引文格式: 冯炜明,张新长,孙颖,等. 融合 Transformer 结构的高分辨率遥感影像变化检测网络[J]. 测绘通报,2022(8):36-40. DOI: 10.13474/j.cnki.11-2246.2022.0229.

融合 Transformer 结构的高分辨率遥感影像变化检测网络

冯炜明[1],张新长[1],孙 颖[2],姜 明[1],甘 巧[1],侯幸幸[1]
(1. 广州大学地理科学与遥感学院,广东 广州 510006; 2. 中山大学地理科学与规划学院,广东 广州 510275)

摘要: 为解决遥感影像变化检测全局上下文信息捕获的问题,本文提出了基于孪生结构、跳跃连接结构及 Transformer 结构的 TSU-Net。该模型编码器采用混合 CNN-Transformers 结构,借助自注意力机制捕获遥感影像的全局上下文信息,增强了模型对于像素级遥感影像变化检测任务的长距离上下文建模能力。该模型在 LEVIR-CD 数据集和 CDD 数据集进行测试,$F1$ 得分分别为 90.73 和 93.14,优于各对比模型。

关键词: 深度学习; 遥感影像变化检测; Transformer; TSU-Net
中图分类号: P237 **文献标识码:** A **文章编号:** 0494-0911(2022)08-0036-05

High-resolution remote sensing image change detection network with Transformer structure

FENG Weiming[1],ZHANG Xinchang[1],SUN Ying[2],JIANG Ming[1],GAN Qiao[1],HOU Xingxing[1]
(1. School of Geography and Remote Sensing, Guangzhou University, Guangzhou 510006, China; 2. School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: In order to solve the problem of global context information capture in remote sensing image change detection, this paper proposes TSU-Net based on twin structure, jump link structure and Transformer structure. The model encoder adopts a hybrid CNN-Transformers structure to capture the global context information of remote sensing images with the help of self-attention mechanism, which enhances the model's ability of long distance context modeling for pixel-level remote sensing image change detection task. The model is tested in the LEVIR-CD dataset and the CDD dataset, and the $F1$ scores are 0.907 3 and 0.931 4, respectively, which are superior to the comparison models.

Key words: deep learning; remote sensing image change detection; Transformer; TSU-Net

遥感影像变化检测是指在地理实体或场景中,通过多种手段检测并提取变化信息的技术[1]。深度学习具有强大的特征提取能力,可以自动、全面地提取依靠人工设计无法取得的高维抽象地物特征,因此在遥感影像变化检测领域已成为研究热点[2-3]。

文献[4]提出了基于早期融合方法的 FC-EF 及具有孪生结构的 FC-Siam-Conc,这两个网络在遥感影像变化检测领域具有重要地位。文献[5]提出了改进型 U-Net++,将密集连接机制与深监督机制融入 U-Net++,其融合了不同尺度的输出,增强了预测图的边缘细节。文献[6]提出了融合孪生结构与时空注意力机制的 STA-Net,使影像变化检测增强了对网络捕捉多尺度时空注意力的依赖。文献[7]提出了 BIT,将 Transformer 结构作为模型的主要编码与解码器,提升了模型对于时空信息的全局上下文建模能力。为了解决多尺度目标的漏检和边界粗糙问题,文献[8]提出了注意力金字塔网络。上述网络在遥感影像变化检测任务中性能优异,但仍不可避免地受到地物形态、场景复杂度、成像条件及配准误差等噪声干扰,出现漏检及虚警,而影像的时空上下文信息可缓解上述干扰[7]。

Transformer 结构可有效捕捉全局上下文信息,融合 Transformer 结构的 CNN 模型,其特征提取能力、全局感受野及长范围上下文依赖建模能力均得到了提升[7,9-10],更适用于遥感影像变化检测任务。

综合以上分析,本文提出具有孪生结构、跳跃

收稿日期: 2022-02-17
基金项目: 国家重点研发计划(2018YFB2100702); 国家自然科学基金面上项目(42071441); 智慧广州时空信息云平台关键技术——时空大数据一体化整合与自适应动态更新模块研发与咨询服务项目(GZIT2016-A5-147)
作者简介: 冯炜明(1997—),男,硕士生,主要研究方向为智慧城市与城市遥感。E-mail: 942846432@qq.com
通信作者: 张新长。E-mail: zhangxc@gzhu.edu.cn

连接结构的 Transformer 网络（Transformer Siamese U-Net，TSU-Net），结合 CNN 结构的特征提取能力与 Transformer 结构的全局上下文信息提取能力，增强模型对于像素级遥感影像变化检测任务的长距离上下文建模能力，并输出高分辨率的变化预测图。

1　原理和方法

1.1　跳跃连接结构

U-Net[11] 是在全卷积网络（FCN）基础上改进的端到端网络，其跳跃连接结构将浅层特征与深层特征相融合，可获得高分辨率的像素级语义分割影像，对高分辨率的语义分割网络设计有重要参考意义。

1.2　孪生结构

SiameseNet[12] 具有其权值共享的孪生主干特征提取网络，可对两个输入数据分别进行特征提取，有助于神经网络生成影像差异图，目前已有研究将其应用于遥感影像变化检测[4,13-14]。

1.3　Transformer 结构

Transformer 结构源于自然语言处理领域，该结构的优点是具有全局自注意力机制，可有效捕获输入数据的全局上下文信息。在计算机视觉领域应用该结构时，需要先进行 Patch Embedding 操作，如图 1 所示。首先将图片转换为图块序列，通过输入编码操作，将各图块含有的信息转换为高维向量；然后提取该向量的语义信息并通过位置编码操作新增位置信息。

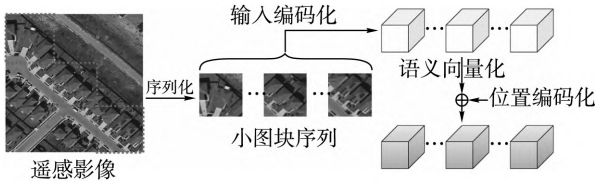


图 1　Patch Embedding 操作

Transformer 结构包括多头注意力模块、多层感知机模块、标准化层及残差模块，如图 2 所示。其中，多头注意力模块是 Transformer 结构的核心部分；多层感知机模块包括两个全连接层和一个 GeLU 激活函数；标准化层有助于调整特征分布，加快模型收敛速度，参考 ViT 网络[15]，置于多头注意力模块及多层感知机模块前；残差模块建立了短路连接，使得 Transformer 结构内部拥有恒等映射能力，避免网络退化。

Transformer 结构的多头注意力模块由多个自注意力机制组成，各个自注意力机制可捕获不同类型的上下文信息，有助于 Transformer 结构得到多层次的长范围上下文依赖信息。自注意力机制输入的为 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 矩阵，可通过输入矩阵分别与权重矩阵 $\boldsymbol{W}^q$、$\boldsymbol{W}^k$、$\boldsymbol{W}^v$ 相乘求得，公式为

$$\boldsymbol{Q} = \boldsymbol{T}^{(l-1)}\boldsymbol{W}^q \tag{1}$$

$$\boldsymbol{K} = \boldsymbol{T}^{(l-1)}\boldsymbol{W}^k \tag{2}$$

$$\boldsymbol{V} = \boldsymbol{T}^{(l-1)}\boldsymbol{W}^v \tag{3}$$

式中，$\boldsymbol{T}^{(l-1)}$ 表示前一层 Transformer 输出或 Patch Embedding 输出；$\boldsymbol{W}^q$、$\boldsymbol{W}^k$、$\boldsymbol{W}^v$ 表示对应权重矩阵，是 3 个可训练线性参数层。

自注意为模块计算公式为

$$\mathrm{Att}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \sigma\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\cdot\boldsymbol{V} \tag{4}$$

式中，σ 表示 softmax 激活层；d 表示自注意力机制的维度。

将多个自注意力机制的输出值叠加并经过线性转换后，最终得到多头注意力模块的输出值，公式为

$$\mathrm{head}_j = \mathrm{Att}(\boldsymbol{T}^{(l-1)}\boldsymbol{W}_j^q, \boldsymbol{T}^{(l-1)}\boldsymbol{W}_j^k, \boldsymbol{T}^{(l-1)}\boldsymbol{W}_j^v) \tag{5}$$

$$\mathrm{MSA} = \mathrm{MSA}(\boldsymbol{T}^{(l-1)}) = \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, \cdots, \mathrm{head}_n)\boldsymbol{W}^o \tag{6}$$

中，$\boldsymbol{W}_j^q$、$\boldsymbol{W}_j^k$、$\boldsymbol{W}_j^v$ 分别表示第 j 个自注意力头的 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 权重矩阵；$\boldsymbol{W}^o$ 表示可训练的线性权重矩阵；n 表示自注意力头的总数。

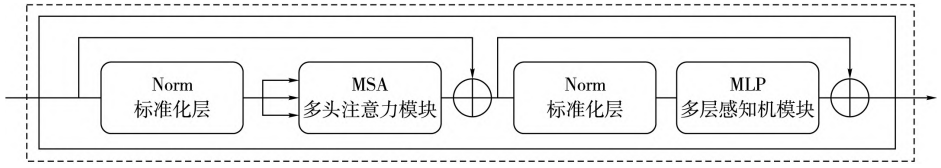


图 2　Trasnformer 结构

1.4　TSU-Net

TSU-Net 的编码器采用混合 CNN-Transformers 结构；解码器部分采用跳跃连接结构。模型整体结构如图 3 所示。

编码器可获得浅层变化特征与深层变化特征。通过孪生结构将两张不同时相的遥感影像，分别输入编码器的两个特征提取分支。TSU-Net 通过孪生 CNN 结构获取 4 对用于解码阶段的高分辨率影像特征图。为减少训练参数量，其中一个分支继续下采样以缩小特征图尺寸，并将该特征图传入由 6 个

Transformer 结构组成的 Transformers 提取影像的全局上下文信息。其中，每个多头注意力模块含有 6 个自注意力机制。

解码器可将浅层变化特征与深层变化特征相融合。首先对 Transformers 输出的全局上下文信息进行双线性插值上采样操作；然后将该特征图与来自编码器上一阶段的特征图叠加，叠加后的变化差异图同时具有来自 CNN 的浅层变化特征和来自 Transformers 的深层变化特征；最后依次进行上采样操作扩大差异图尺寸，借助跳跃连接结构将 4 对来自编码器的特征图融合，逐步将高维变化语义信息还原为高分辨率变化特征图。

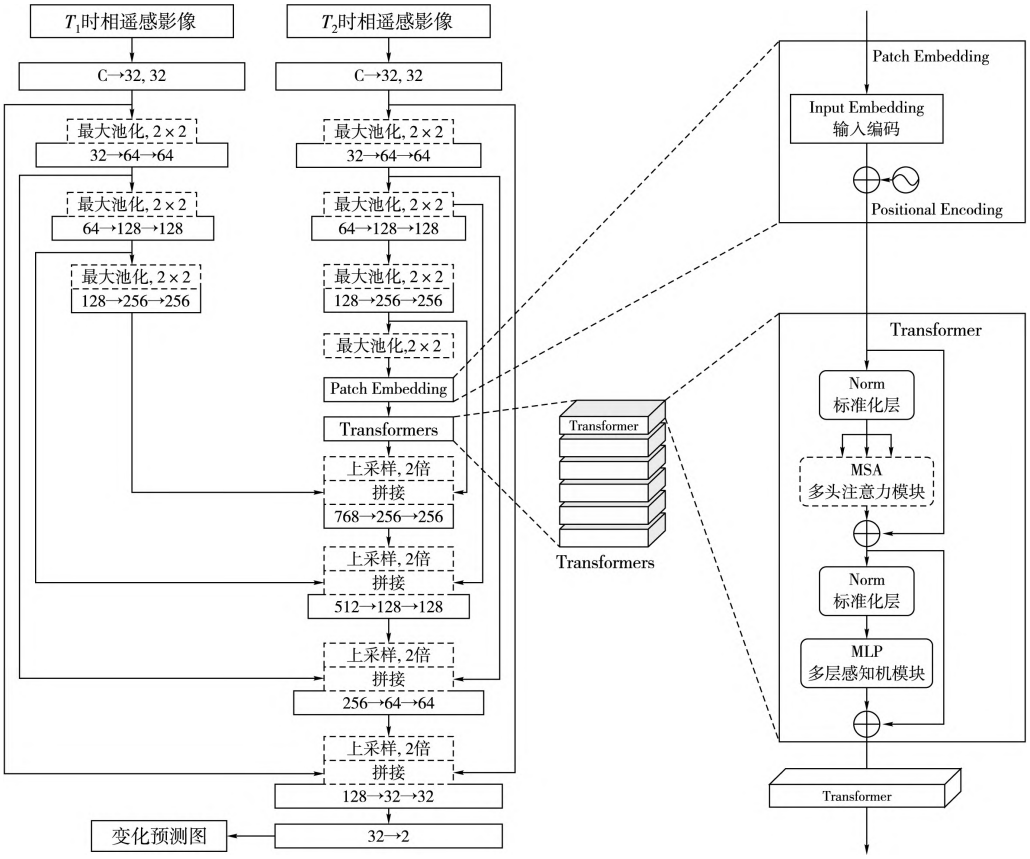


图 3 TSU-Net 模型结构

2 试验和结果

2.1 数据集

该模型采用 LEVIR-CD 数据集和 CDD 数据集进行验证。LEVIR-CD 数据集来自 Google Earth API，影像覆盖 20 个区域，分辨率为 50 cm，尺寸为 1024×1024 像素，数据集每个区域的时间跨度在 5～14 年。该数据集涵盖各种类型的建筑物，如别墅、高层公寓、小型车库及大型仓库等。CDD 数据集来自 Google Earth 影像，影像分辨率在 3～100 cm，尺寸为 256×256 像素，包括 10 000 对训练集、3000 对测试集及 3000 对验证集。地物变化类型丰富多样，包括建筑物、汽车、土地、道路及仓库等。

上述两个数据集的时间跨度较长，均考虑了遥感影像的季节差异及光照差异，可有效检验 TSU-Net 对遥感影像变化检测任务的性能。

2.2 相关参数与评价指标

TSU-Net 模型基于 Pytorch 框架实现，试验所采用的显卡为 NVIDA RTX 6000，内存为 24 GB。模型的超参数如下：训练轮数为 100 轮，选择全局最优模型；优化器为 Adam，学习率为 0.000 5；参考现有研究[5]，损失函数采用由 Balanced Binary Cross-Entropy Loss 与 Dice Coefficient Loss 组合而成的 Hybird_Loss，该混合损失函数考虑了样本类别不平衡的问题，更适用于变化检测任务。

为检验模型性能，本文采用混淆矩阵作为模型性能评价方法，指标包括总体精度、精确度、召回率、MIoU 及 $F1$ 得分。对于变化检测任务而言，精确度越高，表示虚警数量越少；召回率越高，表示误检数量越少；而总体精度、$F1$ 得分及 MIoU 也可揭示模型性能的总体表现，值越高表示模型性能越好。上述指标的详细信息见表 1。

表 1　预测结果评价指标

评价指标	公式	含义
总体精度（OA）	$OA = \frac{TP + TN}{TP + TN + FP + FN}$	正确预测像素在总像素中占比
精确度（P）	$P = \frac{TP}{TP + FP}$	正确预测像素在总正类中占比
召回率（R）	$R = \frac{TP}{TP + FN}$	正确预测像素在正类中占比
$F1$ 得分（$F1$）	$F1 = \times \frac{2PR}{P + R}$	模型综合性能评价
交并比（MIoU）	$MIoU = \frac{TP}{TP + FP + FN}$	全局评价

注：TP 表示正确检测的变化像素数量；TN 表示正确检测的未变化像素数量；FN 表示漏检像素数量；FP 表示虚警的像素数量。

2.3　结果与分析

为检验 TSU-Net 性能，选取了 FC-EF、FC-Siam-Conc、改进型 U-Net++、STANet 及 BIT 与本文模型进行比较。

首先进行评价指标对比，两个数据集的比较结果见表 2—表 3。针对 $F1$ 得分与精确度指标对比评价模型，TSU-Net 在两个数据集中 $F1$ 得分均最高。在 LEVIR-CD 数据集中，$F1$ 得分为 0.907 3，比 FC-EF 提升了 3.72%，比 BIT 提升了 0.86%。在 CDD 数据集中，$F1$ 得分为 0.931 4，比 FC-EF 提升了 6.60%，比 BIT 提升了 0.38%。在精确度方面，对于 LEVIR-CD 数据集，TSU-Net 取得各模型中的最高精确度；对于 CDD 数据集，TSU-Net 的精确度与 BIT 相比，降低了 1.54%，一方面是 CDD 数据集的变化场景较为复杂的原因，另一方面则表明 TSU-Net 对于正类变化像素的检测能力与虚警的处理能力有待提升。

表 2　基于 LEVIR-CD 数据集的不同模型预测结果评价

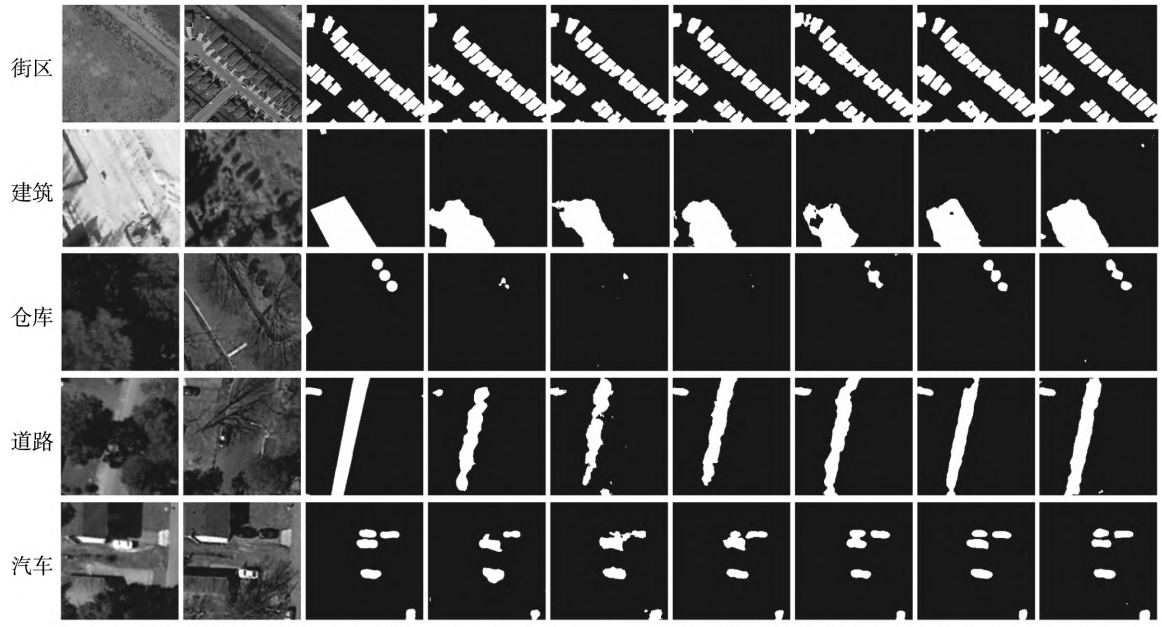
数据集	模型	精确度	召回率	$F1$ 得分	交并比	总体精度
LEVIR-CD 数据集	FC-EF	0.883 0	0.857 6	0.870 1	0.878 2	0.987 0
	FC-Siam-Conc	0.892 8	0.884 7	0.888 7	0.894 0	0.988 7
	改进型 U-Net++	0.904 6	0.878 0	0.891 1	0.896 1	0.989 1
	STANet	0.901 5	0.884 1	0.892 7	0.897 4	0.989 2
	BIT	0.913 1	0.884 7	0.898 7	0.902 7	0.989 9
	TSU-Net（本文模型）	**0.917 0**	**0.897 8**	**0.907 3**	**0.910 2**	**0.990 7**

表 3　基于 CDD 数据集的不同模型预测结果评价

数据集	模型	精确度	召回率	$F1$ 得分	交并比	总体精度
CDD 数据集	FC-EF	0.908 4	0.826 3	0.865 4	0.863 7	0.968 3
	FC-Siam-Conc	0.925 5	0.841 4	0.881 5	0.878 4	0.972 0
	改进型 U-Net++	0.951 2	0.870 1	0.908 8	0.904 4	0.978 5
	STANet	0.941 1	0.876 7	0.907 8	0.903 2	0.978 0
	BIT	**0.979 2**	0.881 0	0.927 6	0.923 0	0.983 0
	TSU-Net（本文模型）	0.963 8	**0.901 0**	**0.931 4**	**0.926 6**	**0.983 6**

注：黑体加粗表示最佳结果，下划线表示次优结果。

然后进行房屋、小型建筑、汽车及道路场景的可视化比较，结果如图 4 所示。整体而言，TSU-Net 的预测效果（如图 4（i）所示）与真实标签（如图 4（c）所示）最为接近。与其他模型相比，TSU-Net 不仅可准确地预测出建筑物变化的边缘部分，而且预测图的空洞更少。TSU-Net 对于小型建筑和汽车等小型地物的变化检测效果较优，能较为准确地刻画出边界。此外，道路变化也是遥感影像变化检测的重要部分，通过对比模型的预测结果可以发现，TSU-Net 道路变化预测图的连续性及平滑性与标签最接近。



（a）T_1时相 （b）T_2时相 （c）标签 （d）FC-EF （e）FC-Siam-Conc （f）改进型 U-Net++ （g）STANet （h）BIT （i）TSU-Net

图 4 5 个场景对比试验

3 结 语

针对高分辨率遥感影像变化检测任务，本文提出了 TSU-Net。该网络融合具有自注意力机制的 Transformer 结构、跳跃连接结构及孪生结构。模型可从遥感影像中获取全局上下文信息，并通过跳跃连接结构将 CNN 获取的高分辨率特征图与全局上下文信息相融合，得到具有精确边界轮廓的遥感影像变化检测图。使用 LEVIR-CD 公开数据集和 CDD 公开数据集对模型性能进行检测。对于 LEVIR-CD 数据集，$F1$ 得分为 0.907 3；对于 CDD 数据集，$F1$ 得分为 0.931 4。由此可见，TSU-Net 在高分辨率遥感影像变化检测中效果较好。

在今后的工作中，可继续改良网络结构，考虑为模型加入深监督机制、金字塔机制及密集连接机制等，进一步提升 TSU-Net 对于正类变化像素的检测能力及虚警的处理能力，增强模型对于复杂场景的变化检测水平。

参考文献：

[1] SINGH A. Review article digital change detection techniques using remotely-sensed data [J]. International Journal of Remote Sensing, 1989, 10(6): 989-1003.

[2] 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)：1885-1898.

[3] 张良培，武辰. 多时相遥感影像变化检测的现状与展望 [J]. 测绘学报，2017，46(10)：1447-1459.

[4] CAYE DAUDT R, LE SAUX B, BOULCH A. Fully convolutional Siamese networks for change detection [C]//Proceedings of the 25th IEEE International Conference on Image Processing. Greece: IEEE, 2018: 4063-4067.

[5] PENG Daifeng, ZHANG Yongjun, GUAN Haiyan. End-to-end change detection for high resolution satellite images using improved UNet++ [J]. Remote Sensing, 2019, 11(11): 1382.

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

[7] CHEN Hao, QI Zipeng, SHI Zhenwei. Remote sensing image change detection with transformers [J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-14.

[8] 田青林，秦凯，陈俊，等. 基于注意力金字塔网络的航空影像建筑物变化检测 [J]. 光学学报，2020，40(21)：47-56.

[9] CHEN01 Jieneng, LU Yongyi, YU Qihang, et al. TransUNet: transformers make strong encoders for medical image segmentation [EB/OL]. [2022-02-01]. https://arxiv.org/abs/2102.04306.

[10] CHEN Kaiqiang, FU Kun, GAO Xin, et al. Building extraction from remote sensing images with deep learning in a supervised manner [C]//Proceedings of 2017 IEEE International Geoscience and Remote Sensing Symposium. Fort Worth: IEEE, 2017: 1672-1675.

（下转第 92 页）

integrated system of video surveillance and GIS [J]. IOP Conference Series: Earth and Environmental Science, 2018, 170: 022088.

[5] 刘垒. 基于地理约束场景的动态目标检测及行为识别方法研究 [D]. 桂林: 桂林理工大学, 2020.

[6] 韩世静, 苗书锋, 郝向阳, 等. 监控视频与地理信息融合技术发展综述 [J]. 测绘通报, 2022(5): 1-6.

[7] HAN Shijing, DONG Xiaorui, HAO Xiangyang, et al. Extracting objects' spatial-temporal information based on surveillance videos and the digital surface model [J]. ISPRS International Journal of Geo-Information, 2022, 11(2): 103.

[8] MILOSAVLJEVIĆ A, RANČIĆ D, DIMITRIJEVIĆ A, et al. A method for estimating surveillance video georeferences [J]. ISPRS International Journal of Geo-Information, 2017, 6(7): 211.

[9] 徐汝东, 刘进. 可量测视频目标动态轨迹生成及 GIS 应用 [J]. 武汉大学学报(信息科学版), 2016, 41(6): 818-824.

[10] XIE Yujia, WANG Meizhen, LIU Xuejun, et al. Integration of GIS and moving objects in surveillance video [J]. ISPRS International Journal of Geo-Information, 2017, 6(4): 94.

[11] XIE Yujia, WANG Meizhen, LIU Xuejun, et al. Integration of multi-camera video moving objects and GIS [J]. ISPRS International Journal of Geo-Information, 2019, 8(12): 561.

[12] 张旭, 郝向阳, 李建胜, 等. 监控视频中动态目标与地理空间信息的融合与可视化方法 [J]. 测绘学报, 2019, 48(11): 1415-1423.

[13] 李景文, 韦晶闪, 姜建武, 等. 多视角监控视频中动态目标的时空信息提取方法 [J]. 测绘学报, 2022, 51(3): 388-400.

[14] MILOSAVLJEVIĆ A, DIMITRIJEVIĆ A, RANČIĆ D. GIS-augmented video surveillance [J]. International Journal of Geographical Information Science, 2010, 24(9): 1415-1433.

[15] 张兴国. 地理场景协同的多摄像机目标跟踪研究 [D]. 南京: 南京师范大学, 2014.

[16] 唐丽玉, 王熠中, 陈崇成, 等. 视频图像中监控目标的空间定位方法 [J]. 福州大学学报(自然科学版), 2014, 42(1): 55-61.

[17] HARTLEY R, ZISSERMAN A. Multiple view geometry incomputer vision [M]. 2nd ed. Cambridge, UK: Cambridge University Press, 2003.

[18] ZHANG Z. A flexible new technique for camera calibration [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(11): 1330-1334.

[19] LEPETIT V, MORENO-NOGUER F, FUA P. EPnP: an accurate O(n) solution to the PnP problem [J]. International Journal of Computer Vision, 2009, 81(2): 155.

[20] PENATE-SANCHEZ A, ANDRADE-CETTO J, MORENO-NOGUER F. Exhaustive linearization for robust camera pose and focal length estimation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(10): 2387-2400.

(责任编辑: 杨瑞芳)

(上接第 40 页)

[11] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation [M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015: 234-241.

[12] BROMLEY J, GUYON I, LECUN Y, et al. Signature verification using a "Siamese" time delay neural network [C]//Proceedings of the 6th International Conference on Neural Information Processing Systems. New York: ACM, 1993: 737-744.

[13] 向阳, 赵银娣, 董霁红. 基于改进 UNet 孪生网络的遥感影像矿区变化检测 [J]. 煤炭学报, 2019, 44(12): 3773-3780.

[14] 倪良波, 卢涵宇, 卢天健, 等. 基于孪生残差神经网络的遥感影像变化检测 [J]. 计算机工程与设计, 2020, 41(12): 3451-3457.

[15] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale [EB/OL]. 2020: arXiv: 2010.11929. https://arxiv.org/abs/2010.11929.

(责任编辑: 胡 淼)