWSFE、ACmix+FE、Swin Transformer+SWSFE。结果表明滑窗式增强结构（SWSFE）和ACmix模块的组合优于其他两种结构组合。研究发现，多源影像在分辨率一致后，影像间仍存在梯度变化率的差异，从而影响模型对目标地物的识别，针对该问题，提出通过平滑统一原始训练数据集的梯度变化率方法，经SWSACNet、FTN、STANet、DASNet、FC-EF五个模型测试，各模型识别精度均得到有效提升，证明了适当降低图像的梯度变化率，保持震前震后样本对的一致性有助于模型对目标地物的提取。

本文所构建模型精度相对同类模型有所提升，但震害数据集有限，模型在应用时仍有较大的提升空间。后续将收集更多国内外震例影像，扩增多源遥感影像建筑物震害数据集，并采用一对多的输入模式，即一个时相的影像对应第二个时相经数据增强后的多个影像作为输入，使模型最大程度地适应同地区不同时相在不同拍摄条件下的影像差异；考虑到数据源间的异质性，尝试将编码部分的孪生网络结构替换为震前震后影像的双编码结构并采用自适应的特征匹配等方法增强模型的鲁棒性。

志　谢　此次实验的影像数据由中国资源卫星应用中心、二十一世纪空间技术应用股份有限公司以及Maxar Technologies Inc.提供，在此表示衷心的谢意！

参考文献（Reference）

B. Adriano, N. Yokoya, J. Xia, H. Miura, W. Liu, M. Matsuoka, S. Koshimura, Learning from multimodal and multitemporal earth observation data for building damage mapping, ISPRS Journal of Photogrammetry and Remote Sensing 175 2021 132 – 143.

Bianet. Erdoğan: Earthquakes to cost Turkey 104 billion dollars[EB/OL]. (2023-3-21) [2024-5-17]. https://bianet. org/haber/erdogan-earthquakes-to-cost-turkey-104-billion-dollars-276043.

CGTN. Death toll from earthquakes rises to 50, 500 in Türkiye[EB/OL]. (2023-4-14) [2024-5-17]. https://news. cgtn. com/news/2023-04-14/Death-toll-from-earthquakes-rises-to-50-500-in-T-rkiye-1iZRoacvs2c/index.html.

Chen, Hao & Shi, Zhenwei. (2020). A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. Remote Sensing. 12. 1662. 10.3390/rs12101662.

Chen Meng, Wang Xiaoqing. The study on extraction of seismic damage of buildings from remote sensing image based on fully convolutional neural network[J]. Technology for Earthquake Disaster Prevention, 2019, 14(4): 810-820. (doi: 10.11899/zzfy20190412. 陈梦,王晓青.全卷积神经网络在建筑物震害遥感提取中的应用研究[J].震灾防御技术,2019,14(04):810-820.)

Daudt R C, Le Saux B, Boulch A, et al. Urban change detection for multispectral earth observation using convolutional neural networks[C]//IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. Ieee, 2018: 2115-2118.

Dong L., & Shan J. (2013) A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. ISPRS Journal of Photogrammetry and Remote Sensing, 84, 85-99.

Gao, M T. Turkey Earthquake: Another Alarm Bell for the Big Quake and Catastrophe[J]. China Emergency Management, 2023(02):34-36. (高孟潭.土耳其地震:大震巨灾再敲警钟[J].中国应急管理,2023(02):34-36.)

Ge J.; Tang H.; Ji C. Self-Incremental Learning for Rapid Identifica-

tion of Collapsed Buildings Triggered by Natural Disasters. Remote Sens. 2023, 15, 3909. https:// doi.org/10.3390/rs15153909.

Girshick R., Donahue J., Darrell T., & Malik J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

J. Chen et al.,"DASNet: Dual Attentive Fully Convolutional Siamese Networks for Change Detection in High-Resolution Satellite Images," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 1194-1206, 2021, doi: 10.1109/JSTARS.2020.3037893.

Krizhevsky A., Sutskever I., & Hinton G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

Pan X, Ge C, Lu R, et al. On the integration of self-attention and convolution[C]Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 815-825.

Rasika A.K., Kerle N., Heuel S.R.K.R., 2006. Multi-scale texture and color segmentation of oblique airborne video data for damage classification. ISPRS mid-term symposium remote sensing: from pixels to processes.

Reuters . Turkey quake kills 912 in historic disaster, Erdogan says [EB/OL] . (2023-2-6)[2024-5-17].https://www.reuters.com/world/middle-east/turkey-quake-kills-912-historic-disaster-erdogan-says-2023-02-06/.

Shorten C, Khoshgoftaar T M. A survey on image data augmentation for deep learning[J]. Journal of big data, 2019, 6(1): 1-48.

T. Miyamoto and Y. Yamamoto, "Using 3-D Convolution and Multimodal Architecture for Earthquake Damage Detection Based on Satellite Imagery and Digital Urban Data," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 8606-8613, 2021, doi: 10.1109/JSTARS.2021.3102701.

Wang T L, Jin Y Q. 2012. Evaluation of multiple mutual information for building damages after earthquake using preevent optical image and post-event SAR image. Journal of Remote Sensing, 16(2): 248-261. (王天涵,金亚秋.地面建筑物破坏状态检测的多类互信息量评估——震前光学图像与震后SAR图像的融合[J].遥感学报,2012,16(02):248-261.)

Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE transactions on image processing, 2004, 13(4): 600-612.

Wu Y. Research on the Method of Extracting the Building Seismic Damage Information Based on Improved U-NET[D]. Southwest Jiaotong University, 2020. DOI: 10.27414/d. cnki. gxnju. 2020. 001661.(伍燚垚.基于改进U-Net的建筑物震害信息提取方法研究[D].西南交通大学,2020.DOI: 10.27414/d.cnki.gxnju.2020.001661.)

Yan T, Wan Z, Zhang P. Fully Transformer Network for Change Detection of Remote Sensing Images[C]Proceedings of the Asian Conference on Computer Vision. 2022: 1691-1708.

Y. Shen et al., "BDANet: Multiscale Convolutional Neural Network With Cross-Directional Attention for Building Damage Assessment From Satellite Images," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-14, 2022, Art no. 5402114, doi: 10.1109/TGRS.2021.3080580.

Zheng Z, Zhong Y F, Wang J, Ma A L and Zhang L P. 2021. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: from natural disasters to man-made disasters. Remote Sensing of Environment, 265: 112636 [DOI: 10.1016/j.rse.2021.112636]

SWSACNet：A change detection network for collapsed buildings extraction using multi−source images

Long Ying[1], Dou Aixia[1*], Wang Feifei[2], Wang Shumin[1]

1.Institute of Earthquake Forecasting, China Earthquake Administration, Beijing, 100036, China;
2.Henan Earthquake Agency, Zhengzhou, 450016, China

Abstract：Objective Change detection networks based on deep learning are widely used in water monitor, urban change, etc.. However, collapsed buildings, as one of the change objectives, are rarely targeted for change detection networks. This study proposes an end-to-end collapsed building extraction model based on a change detection network including the sliding-window feature enhancement and convolution attention mix mechanism, which called SWSACNet (Sliding-Window-Shift Attention Convolution mix Network). Method SWSACNet is an improvement of Fully Transformer Network (FTN). FTN is a network completely composed of Swin Transformer. Besides, it has a unique frame, which involves four parts: SFE (Siamese Feature Extraction), DFE (Deep Feature Enhancement), PCP (Progressive Change Prediction), DS (Deep Supervision). By encoding and decoding the feature of change objects deeply, FTN is able to learn what the collapsed buildings has changed in two temporal images and suppress irrelevant information. ACmix, a blend of convolution and attention mechanism, has been proved better performance than Swin Transformer in mainstream datasets. However, due to the different

sensors, platforms, etc., the spatial heterogeneity of target features in different source remote sensing images will affect the accuracy of change detection. Concerning this problem, we designed a similarity sliding window to match the feature maps of two temporal images. Hence, we replace Swin Transformer with ACmix to extract and restore earthquake-damaged features efficiently in the phase of SFE and PCP, and using similarity sliding window to reduce misidentifications of collapsed buildings in different source image pairs before the phase of DFE. Result Taking the earthquake with 7.8 magnitude on February 6th, 2023, in Turkey as an example, establish a building seismic damage change detection dataset which consists of pre-earthquake Gaofen-2, Google images and post-earthquake Beijing-3 images, and collapsed buildings were extracted based on the SWSACNet, FTN, STANet based on the Siamese self-attention mechanism, DASNet based on a dual-attention fully-convolutional neural network, and the conventional fully-convolutional early fusion FC-EF network. The experimental results show that SWSACNet achieves the highest accuracy with F1 score of 80.8% and mIoU of 67.8%. The ablation experiments of the improved model indicates that SWSACNet obtains highest precision among three structure combinations. Beyond that, by smoothing the BJ-3 image that has higher spatial resolution to make the gradient change rate of image pairs closer, we acquire a new dataset and use it to retrain the five models. We found that the precision of five retrained models increases 1% at least, which also illustrates that appropriately narrowing the gap of gradient change rate of image pairs is an effective preprocessing for models to recognize the collapsed buildings. Finally, applying SWSACNet to three different data combinations covering Fevaipasa, Nurdagi and Islahiye area, the results show that it achieves 60.84% of average F1 score. Conclusion The application of SWSACNet indicates that the model needs richer pre- and post-earthquake training dataset and structural improvement to enhance its generalization.

Key words: multi-source images, deep learning, change detection, collapsed building extraction

Supported by Supported by National Natural Science Foundation of China (No.42271090); Fundamental Research Funds of the Institute of Earthquake Forecasting, CEA (CEAIEF2022050504, CEAIEF20230202)