



BIM 增强的超大基坑施工场地图像语义分割算法

杨逸枫¹, 李康¹, 刘颖彬², 廖少明², 王威³

(1. 国网电力工程研究院有限公司, 北京 100055; 2. 同济大学土木工程学院, 上海 200092;

3. 上海城建市政工程(集团)有限公司, 上海 200065)

摘要:为解决现有的基于图像分类、目标检测等算法的施工场地图像智能识别方法无法识别工人、施工机械等目标对象准确数量与精确边界的问题,引入语义分割算法,以超大规模基坑为应用场景,构建一套建筑信息模型(BIM)增强的语义分割图像数据集、开发基于DeepLabv3+的语义分割算法,并提出基于语义分割结果的目标对象识别方法。首先,基于宁波市某超大基坑工程采集施工场地实拍图像,并通过BIM技术生成大量与实际施工场景类似的效果图像以增强数据集。选取工人、挖掘机、卡车、混凝土搅拌车、混凝土泵车、吊车、围栏、塔吊、混凝土支撑及背景共10类基坑施工场地中常见的对象作为语义分割的目标对象,对所有图像进行像素级的精细标注。在构建的图像数据集上以DeepLabv3+网络为基础开发语义分割算法,并通过对比实验研究不同主干特征提取网络与损失函数组合对于算法性能的影响,以及效果图像对于算法性能的提升效果。结果表明:采用Xception作为主干特征提取网络和Dice loss+CE loss作为损失函数的DeepLabv3+模型分割效果最佳,平均交并比($\bar{I}_{ou}$)值达到68.24%,能基本实现在近景、远景、不同的天气及环境光线条件下对工人、机械设备、结构等对象的像素级分割。采用基于BIM生成的效果图训练模型对模型分割效果具有显著的提升作用,与仅采用实拍图相比模型的 $\bar{I}_{ou}$ 值提高了26.27%。基于语义分割结果的目标对象识别方法能准确输出基坑施工场地图像中的支撑、机械设备等对象的类型与数量。

关键词:超大基坑;语义分割;深度学习;建筑信息模型;DeepLabv3+

中图分类号:TU94;TU17 **文献标志码:**A **文章编号:**2096-6717(XXXX)XX-0001-12

BIM-enhanced semantic segmentation algorithm for construction scenes of oversized excavation sites

YANG Yifeng¹, LI Kang¹, LIU Yingbin², LIAO Shaoming², WANG Wei³

(1. State Grid Electric Power Engineering Research Institute Co., Ltd., Beijing 100055, P. R. China; 2. College of Civil Engineering, Tongji University, Shanghai 200092, P. R. China; 3. Shanghai Urban Construction Municipal Engineering (Group) Co., Ltd., Shanghai 200065, P. R. China)

Abstract: To address the problem that existing image classification and object detection based intelligent

收稿日期:2025-07-07

基金项目:国家自然科学基金(52090082)

作者简介:杨逸枫(1998-),男,博士,主要从事深大基坑智能监控与风险评估研究,E-mail:yangyifeng@ei.sgcc.com.cn。

刘颖彬(通信作者),男,博士,E-mail:liuyb_sd@foxmail.com。

Received: 2025-07-07

Foundation items: National Natural Science Foundation of China (No. 52090082)

Author brief: YANG Yifeng (1998-), PhD, main research interest: intelligent monitoring and risk assessment for deep excavations, E-mail: yangyifeng@ei.sgcc.com.cn.

LIU Yingbin (corresponding author), PhD, E-mail: liuyb_sd@foxmail.com.

recognition methods cannot accurately identify the accurate quantities and precise boundaries of target objects such as workers and equipment in construction site images, this paper introduces semantic segmentation, creates an image dataset enhanced with building information modeling (BIM), develops a semantic segmentation algorithm based on DeepLabv3+ and proposes a target object recognition method based on semantic segmentation results, by taking oversized excavation sites as the application scenario. First, real images of excavation sites were captured from an oversized excavation project in Ningbo City, and a large number of synthetic images similar to actual excavation scenes were generated using BIM technology to enhance the dataset. Worker, excavator, truck, mixer truck, pump truck, crane, fence, tower crane, strut, and background, which are common objects in the 10 categories of excavation sites, were selected as target objects for semantic segmentation, and all images were accurately labeled at the pixel level. Based on the DeepLabv3+ network, a semantic segmentation algorithm was developed on the created image dataset, and the impact of different backbone networks and loss function combinations on algorithm performance, as well as the effect of synthetic images on algorithm performance improvement, were studied through comparative experiments. The results show that the DeepLabv3+ model with Xception as the backbone network and Dice loss+CE loss as the loss function performs the best, with a mean intersection over union ($\bar{I}_{ou}$) value of 68.24%, which can achieve pixel-level segmentation of workers, equipment, structures, and other objects under different weather and environmental lighting conditions in close-up and distant views. The model trained on synthetic images generated based on BIM has a significant improvement effect on the model segmentation performance, with an $\bar{I}_{ou}$ value increase of 26.27% compared to the model trained only on real images. The object recognition method can accurately output the types and quantities of objects such as supports and mechanical equipment in excavation site images.

Keywords: oversized excavation; semantic segmentation; deep learning; building information modeling(BIM); DeepLabv3+

随着新型城镇化建设的不断推进,城市地下空间开发作为实现城市可持续发展目标的重要途径受到了广泛重视,同时促生了大量超大规模的基坑工程。与普通基坑工程相比,大规模基坑工程面积大,建设周期长,风险因素多、分布广,实现实时、无死角的现场监控对于其潜在风险的识别与管理至关重要。传统的人工巡检方法由于效率低、主观性强且存在巡检盲点,已难以满足大规模基坑工程现场监控的需求。

基坑施工场地图像是—类获取成本低廉又蕴含丰富信息的数据。—方面,近年来随着手机、摄像头、无人机等图像采集设备的迅速发展,基坑工程施工场地的实景图像可以很方便地获取。另—方面,施工场地图像蕴含丰富的信息,如工程施工进度、效率、安全状态等,基于图像数据开发智能识别算法是实现大规模基坑工程现场实时、无死角监控的有效途径。目前已经有大量研究提出了基于计算机视觉的施工场地图像智能识别方法,用于识别场地异常施工活动^[1-2],检测工人是否佩戴安全帽^[3]等。

尽管目前已有大量研究涉及施工场地图像的智能识别,但是其主要局限在图像分类、目标检测

等较为简单的计算机视觉任务,无法得到图像中目标对象的精确边界信息。例如,传统研究基于目标检测识别施工场地图像中目标对象之间的距离,其—般方法是将目标对象视为质点,通过计算目标检测算法得到的目标对象边界框中心之间的距离,作为目标对象之间的距离。然而目标对象之间的实际距离应为边缘之间的距离,对于车辆、设备等大型施工设备来说采用这种传统的基于质心的距离识别方法会带来显著的误差。因此,引入语义分割算法对施工场地图像进行像素级的分类,实现对目标对象具体形状的精确识别,对于施工场景深入、精细化的智能识别具有重要意义。

目前大规模基坑工程施工场地图像的语义分割研究主要难点在于:缺乏与基坑施工场景相关的采用语义分割级别标注的公开图像数据集;大规模基坑工程施工场景十分复杂,包含人员、机械、材料、结构等各类对象,且图像中存在各类对象互相遮挡情况,难以精确识别;大规模基坑工程由于面积大,远距离获取的图像或视频中存在很多细小物体难以精确识别;与目标检测相比,语义分割任务计算成本显著增加,实现实时的识别速度有很大挑战。

笔者首先针对工程领域缺乏开源的大规模基坑施工场地图像语义分割数据集问题,以宁波市某超大基坑工程为背景,采用手机拍摄实景图像、基于 BIM 的计算机软件合成效果图像两种方式收集并标注了大量图像,构建基坑施工场地图像语义分割数据集;在图像数据集上以 DeepLabv3+ 网络为基础开发语义分割算法;通过对比实验研究不同主干特征提取网络与损失函数组合对算法性能的影响,以及效果图像对算法性能的提升效果;提出基于语义分割结果的目标对象类别与数量识别方法,并实际验证了其识别效果。

1 相关研究

1.1 施工场地图像的智能识别

目前施工场地图像的智能识别研究主要基于图像分类、目标检测及语义分割 3 类算法。图像分类算法是最简单的一类计算机视觉任务,其目的是输出待识别图像中主要对象的类别,经典的算法包括 VGG^[4], ResNet^[5]等。Ding 等^[1]结合卷积神经网络(CNN)和长短期记忆网络(LSTM)实现了对施工场地中工人危险动作的识别。Gao 等^[6]基于 VGG 网络,并借助迁移学习技术,实现了混凝土结构图像病害类别的识别。图像分类算法简单易实现,但无法给出图像中所关心对象的具体位置,在计算机视觉领域通常被用作目标检测算法和语义分割算法的主干特征提取网络。

在图像分类算法的基础上,很多学者采用目标检测算法对施工场地的图像进行智能识别。目标检测算法不仅能给出图像中物体的所属类别,而且能框选出其具体位置,经典的算法有 RCNN^[7], YOLO 系列网络^[8]等。Fang 等^[3]采用 Faster RCNN 算法实现了对施工场地中工人是否佩戴安全帽的识别。Lin 等^[2]采用 Faster RCNN 算法对施工场地中机械设备的异常施工活动识别进行了相关研究。相较于图像分类算法,目标检测算法虽然能实现目标对象的定位,但是无法给出其精确的边界信息,在实际的工程应用中仍有局限性。因此,为了对施工场地图像进行更深层次的视觉理解,应用语义分割算法是有前景的研究方向。

1.2 语义分割算法及其工程应用

语义分割算法的目的是对输入图像进行像素级分类,即输出图像中每一个像素所属的类别,能给出拟识别对象的精确边界信息。早期的语义分割算法主要是基于图像数据的统计特征,如阈值算法、边缘检测算法、k-means 聚类算法等^[9]。近年来随着深度学习技术的发展,基于 CNN 的 FCN^[10]网

络,基于编码-解码网络结构的系列网络(U-Net^[11]、V-Net^[12]、SegNet^[13]等)相继被提出,大幅提高了图像语义分割的效果。其中,Chen 等^[14]提出的 DeepLabv3+ 算法被认为是语义分割算法的新高峰,在开源的数据集上达到了十分优越的效果。

目前语义分割算法在土木工程领域中的应用主要集中在结构健康检测,包括对路面缺陷^[15-17]的检测,钢结构^[18]和混凝土结构缺陷^[19-22]的检测,盾构隧道渗漏水^[23-25]的检测,以及输水管道的损伤检测^[26-27]等。除此之外,Golparvar-Fard 等^[28]采用语义分割算法对高速公路道路标识、信号灯等进行识别以评估其状态。Zamani 等^[29]采用 DeepLabv3+ 算法对土体的地质剖面图像进行识别,实现了对土层边界以及土体类型的准确识别。

虽然已有大量研究涉及语义分割算法的工程应用,然而针对施工场地图像的语义分割研究鲜见于文献中。施工场地环境十分复杂,遍布各类对象,且对象之间存在复杂的遮挡关系,采用语义分割算法对其进行精准分割具有很大难度。除此之外,目前工程领域尚缺乏与基坑施工场景相关的采用语义分割级别标注的公开图像数据集,很大程度上阻碍了基坑施工场地图像语义分割模型的开发。

2 基坑施工场地图像数据集

为了训练基于深度学习的基坑施工场地图像语义分割模型,首要任务是建立足够规模的采用语义分割级别标注的图像数据集。笔者采用手机拍摄、基于 BIM 的计算机软件合成两种方式采集并标注大量基坑施工场地图像,用于后续语义分割模型的开发。

2.1 图像采集

首先基于宁波市某超大基坑工程通过手机拍摄的方式采集 150 张实拍图像,图像的像素范围为 2 268~4 608,均保存为 Joint Photographic Experts Group (JPEG)格式。为保证图像数据集的多样性,实拍图像从不同角度、不同距离,在不同的天气及环境光线条件下采集。考虑到实际基坑工程中视频监控设备大都安装在塔吊等高处进行远距离监控,且在大面积施工场地中近距离拍摄图像费时费力,在采集图像时均从施工场地围栏外进行远距离拍摄,以保证实拍图像能真实地反映实际的基坑施工场地监控场景。

由于实拍图像缺乏对基坑施工场地各类对象的近景信息,拍摄角度也较为单一,为补充近景信息及进一步扩充数据集,提出了一套基于 BIM 生成施工场地效果图的流程,如图 1 所示。效果图生成

过程中最为关键的是要尽可能真实地模拟基坑施工场景,一方面需要模拟基坑周边道路、开挖区域、围护结构等静态场景要素,另一方面还需要模拟工人、车辆、机械等在施工过程中不断变化的移动场景要素。为此,首先以实际基坑工程为原型采用 BIM 建模软件 Autodesk Revit 建立素三维 BIM 模型,该模型主要包括整个基坑开挖区域、围护结构、塔吊等固定对象,用于模拟基坑施工静态场景。在此基础上,通过 Lumion 软件对三维 BIM 模型进行场景增强,并借助于 Lumion 的资源库在模型中灵活设置工人、车辆和机械等移动场景要素。为进一步增强场景的真实感,对每个构件的材质进行精细设置,将土层表面、格构柱表面、混凝土表面和承台表

面的材质替换为现场实拍材质图像,使其在视觉效果上与实际场景更加贴合。此外,根据实拍图像中的环境光线,利用 Lumion 中的光线和阴影设置功能调整环境光照效果,使整个场景的光影表现更加自然。最后通过自定义相机高度、视角和焦距,以不同的视点呈现施工场景的全景效果和局部细节,导出不同角度与尺度下的基坑施工场地效果图。需要说明的是,该效果图生成方法是一种通用方法,图像生成十分灵活方便,理论上可生成任意数量的效果图。然而考虑到图像的语义分割级别标注是项极其耗费时间的工作,因此,共收集 180 张具有代表性的基坑施工场地效果图,所有效果图的像素均设置为 $1\,280 \times 720$,保存为 JPEG 格式。



图 1 基于 BIM 的施工场地效果图生成流程

Fig.1 Flowchart for generating synthetic images of construction sites based on BIM

2.2 图像标注

选取 10 类基坑施工场地中常见的对象作为语义分割的目标对象,分别为工人、挖掘机、卡车、混凝土搅拌车、混凝土泵车、吊车、围栏、塔吊、混凝土支撑及背景。采用开源的 LabelMe^[30]标注工具对图像数据集中的目标对象进行像素级的精细标注,形成包含原始图像信息和标签信息的 JavaScript Object Notation (JSON) 文件。标注过程如图 2 所示,使用多边形工具将出现在图像中的目标对象框选并赋予相应的标签信息,若不同目标对象之间存在遮挡,则先标注被遮挡对象,后标注遮挡对象。通过内置函数提取标签信息并将其转换为与原始对象对应的标签图像,用于后续语义分割模型的建立。图 3 展示了所构建的图像数据集中的部分原始图像与对应的标签图像,各类目标对象分别用不同颜色表示。表 1 统计了图像数据集中各目标对象的

实例数和像素数。从表中可以看出实拍图中卡车、混凝土搅拌车、混凝土泵车和吊车对象的实例数相对较少,因此,在效果图中对其进行了一定的补充以保证语义分割模型的识别效果。另外,考虑到工人对象在像素点占比上显著小于其他目标对象,可能会导致模型难以识别,因此,在效果图中对其进行了大量的补充。

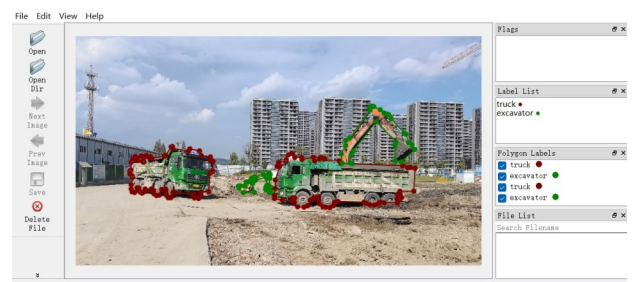


图 2 图像标注

Fig.2 Image annotation

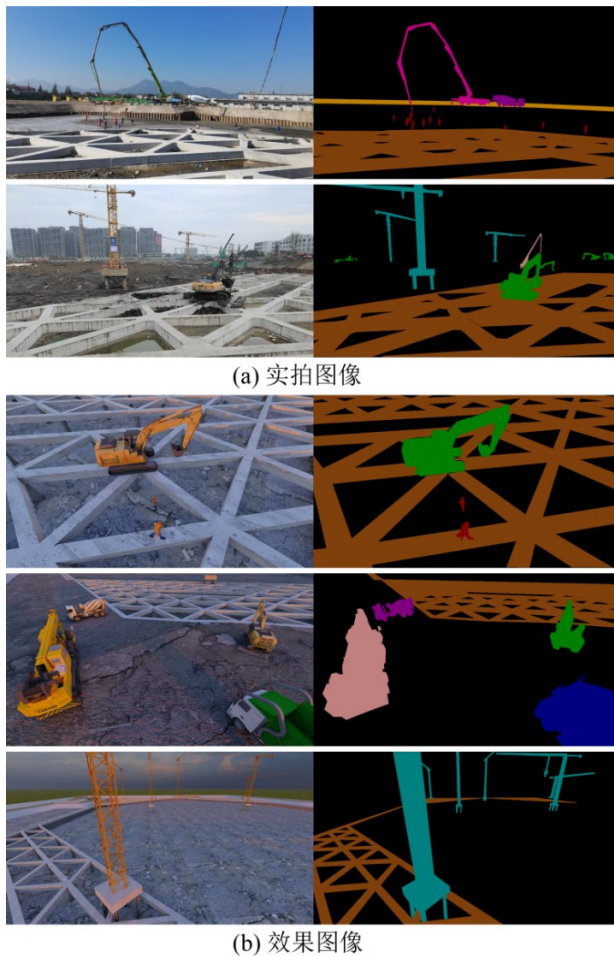


图3 部分原始图像与对应的标签图像

Fig.3 Sample images and corresponding labeled images

表1 图像数据集统计

Table 1 Statistics of the image dataset

对象类别	实拍图		效果图	
	实例数	像素数	实例数	像素数
工人	94	421 039	167	417 804
挖掘机	345	18 857 317	65	2 466 777
卡车	23	4 439 030	73	3 698 149
混凝土搅拌车	20	1 263 911	44	994 162
混凝土泵车	9	478 586	41	1 878 417
吊车	20	333 885	52	1 953 813
围栏	156	34 306 615	104	5 732 909
塔吊	304	49 983 248	106	1 625 247
混凝土支撑	196	319 847 033	182	9 049 098

3 深度语义分割算法

DeepLabv3+算法^[14]是一种语义分割算法,其网络结构采用编码器-解码器形式,相对于其他语义分割算法,具有更优越的性能和鲁棒性。因此,以DeepLabv3+算法为基础实现对基坑施工场地图像的语义分割。图4展示了DeepLabv3+网络的结构,主要由3部分构成:主干特征提取网络,用于提取输入图像的高级特征;空洞卷积池化金字塔

(atrous spatial pyramid pooling, ASPP),用于提取不同尺度的图像语义信息;解码器结构,用于提取目标对象的低级特征,并恢复编码过程中丢失的边界信息,从而提升目标对象分割的细节。

3.1 主干特征提取网络

主干特征提取网络一般选择经典的卷积神经网络结构,如ResNet^[5]、MobileNet^[31]、Xception^[32]等。在进行特征提取前,首先通过预处理将输入图像的大小统一调整为 $1\ 280 \times 720 \times 3$,分别代表图像的宽、高和通道数。为了避免图像尺寸变化带来失真问题,在尺寸归一化的过程中首先保持原图像的长宽比不变,对图像进行等比缩小,之后再对剩余区域进行纯色填充。处理后的图像输入主干特征提取网络后输出一个低级特征层和一个高级特征层。其中,高级特征层作为下一步ASPP结构的输入,低级特征层被直接输出到解码器中。

3.2 ASPP

ASPP结构首先通过不同空洞率的空洞卷积层对输入的高级特征层进行不同尺度的采样,形成融合多尺度语义信息的加强特征层。其中,空洞卷积通过在卷积核中引入空洞机制,相比于传统卷积能够在不改变输出特征图尺寸的前提下有效增大模型的感受野,减少DeepLabv3+网络在特征提取过程中的目标信息丢失。该ASPP结构中包含3个并行设置空洞卷积层,其空洞率分别为6、12和18。此外,ASPP还包含一个全局池化和上采样的分支用于提取加强特征。将所有加强特征层融合并通过一个 1×1 卷积层降维后得到最终的编码器提取的特征层,输出到解码器中。

3.3 解码器结构

在解码器中,主干特征提取网络输出的低级特征层首先经过 1×1 卷积层,然后与通过4倍上采样放大的特征层融合得到新的特征层。新特征层在经过一个 3×3 卷积层后,通过4倍上采样输出得到语义分割结果。DeepLabv3+算法采用高效的编码-解码结构,在编码过程中没有进行特征膨胀,有效地提高了计算速度;在解码过程中通过融合特征与上采样操作,逐渐恢复目标对象的精细边界信息,进一步提升了图像的分割精度。

3.4 损失函数及评价指标

损失函数用于计算在模型训练过程中预测的语义分割结果与实际标注结果之间的差别,通过反向传播算法优化更新模型参数。已有研究^[33]表明,损失函数对于语义分割模型效果的影响十分显著。考虑到构建的图像数据集中不仅各类目标对象实例数差异较大,且存在工人等细小的目标对象,可

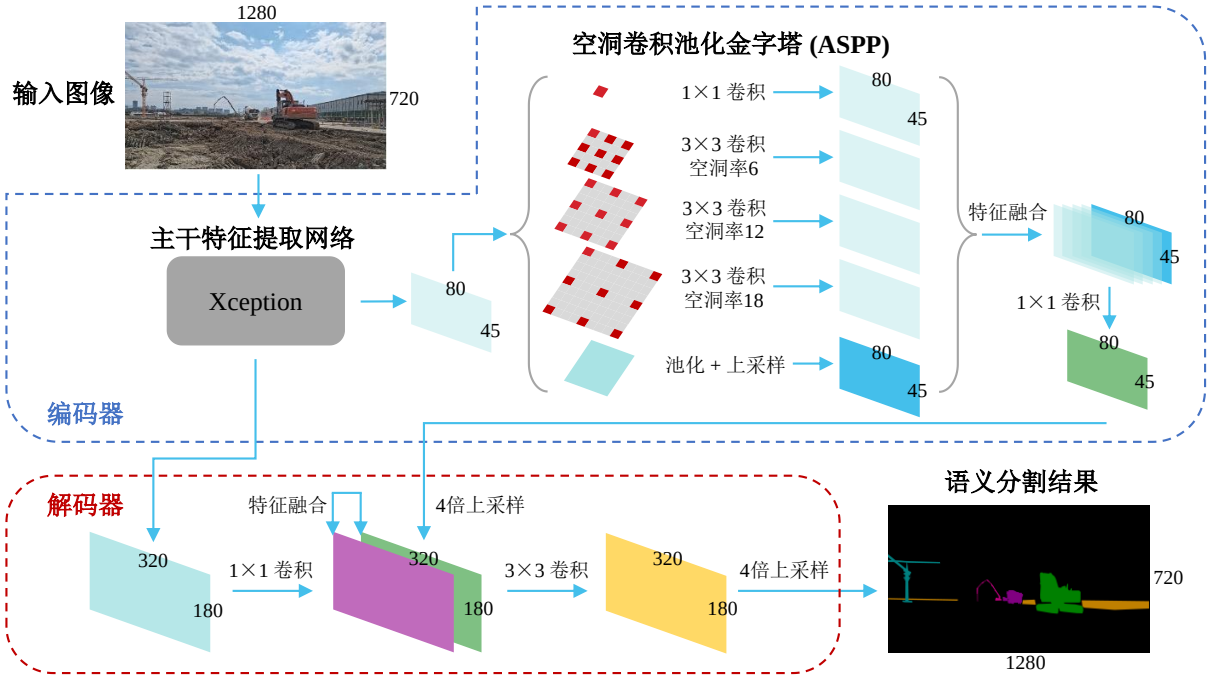


图4 DeepLabv3+网络结构

Fig.4 Architecture of DeepLabv3+ network

能会显著降低模型的预测效果。为了提升语义分割模型的性能,选取 Cross Entropy (CE) loss、weighted Cross Entropy (wCE) loss、Focal loss、Dice loss 及其组合,探究用于模型训练的最佳损失函数形式。

CE loss 的计算式为

$$\ell_{\text{CE}}(g_{ij}, \hat{P}_{ij}) = - \sum_k \delta(g_{ij} = k) \ln \hat{p}_{ij}^{(k)} \quad (1)$$

式中: g_{ij} 为像素点 (i, j) 的真实类别; $\hat{P}_{ij}$ 为预测的各类别概率, $\hat{P}_{ij} = (\hat{p}_{ij}^{(1)}, \hat{p}_{ij}^{(2)}, \dots, \hat{p}_{ij}^{(10)})^\top$; k 为各类别的编号, $k = 1, 2, \dots, 10$; δ 为 Kronecker 符号。

wCE loss 在 CE loss 的基础上引入权重,以缓解数据集集中目标对象样本不均引起的训练困难问题,其计算式为:

$$\ell_{\text{wCE}}(g_{ij}, \hat{P}_{ij}) = - \sum_k \delta(g_{ij} = k) \beta_k \ln \hat{p}_{ij}^{(k)} \quad (2)$$

式中: β_k 为分配给各个类别的权重,通常取各类别样本占比的倒数

$$\beta_k = \frac{1}{r_k}, r_k = \frac{\text{类别}k\text{的像素数}}{\text{总像素数}} \quad (3)$$

Focal loss 在 CE loss 的基础上引入调节因子 γ , 其计算式为

$$\ell_{\text{Focal}}(g_{ij}, \hat{P}_{ij}) = - \sum_k \delta(g_{ij} = k) (1 - \hat{p}_{ij}^{(k)})^\gamma \ln \hat{p}_{ij}^{(k)} \quad (4)$$

式中通过设定调节因子 γ 的值改变各类别预测结果在损失函数中的占比,以提升模型在较难识别对象中的性能。

由以上在像素点 (i, j) 计算的损失值,模型在训

练中的总损失为

$$L(g, \hat{P}) = \frac{1}{1280 \times 720} \sum_{i=1}^{1280} \sum_{j=1}^{720} \ell(g_{ij}, \hat{P}_{ij}) \quad (5)$$

除此之外,还引入了基于区域的损失函数 Dice loss,用于提升模型对工人等细小对象的识别效果。其计算式为

$$L_{\text{Dice}} = 1 - \frac{1}{10} \sum_k D_c^k, D_c = \frac{2T_p}{2T_p + F_p + F_n} \quad (6)$$

式中: D_c 为 Dice 系数,其值可由混淆矩阵中的 T_p (True Positive)、 F_p (False Positive)、 F_n (False Negative) 数量计算得到。

为了评价模型对施工场地图像的分割效果,选取平均交并比 (mean Intersection over Union, $\bar{I}_{oU}$) 作为评价指标。交并比 I_{oU} 是指预测值与真实值的交集与并集之比, $\bar{I}_{oU}$ 为每一类交并比的平均值,计算公式如下:

$$\bar{I}_{oU} = \frac{1}{10} \sum_k I_{oU}^k, I_{oU} = \frac{T_p}{T_p + F_p + F_n} \quad (7)$$

4 基坑场地图像语义分割实验及结果

首先通过实验探究不同主干特征提取网络 (MobileNetV2、ResNet-101、Xception) 与不同损失函数 (CE loss、wCE loss、Focal loss、Dice loss+CE loss、Dice loss+wCE loss、Dice loss+Focal loss) 的组合对 DeepLabv3+ 语义分割模型效果的影响,以得到最佳的基坑施工场地图像语义分割模型。将 150 张实拍图按 3:1:1 划分训练验证测试集,180 张

效果图全部划入训练集,最终得到9:1:1的训练验证测试集比例,图像数量分别为270张、30张、30张。除此之外,通过调整训练集中实拍图与效果图的比例,对比验证效果图对基坑施工场地图像语义分割模型的增强效果。

4.1 实验环境及超参数设置

实验计算机操作系统为Windows Server 2016,显卡为NVIDIA Tesla P40,显存24 GB。DeepLabv3+算法通过基于Python计算机语言的开源深度学习库Pytorch实现。

在模型训练前,首先对图像数据集进行随机缩放、翻转、旋转、高斯模糊、色域变换等一系列图像增强操作,扩充训练集大小,有效防止模型的过拟合。为提升模型的训练效果,在训练过程中采用了迁移学习方法。首先将在公开数据集ImageNet^[34]上预训练的网络参数应用到主干特征提取网络上,然后将整个DeepLabv3+网络在PASCAL VOC 2012^[35]数据集上预训练的网络参数作为初始参数,最后在建立的施工场地图像数据集上训练。在模型训练时,所有输入图像像素均预先调整为 $1\,280 \times 720$,对于原始宽高比非16:9的图像,采用灰条填充的方式调整比例。为提高计算效率并节约计算资源,训练的批大小设置为2。初始学习率设置为 5×10^{-4} ,并在训练过程中不断衰减,最小学习率设置为 5×10^{-6} 。采用Adam优化器更新网络参数,网络训练代数设置为200,以保证网络在训练过程中能够达到收敛。

4.2 训练结果

训练过程中的损失与 $\bar{I}_{oU}$ 曲线如图5所示。由图5可以看出,训练集和验证集的损失在前20个训练世代迅速降低。在60个训练世代后,损失与 $\bar{I}_{oU}$

值已基本稳定,表明网络性能已在训练过程中收敛。此外,训练集与验证集的损失值相差很小,表明网络在训练过程中没有出现拟合的现象。

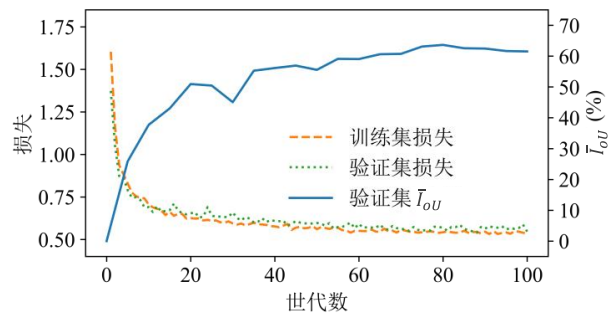


图5 训练过程中的损失与 $\bar{I}_{oU}$ 曲线

Fig.5 Loss and $\bar{I}_{oU}$ curves during the training process

4.3 测试结果

表2展示了不同主干特征提取网络与损失函数组合下DeepLabv3+网络模型在测试集上的 $\bar{I}_{oU}$ 结果。从表2可以看出,采用Xception作为主干网络和Dice loss+CE loss作为损失函数的网络模型在测试集上达到了最好的效果, $\bar{I}_{oU}$ 值为68.24%。主干网络的效果在相同的损失函数设置下优于MobileNetV2和ResNet-101,主要是因为其采用的深度可分离卷积结构^[32]具有显著的优越性。此外,可以发现与Dice loss相组合的损失函数能够显著提高网络模型的性能,主要是因为Dice loss根据目标对象的识别区域进行损失的计算,与像素点占比的大小无关,能显著提高模型对于细小物体的识别能力。值得一提的是,在CE loss中引入权重后显著降低了模型的效果,其主要原因是对于样本高度不平衡的数据集各目标类别的权重差距过大,造成模型在训练过程中难以收敛。

表2 不同配置下DeepLabv3+网络的 $\bar{I}_{oU}$ 值

Table 2 $\bar{I}_{oU}$ values of DeepLabv3+ network with different setups

主干特征提取网络	CE	wCE	Focal	Dice+CE	Dice+wCE	Dice+Focal
MobileNetV2	56.06%	45.65%	57.05%	64.27%	58.59%	63.98%
ResNet-101	57.26%	44.44%	60.17%	66.41%	59.28%	67.70%
Xception	61.59%	46.80%	61.73%	68.24%	59.95%	67.65%

图6为最优DeepLabv3+模型的各目标对象 I_{oU} 值。一般情况下,语义分割任务中拟识别的目标对象 I_{oU} 值超过50%,即可认为模型对其有基本的识别能力。由图6可以看出,训练的模型对于挖掘机、混凝土搅拌车、混凝土泵车、围栏、塔吊和混凝土支撑均有较好的识别效果,但是对工人和吊车的识别效果稍差,尤其是吊车对象, I_{oU} 值仅为32%。其主要原因是工人、吊车对象不仅像素点占比小,而且形状比较细长,使得模型难以准确识别。现有分割模

型主要针对常规形态目标设计,在面对吊车等具有显著线性特征及复杂空间延伸结构的对象时,往往容易出现结构断裂或轮廓不连续的问题。未来研究应进一步优化算法架构,着重提升对线性及细长特征的捕捉与重建能力。

表3详细比较了不同配置下DeepLabv3+网络在参数规模和计算成本方面的表现。结果显示,以ResNet-101和Xception为主干网络的DeepLabv3+参数量约为MobileNetV2版本的10倍,因此,训练

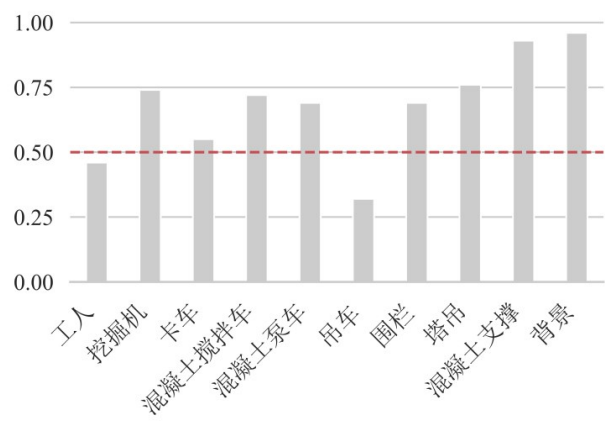


图 6 各目标对象的 I_{ou} 值

Fig.6 I_{ou} values of each target object

时间显著增加。在训练过程中,MobileNetV2 主干网络所占用的内存仅为 Xception 的一半,显著降低了计算资源需求。在推理速度方面,基于 MobileNetV2 的 DeepLabv3+ 达到了 39.84 FPS (Frames Per Second: 每秒帧数),满足了基坑施工场景图像实时语义分割的要求(推理速度 > 30 FPS)。尽管 Xception 主干网络因较高的计算成本其推理速度仅为 15.85 FPS,但通过提升计算平台性能,仍可满足现场实时监控的需求。

为研究基于 BIM 生成的效果图对基坑施工场

地图语义分割模型训练的增强效果,调整训练集中实拍图与效果图的比例来比较模型的性能。表 4 展示了不同训练集下模型的 $\bar{I}_{ou}$ 值以及各目标对象的 I_{ou} 值,其中 DeepLabv3+ 模型统一采用 Xception 作为主干网络,Dice loss+CE loss 作为损失函数。从表中可以看出采用效果图辅助模型训练对于模型预测性能的提升具有显著的效果,与不采用效果图训练相比测试集上的 $\bar{I}_{ou}$ 值提高了 26.27%(从 41.97% 到 68.24%)。几类细小的或是在数据集中数量很少的难以识别的目标对象,例如工人、混凝土搅拌车、混凝土泵车和吊车等,在未混合效果图的情况下模型不具备对其的识别能力。然而,由于在效果图中针对性地对这几类难以识别目标对象的实例数进行了补充,并通过特定的近景图像补充了目标对象的细节信息,随着效果图混合比例的增大,模型逐渐具备了对这几类难以识别目标对象的识别能力。在混合 40% 数量的总效果图时模型已能识别这几类难以识别的目标对象,且随着效果图混合比例的增大,其 I_{ou} 值进一步提高。除了这几类难以识别的目标对象外,效果图的引入对于模型对其他目标对象的识别效果也均有提升作用。在效果图数量达到 40% 后,后续继续增加效果图的数量对模型效果的提升作用开始减弱,表明所收集的效果图数量基本足够用于模型的训练。

表 3 不同配置下 DeepLabv3+ 网络的参数与计算成本比较

Table 3 Comparison of parameters and computational cost of DeepLabv3+ network with different setups

主干特征提取网络	参数/ 10^6	GFLOPs	显存占用/MB	训练时间/s	验证时间/s	推理速度/FPS
MobileNetV2	5.815	186.095	7 072	36.30	6.42	39.84
ResNet-101	59.344	625.679	11 256	55.33	7.11	19.52
Xception	54.711	586.784	14 766	71.75	7.15	16.66

注:实验环境为 Ubuntu 18.04 操作系统,NVIDIA RTX 3090 24GB GPU、Intel Xeon Platinum 8358P CPU。GFLOPs (Giga Floating-point Operations Per Second): 每秒十亿次的浮点运算数。FPS(Frames Per Second): 每秒帧数。

表 4 不同训练集下 DeepLabv3+ 网络的 $\bar{I}_{ou}$ 值

Table 4 $\bar{I}_{ou}$ values of DeepLabv3+ network with different training sets

训练集		验证集		测试集		$\bar{I}_{ou}$	工人	挖掘机	卡车	混凝土搅拌车	混凝土泵车	吊车	围栏	塔吊	混凝土支撑	背景
实拍图比例	效果图比例	实拍图比例	实拍图比例	效果图比例	效果图比例											
60%	0%	20%	20%	20%	20%	41.97%	0	66%	37%	0	0	0	70%	73%	81%	93%
60%	20%	20%	20%	20%	20%	47.92%	27%	69%	50%	0	0	0	71%	74%	92%	96%
60%	40%	20%	20%	20%	20%	62.74%	39%	69%	54%	64%	46%	21%	70%	76%	93%	96%
60%	60%	20%	20%	20%	20%	65.10%	44%	74%	55%	62%	57%	25%	71%	76%	92%	96%
60%	80%	20%	20%	20%	20%	66.77%	47%	73%	54%	69%	57%	36%	65%	77%	93%	96%
60%	100%	20%	20%	20%	20%	68.24%	46%	74%	55%	72%	69%	32%	69%	76%	93%	96%

图 7 为不同训练集下 DeepLabv3+ 网络对测试集图像的分割效果。从图中可以看出,对于图 7(a)和(b)这类基本不包含细小物体的近景图像,训练的网络模型能够有效地分割出各类不同的目标对

象,并且能够正确识别图像中对象之间的遮挡关系,如模型能正确识别图 7(b)中遮挡了卡车的挖掘机对象。随着训练集中效果图比例的增加,模型对于各类目标对象轮廓的分割越来越准确。对于图 7

(c)和(d)这类包含工人、混凝土泵车臂、吊车臂等细小对象的远景图像,训练的网络模型分割效果显著降低,仅在实拍图上训练的模型甚至完全无法识别出部分目标对象。随着训练集中效果图比例的增加,模型对细小对象的分割效果显著提升,主要原因是在效果图中对工人等细小对象的近景信息进

行了大量补充。上述结果表明,在基于150张实拍图和180张效果图构建的基坑施工场地图像数据集上训练的语义分割模型已基本实现在近景、远景、不同的天气及环境光线条件下对工人、机械设备、结构等对象的像素级分割。

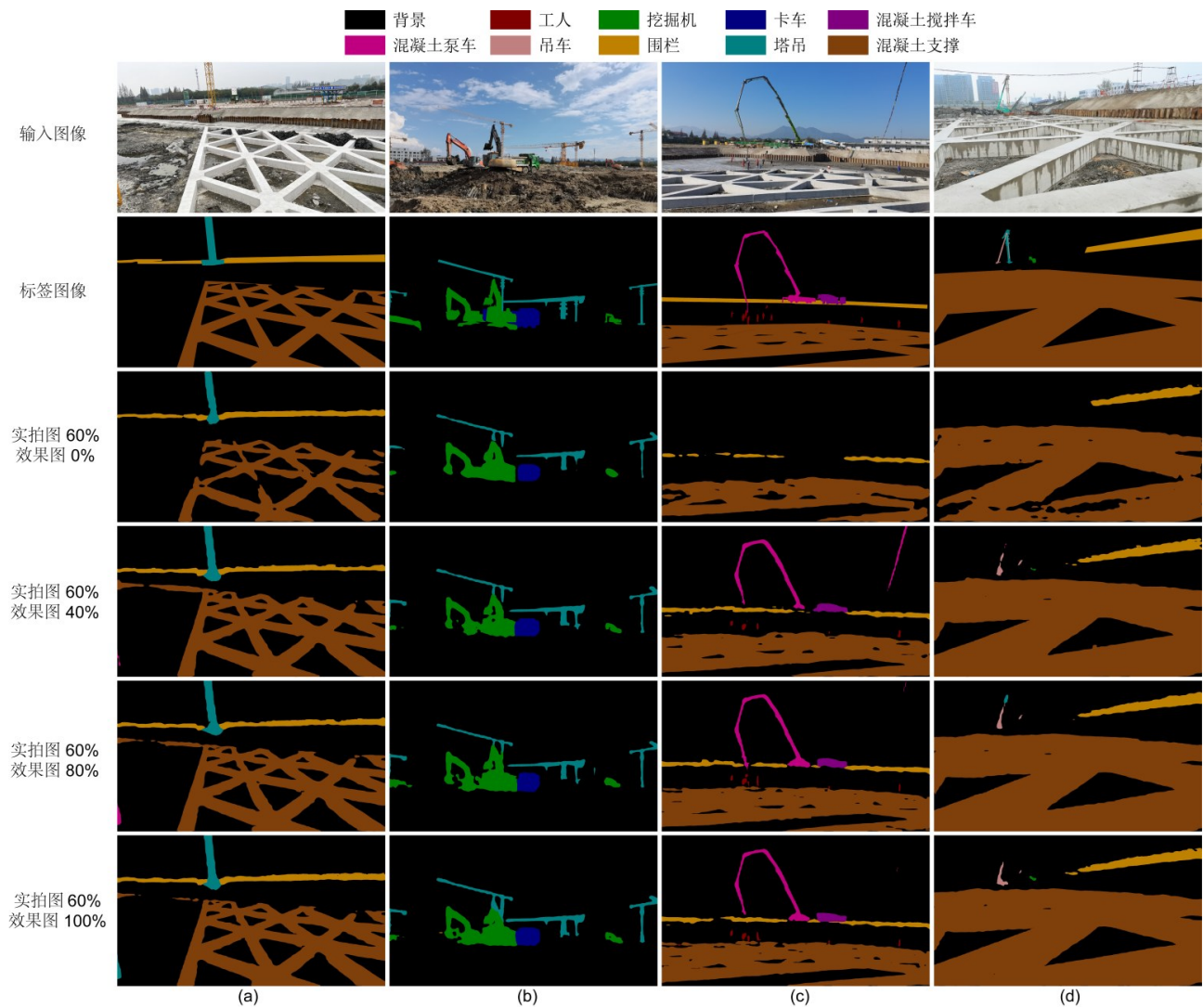


图7 不同训练集下DeepLabv3+网络的图像分割结果

Fig.7 Segmentation results of DeepLabv3+ network with different training sets

4.4 效果图优缺点讨论

实验结果表明,在训练基坑施工场地图像语义分割模型时,效果图的引入显著提升了模型性能,与仅使用实拍图进行训练相比,加入效果图后模型的 $\bar{I}_{ou}$ 值提高了26.27%。基于BIM生成的效果图能够有效弥补实拍图中个别对象实例数或像素占比不足的问题,从而增强模型对这些对象的识别能力。提出的基于BIM的效果图生成方法能够生成多角度、多尺度的基坑施工场地效果图,这些图像在视觉效果上高度逼真,能够模拟复杂的基坑施工场景。此外,该方法支持对施工场景中各类移动对象的自定义设置,从而满足各种特定需求。更为重

要的是,该方法不仅可以生成当前施工阶段的场景,还能够模拟实际现场尚未发生的施工场景。效果图生成过程灵活高效、成本低廉、操作便捷,为基坑施工场地图像的获取提供了一种创新且实用的解决方案。

尽管效果图在提升模型性能方面表现出显著效果,但其仍无法完全替代实拍图。主要原因是效果图在信息完整性和细节表现方面存在不足,而这些缺陷源于三维建模过程的局限性。其中主要的问题是基于一BIM的三维模型难以准确还原复杂施工现场的环境细节,导致效果图中缺乏真实场景中的背景信息。此外,目标对象的建模受制于Revit

和 Lumion 软件中预设元素的限制。尽管工人与施工机械(如挖掘机、卡车和吊车等)可以通过嵌入式元素进行建模,但这些虚拟模型与真实图像中的对象仍存在差异。一方面体现在几何形状与视觉风格上,例如,Lumion 虽可模拟不同天气条件(如晴天/阴天)及其对应的光照环境,但其生成的图像在光线质感、阴影层次和色彩饱和度等方面仍与真实场景存在可见差别,如图 8(a)、(b)所示。另一方面,差异还体现在工人与机械设备的动作状态上。例如,真实施工场景中工人的典型行为既包括静态休息状态,也包括动态作业状态(如图 8(c)所示);而挖掘机在实际作业中则可能处于闲置、移动、旋转、挖掘或卸载等多种动作模式^[36],这些动态行为在现有效果图中尚未得到充分还原。未来研究可以聚焦于开发与施工现场目标对象高度匹配的定制化元素,并将其集成到三维 BIM 模型建模过程中,有助于生成更接近真实场景的效果图,从而提升效果图在实际应用中的有效性。

5 基于语义分割的基坑场地对象识别

首先,基于 OpenCV 库采用图像形态学方法(开运算、闭运算等)对语义分割结果中的各目标对象像素进行处理,去除孤立像素和孔洞,以增强像素区域的完整性和连贯性。然后,通过筛选特定面积



图 8 效果图与实拍图对比

Fig.8 Comparison between synthetic and real images

以下的区域进一步去除小面积的误检区域。最后,利用 scikit-image 库对语义分割结果进行连接组件分析,将相互连通的像素区域进行聚合,并统计图像中各像素区域(即有效连接组件)的类型与数量。图 9 分别以支撑、机械设备为例,展示了基于语义分割结果的基坑场地对象识别流程。

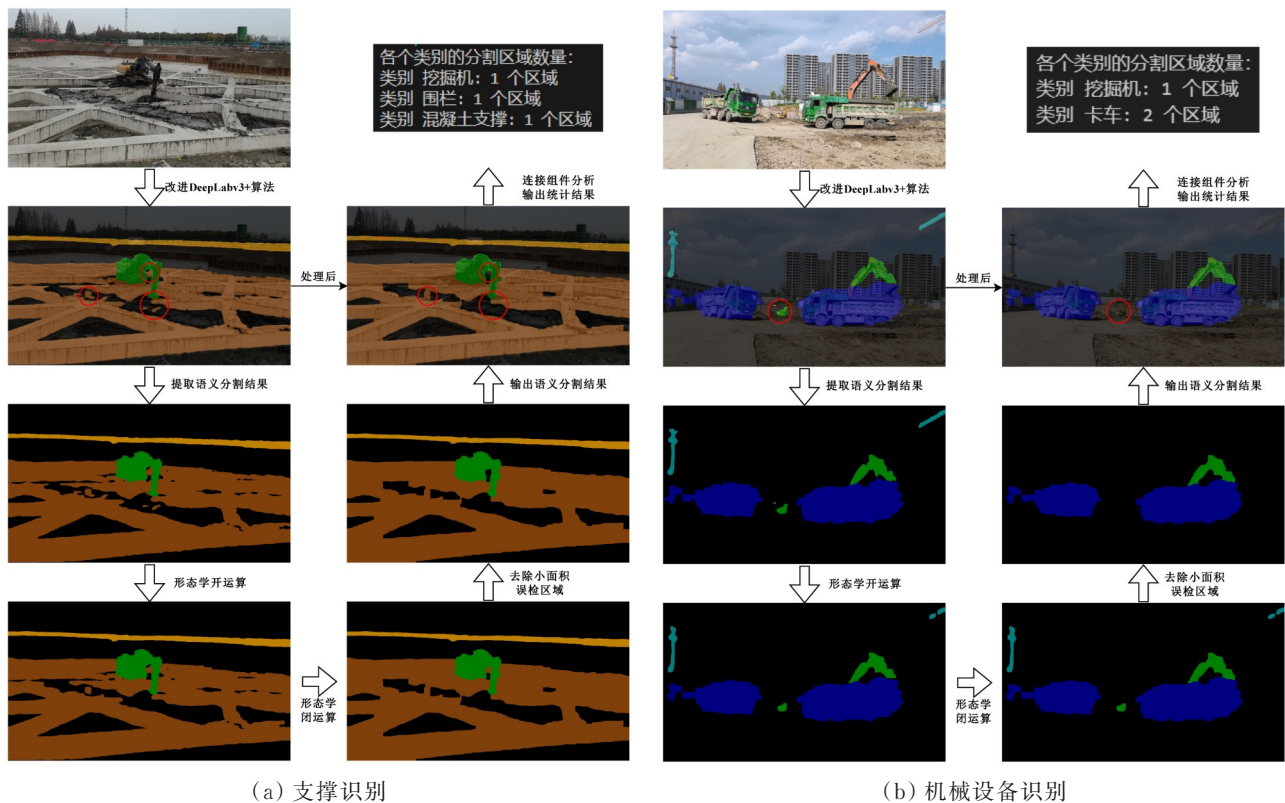


图 9 基于语义分割结果的基坑场地目标对象识别

Fig.9 Recognition of target objects at excavation sites based on semantic segmentation results

从图9中可以看出,原始语义分割结果中存在很多小的误检区域以及内部孔洞。经过图像形态学运算以及筛选后,误检区域与孔洞显著减少,有效避免了其对后续像素区域统计结果的干扰。从图中输出的统计结果可以看出,所提出的识别流程能够准确输出基坑施工场地图像中的支撑、机械设备等目标对象的类型与数量。

6 结论

以超大规模基坑工程为应用场景,构建了一套BIM增强的语义分割图像数据集并开发了基于DeepLabv3+的语义分割算法,实现了对复杂环境条件下基坑施工场地图像中工人、机械设备、结构等对象的像素级分割,并提出了基于语义分割结果的目标对象类别与数量识别方法。主要结论如下:

1)采用Xception作为主干网络和Dice loss+CE loss作为损失函数的DeepLabv3+模型分割效果最佳, $\bar{I}_{ou}$ 值达到68.24%。采用与Dice loss相结合的损失函数能够显著提升模型对于施工场地图像中细小对象的分割效果。

2)采用基于BIM生成的效果图训练模型对模型分割效果具有显著的提升作用,与仅采用实拍图相比模型的 $\bar{I}_{ou}$ 值提高了26.27%。所提出的效果图生成方法简单、灵活、高效,可用于增强类似工程领域的图像数据集。

3)基于语义分割结果,采用图像形态学运算与连接组件分析的目标对象识别方法能够准确输出基坑施工场地图像中的支撑、机械设备等目标对象的类型与数量。

4)相较于图像分类与目标检测算法,开发的语义分割算法能分割出复杂场景下施工场地各目标对象的边缘细节,在对象类型数量识别、目标物体间距离识别等方向具有实际的应用前景,可为施工场景智能识别向着深入、精细化方向发展提供技术支持。

参考文献

- [1] DING L Y, FANG W L, LUO H B, et al. A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory [J]. Automation in Construction, 2018, 86: 118-124.
- [2] LIN Z H, CHEN A Y, HSIEH S H. Temporal image analytics for abnormal construction activity identification [J]. Automation in Construction, 2021, 124: 103572.
- [3] FANG Q, LI H, LUO X C, et al. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos [J]. Automation in Construction, 2018, 85: 1-9.
- [4] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. 2014: arXiv: 1409.1556. <https://arxiv.org/abs/1409.1556>
- [5] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas, NV, USA. IEEE, 2016: 770-778.
- [6] GAO Y Q, KONG B Y, MOSALAM K M. Deep leaf-bootstrapping generative adversarial network for structural image data augmentation [J]. Computer-Aided Civil and Infrastructure Engineering, 2019, 34(9): 755-773.
- [7] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. June 23-28, 2014, Columbus, OH, USA. IEEE, 2014: 580-587.
- [8] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas, NV, USA. IEEE, 2016: 779-788.
- [9] LIU X L, DENG Z D, YANG Y H. Recent progress in semantic image segmentation [J]. Artificial Intelligence Review, 2019, 52(2): 1089-1106.
- [10] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 7-12, 2015, Boston, MA, USA. IEEE, 2015: 3431-3440.
- [11] RÖNNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation [C]//Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Cham: Springer, 2015: 234-241.
- [12] MILLETARI F, NAVAB N, AHMADI S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation [C]//2016 Fourth International Conference on 3D Vision (3DV). October 25-28, 2016, Stanford, CA, USA. IEEE, 2016: 565-571.
- [13] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.
- [14] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [C]//Computer Vision -

- ECCV 2018. Cham: Springer, 2018: 833-851.
- [15] ZHONG J T, ZHU J Q, JU H Y, et al. Multi-scale feature fusion network for pixel-level pavement distress detection [J]. *Automation in Construction*, 2022, 141: 104436.
- [16] ZHANG A A, WANG K C P, LIU Y, et al. Intelligent pixel-level detection of multiple distresses and surface design features on asphalt pavements [J]. *Computer-Aided Civil and Infrastructure Engineering*, 2022, 37 (13): 1654-1673.
- [17] CHEN J, HE Y. A novel U-shaped encoder-decoder network with attention mechanism for detection and evaluation of road cracks at pixel level [J]. *Computer-Aided Civil and Infrastructure Engineering*, 2022, 37 (13): 1721-1736.
- [18] PAN Y, ZHANG L. Dual attention deep learning network for automatic steel surface defect segmentation [J]. *Computer-Aided Civil and Infrastructure Engineering*, 2022, 37(11): 1468-1487.
- [19] ÇELİK F, KÖNIG M. A sigmoid-optimized encoder-decoder network for crack segmentation with copy-edit-paste transfer learning [J]. *Computer-Aided Civil and Infrastructure Engineering*, 2022, 37(14): 1875-1890.
- [20] CHAIYASARN K, BUATIK A, MOHAMAD H, et al. Integrated pixel-level CNN-FCN crack detection via photogrammetric 3D texture mapping of concrete structures [J]. *Automation in Construction*, 2022, 140: 104388.
- [21] CHU H H, WANG W, DENG L. Tiny-Crack-Net: A multiscale feature fusion network with attention mechanisms for segmentation of tiny cracks [J]. *Computer-Aided Civil and Infrastructure Engineering*, 2022, 37(14): 1914-1931.
- [22] LI Y T, BAO T F, HUANG X J, et al. Underwater crack pixel-wise identification and quantification for dams via lightweight semantic segmentation and transfer learning [J]. *Automation in Construction*, 2022, 144: 104600.
- [23] HUANG H W, LI Q T, ZHANG D M. Deep learning based image recognition for crack and leakage defects of metro shield tunnel [J]. *Tunnelling and Underground Space Technology*, 2018, 77: 166-176.
- [24] XUE Y D, CAI X Y, SHADABFAR M, et al. Deep learning-based automatic recognition of water leakage area in shield tunnel lining [J]. *Tunnelling and Underground Space Technology*, 2020, 104: 103524.
- [25] 周中, 张俊杰, 龚琛杰, 等. 基于深度语义分割的隧道渗漏水智能识别[J]. *岩石力学与工程学报*, 2022, 41 (10): 2082-2093.
- ZHOU Z, ZHANG J J, GONG C J, et al. Automatic identification of tunnel leakage based on deep semantic segmentation [J]. *Chinese Journal of Rock Mechanics and Engineering*, 2022, 41(10): 2082-2093. (in Chinese)
- [26] WANG M Z, LUO H, CHENG J C P. Towards an automated condition assessment framework of underground sewer pipes based on closed-circuit television (CCTV) images [J]. *Tunnelling and Underground Space Technology*, 2021, 110: 103840.
- [27] ZHOU Q Q, SITU Z, TENG S, et al. Automatic sewer defect detection and severity quantification based on pixel-level semantic segmentation [J]. *Tunnelling and Underground Space Technology*, 2022, 123: 104403.
- [28] GOLPARVAR-FARD M, BALALI V, DE LA GARZA J M. Segmentation and recognition of highway assets using image-based 3D point clouds and semantic texton forests [J]. *Journal of Computing in Civil Engineering*, 2015, 29: 04014023.
- [29] ZAMANI V, TAGHADDOS H, GHOLIPOUR Y, et al. Deep semantic segmentation for visual scene understanding of soil types [J]. *Automation in Construction*, 2022, 140: 104342.
- [30] RUSSELL B C, TORRALBA A, MURPHY K P, et al. LabelMe: A database and web-based tool for image annotation [J]. *International Journal of Computer Vision*, 2008, 77(1): 157-173.
- [31] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: Inverted residuals and linear bottlenecks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-23, 2018, Salt Lake City, UT, USA. IEEE, 2018: 4510-4520.
- [32] CHOLLET F. Xception: Deep learning with depthwise separable convolutions [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). July 21-26, 2017, Honolulu, HI, USA. IEEE, 2017: 1800-1807.
- [33] CHERN W C, NGUYEN T V, ASARI V K, et al. Impact of loss functions on semantic segmentation in far-field monitoring [J]. *Computer-Aided Civil and Infrastructure Engineering*, 2023, 38(3): 372-390.
- [34] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge [J]. *International Journal of Computer Vision*, 2015, 115(3): 211-252.
- [35] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The pascal visual object classes (VOC) challenge [J]. *International Journal of Computer Vision*, 2010, 88(2): 303-338.
- [36] JUNG S, JEOUNG J, KANG H, et al. 3D convolutional neural network-based one-stage model for real-time action detection in video of construction equipment [J]. *Computer-Aided Civil and Infrastructure Engineering*, 2022, 37(1): 126-142.