E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:P23　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)06-1808-20

论文引用格式:Zhang G Y, Zhang R T, Zhang Y, Wang Q X, Feng J Q and Jiang H X. 2025. Technology of intelligent interpretation of three-dimensional mesh models for complex urban scenes. Journal of Image and Graphics, 30(6):1808-1827(张广运, 张荣庭, 张余, 王麒雄, 冯家齐, 姜鸿翔. 2025. 复杂城市场景三维网格模型智能解译技术综述. 中国图象图形学报, 30(6):1808-1827)[DOI:10.11834/jig.240778]

复杂城市场景三维网格模型智能解译技术综述

张广运[1],张荣庭[1*],张余[2],王麒雄[2],冯家齐[2],姜鸿翔[2]

1. 南京工业大学测绘科学与技术学院,南京　211816; 2. 北京航空航天大学宇航学院,北京　100191

摘　要: 城市3D Mesh模型解译是城市级实景三维建设的重要环节,有助于建筑设施、交通设施等城市设施的数字化和智能化、精细化管理,在城市更新、环境整治和城市仿真等城市行动中发挥积极作用。当前城市3D Mesh模型的语义化、单体化仍主要由人工勾勒实体轮廓,通过实体边界将每一个单独地物从城市3D Mesh模型中切割出来并赋予语义信息,然而城市3D Mesh模型通常是以瓦块的形式表达,在进行跨瓦块切割时容易让模型出现破碎、接缝和割裂等问题。为此,学者们开始研究基于深度神经网络的城市3D Mesh模型智能解译。然而,城市3D Mesh模型的智能解译却面临着巨大挑战,如城市3D Mesh模型不规则/非水密,传统卷积网络难以直接应用;城市3D Mesh模型多尺度特征获取困难等。虽然深度神经网络在城市3D Mesh模型解译方面的应用起步较晚,但该领域的研究依然取得了迅猛发展。因此,本文以城市3D Mesh模型智能解译为主线,系统回顾和总结现有面向城市3D Mesh模型解译的深度神经网络方法,根据城市3D Mesh模型表达方式的不同,将面向城市3D Mesh模型解译的深度神经网络方法分为3类,即面向多视图表示的方法、面向质心点云表示的方法和面向3D Mesh模型元素的方法,对这3类方法进行详细比较,并总结了当前面临的挑战;其次,梳理了城市3D Mesh模型智能解译常用的6个基准数据集,比较了多种方法在这些基准数据集上针对城市3D Mesh模型语义分割任务的性能表现;最后,对城市3D Mesh模型解译未来的发展方向和潜在应用前景进行了深入分析和讨论。

关键词: 数字中国;实景三维;深度学习;场景解译;城市三维网格

Technology of intelligent interpretation of three-dimensional mesh models for complex urban scenes: a survey

Zhang Guangyun[1], Zhang Rongting[1*], Zhang Yu[2], Wang Qixiong[2], Feng Jiaqi[2], Jiang Hongxiang[2]

1. School of Geomatics Science and Technology, Nanjing Tech University, Nanjing　211816, China;

2. School of Astronautics, Beihang University, Beijing　100191, China

Abstract: The 3D real scene forms the spatial foundation and provides a unified spatial positioning framework and analysis basis for digital China. According to content and hierarchy, 3D real scene can be categorized into terrain, city, and component levels. The city-level 3D real scene is mainly composed of 3D mesh models derived from oblique photography, LiDAR (light detection and ranging) point clouds, and texture images, which are semantically processed and integrated with real-time perception data. Urban 3D mesh models are primarily composed of vertices, edges, triangular faces, and texture images. Compared to point clouds, 3D mesh models not only display more detailed information about objects but also allow

收稿日期:2024-12-27;**修回日期**:2025-02-19;**预印本日期**:2025-02-26

*通信作者:张荣庭　zrt@njtech.edu.cn

基金项目:江苏省自然科学基金项目(BK20230338)

Supported by:Natural Science Foundation of Jiangsu Province, China (BK20230338)

for easy control of the level of detail through adjusting the parameters of the 3D mesh model. The focus of 3D real scene is on the digital mapping of production and living spaces, which can assist in the fine-grained management of cities and serve intelligent urban planning and construction. Interpreting 3D mesh models of urban scenes, such as semantic and instance segmentation, is a crucial step in constructing city-level 3D real scene. Currently, the semantic and instance segmentation of urban 3D mesh models primarily involves manually drawing the outline of object and cutting out each individual object from the 3D mesh model using object boundaries, followed by assigning semantic information. However, urban 3D mesh models are typically represented in a tile-based format, and cross-tile cutting can easily lead to issues such as fragmentation, seams, and discontinuities in the model. In recent years, deep learning technology has seen rapid development. Deep neural networks, due to their ability to learn discriminative high-level semantic features from given datasets, have been widely applied to the interpretation of image data and three-dimensional data (such as point clouds and 3D meshes). In addition, with the continuous improvement in the performance of graphics processing units and the expansion of annotated datasets, the accuracy of deep neural networks in interpreting 2D images and 3D data has significantly improved. Despite significant progress in the interpretation of 3D mesh models, most of these studies have focused on small-scale, toy, and simulated 3D mesh models. Research on deep neural networks for interpreting complex urban 3D mesh models is still in its early stages and faces many challenges and difficulties, primarily in the following three aspects. 1) Urban 3D mesh models are often irregular and may contain holes or be nonwatertight, making it difficult for traditional deep neural networks to directly apply to these models in extracting highly discriminative features. 2) The efficiency of extracting multiscale feature is low. Traditional 3D mesh simplification methods (such as quadric error metrics), which are used to generated hierarchy 3D mesh, use greedy strategies that are difficult to parallelize. When processing large-scale urban 3D mesh models, these methods inevitably increase computational burden. 3) Compared to benchmark datasets for images and point clouds, publicly available benchmark datasets for urban 3D mesh models are scarce. The structures of buildings, roads, and vegetation in urban scenes are complex and varied, making the annotation of urban 3D mesh models not only require specialized knowledge but also consume a significant amount of time and human resources. Compared to the intelligent interpretation of images and point clouds, the application of deep neural networks in the interpretation of urban 3D mesh models started later but has still seen rapid development. However, currently, few review articles systematically explore and summarize how different deep neural network architectures achieve the interpretation of urban 3D mesh models. Therefore, this paper aims to systematically review and summarize existing deep neural network methods for interpreting urban 3D mesh models and highlight the open challenges currently faced by researchers, providing a reference for future research. For this purpose, we initially survey the vast literature and categorize the intelligent interpretation methods for urban 3D mesh models into three classes, according to the types of representations used in processing urban 3D mesh models. 1) Methods based on multiview images attempt to project 3D mesh models into 2D images from multiple viewpoints and use well-established 2D image deep learning methods to learn discriminative semantic features from the projected images. Subsequently, the semantic features learned from the projected images are mapped back to the 3D mesh models. 2) Methods based on center-of-gravity (COG) point cloud representation convert each face of the urban 3D mesh model into its COG point, thereby abstracting the entire 3D mesh model into COG point clouds. Subsequently, intelligent interpretation algorithms designed for point clouds are used to process these COG point clouds. Different from traditional point clouds, COG point clouds can inherit rich texture and geometric information from the urban 3D mesh model. 3) Methods based on 3D mesh elements aim to define learnable operations (such as convolution and pooling) directly on the 3D mesh elements (vertices, edges, triangular faces). This approach allows for the direct learning and extraction of rich high-level semantic features from the urban 3D mesh model, thereby avoiding information loss that can result from preprocessing steps such as multiview image projection and centroid point cloud abstraction. Subsequently, we conduct a detailed comparison of the four categories of methods and outline their current challenges. Furthermore, we summarize commonly used datasets for intelligent interpretation of urban 3D mesh models and compare the interpreting performance of different methods on these datasets. Finally, based on the systematic survey and comprehensive performance comparison, we discuss some promising future research directions from aspects such as dataset creation, 3D large model construction, and application scenarios.

Key words: digital China; three-dimensional real scene; deep learning; scene interpretation; urban three-dimensional mesh

0 引 言

实景三维中国是数字中国的空间基底和统一的空间定位框架与分析基础，在我国范围内作为新型基础设施得到大力推广（朱庆 等，2022）。按照表达内容和层级，实景三维可划分为地形级、城市级和部件级。城市级实景三维主要由融合了实时感知数据的语义化倾斜摄影三维网格（3D Mesh）模型、激光点云和纹理影像等数据构成，其重点是对生产和生活空间的数字映射，能够助力于城市精细化管理、服务于城市智能规划建设（张广运 等，2021；朱庆 等，2022）。

对城市场景3D Mesh模型进行解译，如语义分割和实例分割，是建设城市级实景三维的重要环节。城市3D Mesh模型主要由顶点、边、三角面片以及纹理影像组成。相比于点云数据，3D Mesh模型不仅可以展示更多的地物细节，而且易于通过调节3D Mesh模型参数来控制表达的精细程度（张力 等，2022）。然而，基于倾斜影像或激光点云构建的3D Mesh模型仅包含一张皮的几何信息，缺乏语义或单体信息，仅适合于城市场景整体上的宏观可视化浏览等简单应用，无法对数据中各个对象进行单独操作和管理，难以满足精细化、智能化城市管理的应用需求。

近年来，深度学习技术得到飞速发展。深度神经网络由于能够从给定的数据集中学到鉴别性强的高阶语义特征，已广泛应用于二维图像数据（Ulku和Akagündüz，2022）和三维数据（如点云、3D Mesh等）（Bronstein 等，2017；Guo 等，2021；Wang 和Zhang，2022；Xiao 等，2020）解译。此外，随着图形处理器（graphic processing unit，GPU）性能的不断提升与标注数据集规模的不断扩大，深度神经网络对二维图像和三维数据的解译精度得到了显著提升。尽管针对3D Mesh模型解译的研究取得了显著的进步（Wang和Zhang，2022），但这些研究主要针对小尺度/小规模的玩具/仿真3D Mesh模型。而面向复杂城市场景3D Mesh模型解译的深度神经网络研究正处于起步阶段，仍面临着许多难题与挑战，主要体现在以下3个方面：1）城市3D Mesh模型不规则、存在孔洞/非水密，传统深度神经网络难以直接应用于城市3D Mesh模型以提取高鉴别性的特征；2）城市3D Mesh模型多尺度特征获取效率低下，传统的3D Mesh简化方法（如二次误差度量等）（Garland和Heckbert，1997）采用的是贪心策略，难以进行并行处理，在处理城市尺度的3D Mesh模型时，不可避免地会增加计算负担；3）城市3D Mesh模型标注困难、训练样本少。相比于二维图像、三维点云基准数据集，公开的城市3D Mesh基准数据集较少。城市场景中的建筑物、道路和植被等结构复杂多样，对城市3D Mesh模型进行标注不仅需要具备专业知识，还需要耗费大量时间和人力资源。

与二维图像、三维点云数据智能解译（仇志江等，2024）相比，虽然深度神经网络在城市3D Mesh模型解译方面的应用起步较晚，但该领域的研究依然取得了迅猛的发展。利用不同深度神经网络架构获取语义化城市3D Mesh模型已广泛应用于城市管理（Skondras 等，2022）、城乡规划（Chen，2022b）、建筑损坏评估（Hong 等，2022）等领域。然而，目前鲜有综述性文章系统探讨、总结各种深度神经网络架构如何实现城市3D Mesh模型解译任务。因此，本文旨在系统性地回顾并总结现有面向城市3D Mesh模型解译的深度神经网络方法，并提出目前研究面临的开放性挑战，为未来研究人员提供参考。

1 国内外研究现状

针对城市3D Mesh模型特殊的数据格式，学者们在研究城市3D Mesh模型智能解译方法时，通常会根据需要将城市3D Mesh模型转换为不同的表达方式（如重心点云、二维图像等）进行处理。本节根据在处理城市3D Mesh模型时的表达方式，对城市3D Mesh模型智能解译方法进行详细分类，如图1所示。本节首先对城市3D Mesh模型解译的传统方法进行简要概述，然后详细回顾面向城市3D Mesh模型解译的深度学习方法。

1.1 传统机器学习方法

随机森林（random forest，RF）（Breiman，2001）是机器学习的主流模型之一，因其高效、鲁棒以及可解释等优点，已广泛应用于分类、回归和异常检测等任务。随机森林是一种基于Bagging策略（Breiman，1996）的集成学习模型，它能够有效地处理非线性问题，并擅长处理大量样本和特征。随机森林由多个决策树组成，通过对所有决策树的预测结果进行投

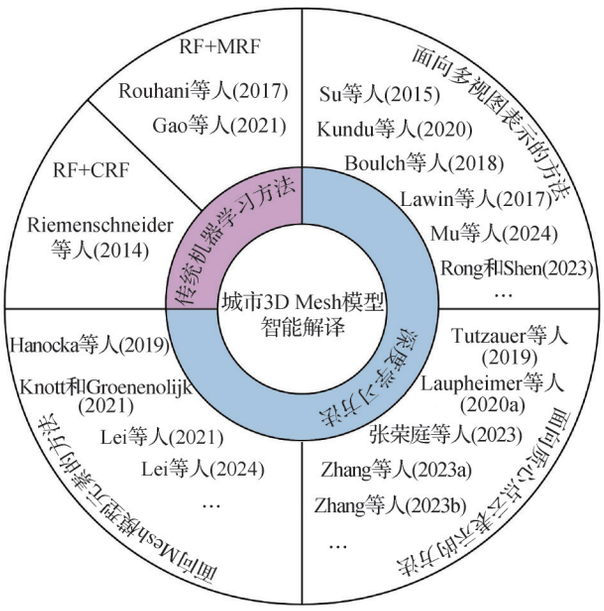


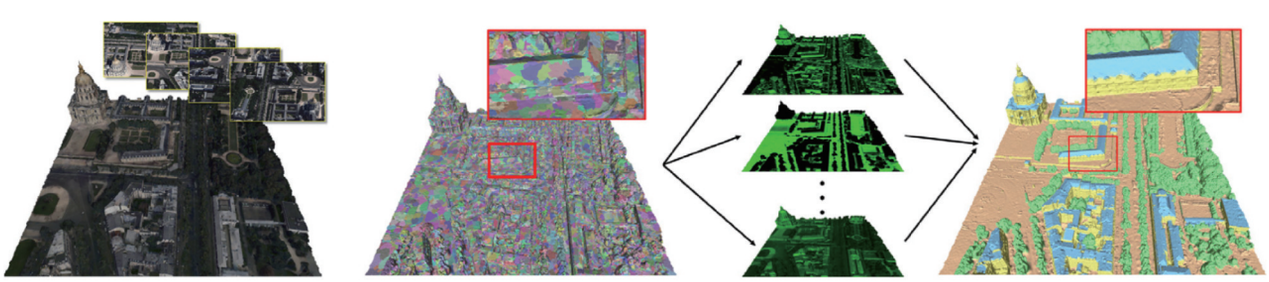
图 1　城市 3D Mesh 模型智能解译方法详细分类

Fig. 1　Detailed classification of intelligent interpretation method for urban 3D Mesh model

票或求平均值来决定最终的预测结果。

因此，在早期的城市 3D Mesh 模型解译研究中，学者们主要基于随机森林对城市 3D Mesh 模型进行语义分割，预测每个三角面片的类别。虽然训练好的随机森林能够预测联合标签，但是当数据中存在几何缺陷和纹理噪声时（于柳和吴晓群，2024），标签预测的空间一致性会受到影响，导致预测精度下降。为了解决上述问题，马尔可夫随机场（Markov random field，MRF）（Cross 和 Jain，1983）及其变种——条件随机场（conditional random field，CRF）（Sutton 和 McCallum，2012）被引入到随机森林模型中，通过能量最小化过程来实现预测标签的平滑。MRF、CRF 能够将局部信息和邻域上下文信息以独立元素（如顶点）和成对元素（如每条边）的形式进行编码，以此来表达城市场景中物体及其组成部分之间的相互依赖关系。具体地，基于 RF-MRF/CRF 方法的城市 3D Mesh 模型解译主要包括 3 个步骤：1）利用无监督聚类算法，如区域生长聚类算法（Cohen-Steiner 等，2004；Lafarge 和 Mallet，2012），对城市 3D Mesh 模型的基元（如三角面片）进行聚类，生成同质的超面片（superfacets），以降低计算时间并提高可扩展性；2）计算超面片的几何和光度特征，并作为 RF 模型的输入，进而预测超面片的标签；3）利用 MRF/CRF 对城市 3D Mesh 模型的预测结果进行平滑处理，得到最终的语义分割结果。如图 2 所示，其中图 2（a）为由多视图立体（multi-view stereo，MVS）图像生成的城市 3D Mesh 模型，图 2（b）为由无监督聚类算法生成的超面片，图 2（c）为超面片的几何、光度特征，图 2（d）为由 RF 预测并经过 MRF/CRF 平滑的城市 3D Mesh 模型解译结果（Rouhani 等，2017）。

作为该研究领域的开拓团队之一，Rouhani 等人（2017）首次提出利用 RF-MRF（random forest Markov random field）模型对由区域生长聚类算法（Cohen-Steiner 等，2004）生成的超面片进行标签预测。该方法在获取城市 3D Mesh 模型的超面片表示后，分别提取了超面片的几何（包括高度、平面度和垂直程度）和光度（包括颜色均值、颜色分布标准差和灰度直方图）特征，并将几何和光度特征进行级联以组成输入向量。该方法利用联合标签空间来表示超面片及其邻居的类别，并通过训练随机森林模型来预测这些联合标签。为了确保超面片类别的空间一致性并细化类别边界，该方法利用 MRF 对预测标签的概率进行了优化。此外，RF 还应用于城市 3D Mesh 模型的标注任务（Gao 等，2021）。Gao 等人（2021）提出



（a）城市3D Mesh模型　（b）超面片　（c）超面片的几何/光度特征　（d）城市3D Mesh模型解译结果

图 2　基于 RF-MRF/CRF 方法的城市 3D Mesh 模型解译整体流程（Rouhani 等，2017）

Fig. 2　Overall workflow for urban 3D mesh interpretation based on the RF-MRF/CRF methods ((a) urban 3D mesh model; (b) super-faces; (c) geometric and photometric features of super-faces; (d) interpretation results of the urban 3D mesh)(Rouhani et al. ,2017)

基于随机森林的半自动城市3D Mesh模型标注方法。该方法首先利用区域生长算法(Lafarge和Mallet,2012)对城市3D Mesh模型的三角面片进行聚类,以获取超面片。超面片的几何(包括高度、面积、三角面片密度和内部半径等)和光度(包括HVS(hue-saturation-value)颜色、绿度)特征被级联为RF模型的输入向量。然后通过训练RF模型,实现对超面片类别的预测。与RF-MRF方法(Rouhani等,2017)不同,Gao等人(2021)通过人工干预的方式对预测的结果进行精细化改正。目前,Gao等人(2021)所发布的SUM(semantic urban meshes)数据集已广泛应用于城市3D Mesh模型智能解译的研究中。与上述方法不同,Riemenschneider等人(2014)利用RF-CRF方法来预测3D Mesh模型中每个三角面片的标签。该方法通过CRF加强了相邻三角面片的空间连接性。

综上所述,基于随机森林的城市3D Mesh模型解译方法具有模型简单、参数可控等优点,并催生了一系列城市3D Mesh模型数据集(如SUM(Gao等,2021)等),促进了城市3D Mesh模型智能解译研究的发展。然而,基于随机森林的城市3D Mesh模型解译方法面临巨大挑战:一方面,基于随机森林的城市3D Mesh模型解译需要人工设计复杂的输入特征,限制了模型的泛化性;另一方面,生成超平面的区域生长算法(Cohen-Steiner等,2004;Lafarge和Mallet,2012)属于不可微的聚类算法,难以实现端到端的学习。随着深度学习技术的不断发展及越来越多的基准数据集被公开,学者们的注意力逐步集中到基于深度学习的城市3D Mesh模型智能解译研究中,旨在实现对大规模城市3D Mesh模型进行端到端的快速、鲁棒解译。

1.2 深度学习方法

深度学习模型具备自动学习复杂特征的能力,能够从原始数据中直接提取鉴别性强的高阶语义信息,而无需依赖于手工设计的特征,这极大地简化了特征工程的复杂性。深度学习已广泛应用于语义分割、实例分割等高层次视觉任务中(Hao等,2020;Ulku和Akagündüz,2022;Yu等,2023)。随着城市3D Mesh模型数据集规模的不断扩大,传统机器学习方法难以高效处理这些大数据量,而深度学习模型则表现出了强大的处理能力。

与规则的二维图像相比,传统的深度卷积网络难以直接处理非规则的3D Mesh数据。因此,为了能够利用二维图像、点云等处理中成熟的深度学习技术,学者们对城市3D Mesh模型进行了相应的转换。本节回顾深度学习在城市3D Mesh模型智能解译中的深度学习方法,并根据表达方式的不同对这些方法进行分类,如图1所示。

1.2.1 面向多视图像表示的方法

在早期基于深度学习的3D Mesh模型智能解译方法中,学者们尝试从多个视角将3D Mesh模型投影为二维图像,并利用成熟的二维图像深度学习方法从投影图像中学习高鉴别性语义特征,然后将从投影图像学到的语义特征/标签映射到3D Mesh模型中。

MVCNN(multi-view convolutional neural network)(Su等,2015)是该类方法的典型代表之一。在MVCNN中,为了获取3D Mesh模型的多视图表示,在不同位置设置了虚拟视点,并利用Phong反射模型(Phong,1998)来渲染3D Mesh模型的多个视图。如何通过多视图的特征来描述3D Mesh面临着巨大挑战。一种简单的方式是,分别为每个视图生成独立的图像描述符,然后基于投票机制或以求平均的方式对这些独立的图像描述符进行合并。为此,对于每个视图,作者验证了两种类型的图像描述符,一种是通过基于Fisher特征向量(Sánchez等,2013)的手工设计的图像描述符,该描述符使用多尺度SIFT(scale-invariant feature transform)表示,通过VLFeat方法(Vedaldi和Fulkerson,2010)来实现;另一种是由VGG-M(Visual Geometry Group-medium)网络(Chatfield等,2014)提取到的特征描述符。相对于SPH(spherical harmonic representation)(Kazhdan等,2003)、LFD(lightfield descriptor)(Chen等,2003)、ShapeNets(Wu等,2015)等传统的3D特征描述符来说,虽然通过多个独立的描述符来表示3D Mesh模型在解译任务中取得了较好的效果,但这种方式效率较低。为此,MVCNN使用一个统一的卷积神经网络(convolutional neural network,CNN)架构来学习如何聚合多个视图的信息。在该架构中,所有的图像特征通过视图汇聚层,以生成一个代表3D形状的单一描述符,并用于后续解译任务。在MVCNN的基础上,Kundu等人(2020)提出一种新的面向3D Mesh语义分割的虚拟多视图融合方法VMVF(virtual multi-view fusion)。VMVF方法的核心思想是使用

从3D场景的“虚拟视角”渲染的合成图像，而不局限于由物理相机获取的原始图像，克服了前人研究（Boulch等，2018；Lawin等，2017）中存在的问题。在获取多个虚拟视图后，VMVF方法使用Xcpetion（Chollet，2017）作为图像特征提取器，并使用DeeplabV3+（Chen等，2018）作为解码器，进而对虚拟视图进行语义特征的提取。为了将2D图像特征投影到3D空间中，VMVF方法首先会在虚拟视图上渲染一个深度通道；然后3D Mesh模型中的顶点被反向投影回每个虚拟视图，只有在像素的深度与该点到相机的距离相匹配的情况下，才对该投影像素的图像特征进行累加。最后对累加的图像特征进行平均，从而得到3D Mesh顶点的融合特征。相较于使用原始视图的多视图融合，VMVF方法中的虚拟视图选择方法（virtual view selection，VVS）有以下优势：1）VVS可以自定义相机的内外参数，以拍摄最有利于2D语义分割任务的图像，并且可以与任何2D数据增强方法相结合。2）通过放松实际相机的物理限制，VVS极大扩展了可以选择的视图集合，允许从不切实际但却有用的相机位置拍摄视图（例如从墙后拍摄），这些位置可以显著提高模型性能。3）VVS允许2D视图捕捉到真实相机难以捕捉的额外通道信息，例如法线等。通过选择和渲染虚拟视图，VMVF方法消除了3D重建过程中常见的相机校准和姿态估计错误。4）在不同尺度上进行一致的视图采样，解决了传统2D CNN的尺度不变性问题。然而，VMVF方法也存在一些局限性。一方面，VMVF方法对3D Mesh模型的解译只使用了虚拟视图的聚合特征，而没有利用3D Mesh模型的几何信息；另一方面，VMVF方法中虚拟视图的选择并没有经过定量的质量评估，因此有可能会导致视图冗余、缺乏信息，甚至会对3D Mesh模型的解译结果产生负面影响。为解决上述问题，Mu等人（2024）提出一种通过选择代表性虚拟视图来实现联合2D-3D场景解译的通用学习框架（learning virtual view selection，LVVS）。LVVS框架主要包括基础模型、评分网络（score network）、回报（reward）模块以及虚拟视图选择模块（virtual view selection module，VVSM）。其中，基础模型可以是任意联合2D-3D场景解译的网络，如BPNet（bidirectional projection network）（Hu等，2021）、MVPNet（multi-view PointNet）（Jaritz等，2019）等。LVVS框架以3D几何信息和2D虚拟视图为输入，经过基础模型后输出3D场景的初始预测结果，然后使用基于强化学习的评分网络来学习关于当前3D场景预测的信息评分图，以指导虚拟视图的选择。

与MVCNN（Su等，2015）、VMVF（Kundu等，2020）、LVVS（Mu等，2024）等使用冗余的多视图图像不同，Rong和Shen（2023）提出基于正射图像的城市3D Mesh模型解译方法（OrthoMeshSeg）。OrthoMeshSeg方法首先通过正射投影从3D Mesh模型中获取RGB图像和高度图像，然后使用改进的ResNet（residual network）（He等，2016）图像语义分割网络对RGB图像和高度图像进行特征提取，并融合输入到预测头，以获得像素级的语义预测，最后将2D图像的解译结果反向投影到3D模型上。在正射投影阶段，OrthoMeshSeg方法从正射图像像素出发，沿着平行于z轴的方向发射一条光线，并利用CGAL（computational geometry algorithms library）工具库来计算该光线与3D Mesh模型的交点，然后将交点信息分配给当前的正射图像像素。在2D图像解译阶段，为了能够提取更具鉴别性的语义特征，Rong和Shen（2023）利用ResNet分别对渲染的RGB图像和高程图像进行特征提取，并在ResNet每一层后利用提出的跨模态特征聚合模块（cross-modality feature aggregation module，CM- FAM）对RGB图像特征和高程特征进行特征融合。ResNet最后一层输出的融合特征被送入基于类别特征的上下文引导模块（CF-CGM），为每个类别计算特征向量，并根据类别特征向量与像素局部特征的相似度将这些类别特征聚合到像素的局部特征中。最后，拼接后的特征被送入DeeplabV3+（Chen等，2018）的解码器进行语义分割。在二维与三维语义融合阶段，OrthoMeshSeg方法结合2D图像的分割结果与3D模型的几何一致性来进行联合优化，以进一步提高城市3D Mesh模型解译的准确性。在InstanceSegMesh（Chen，2022a）中，采用了2D-3D融合技术对城市3D Mesh模型中的建筑进行实例分割。为了实现混合分割过程，Chen（2022a）发布了首个基于3D Mesh模型的实例分割基准数据集——InstanceBuilding，该数据集包含带有标签的无人机图像及其对应的3D Mesh模型。该方法首先使用Swin Transformer（Liu等，2021）对无人机图像中的屋顶进行分割并计算实例掩膜，然后将这些掩膜反向投影到3D Mesh模型中的建筑

物上，生成相应的3D实例。最后，使用MRF来分割3D Mesh模型中建筑物的其余部分。

尽管面向多视图图像表示的方法已经使深度神经网络（deep neural network，DNN）能够处理大规模的城市3D Mesh模型数据，但这类方法仍然面临诸多挑战：一方面，多视图图像生成及语义特征/标签反向投影过程中会引起失真、遮挡等问题，这些因素不可避免地会对城市3D Mesh模型解译结果产生负面影响；另一方面，上述面向多视图图像表示的方法中，主要利用的只有二维图像特征，忽略了3D Mesh模型固有的几何、空间特征，致使学到的语义特征鉴别性不高，影响3D Mesh模型的解译精度。

1.2.2　面向质心点云表示的方法

三维点云是另一种应用广泛的三维城市场景表示形式，其具有数据组织简单的特点。学者们对三维点云解译进行了大量研究，发展了一系列先进的深度学习方法（Gao等，2021；Zhang等，2019）。其中，PointNet（Qi等，2017a）和PointNet++（Qi等，2017b）被认为是点三维云处理中的里程碑方法。PointNet在直接学习无序点云方面具有开创性的意义。在PointNet中，多层感知器（multilayer perceptron，MLP）被用于特征提取，而T-Net则用于处理点云的无序性与几何旋转。尽管PointNet在处理无序点云数据方面表现出色，但也存在一些明显的缺陷，如缺乏局部结构信息、特征表达能力有限以及缺乏上下文信息等。为了克服这些局限性，PointNet++在PointNet的基础上对特征提取模块进行了改进，通过多层次分组来提取不同尺度上的特征，增加了对局部结构信息的捕捉能力。具体而言，PointNet++在每个层次上都会对邻近点进行分组，并在每个分组内应用PointNet子网络来提取局部特征。通过多层次的特征提取及由FPS（farthest point sampling）方法生成均匀分布的种子点，能够在处理大规模点云数据时更好地捕捉局部特征，并且保持较高的计算效率和鲁棒性。在PointNet和PointNet++的基础上，涌现出了一系列基于点的深度学习方法。

与点云相比，城市3D Mesh模型能够提供更多的信息（如高分辨率纹理、显式的表面连接性），并且能够更精细地表达变化剧烈的区域。鉴于三维点云智能解译方法的优异性能及其数据组织的便捷性，学者们尝试将三维点云智能解译方法应用于城市3D Mesh模型解译研究中。一般而言，城市3D Mesh模型中的每个面片通过其质心点来表示，进而将整个3D Mesh模型抽象成质心点云（centre of gravity，COG）。之后，利用三维点云中的智能解译算法对这些质心点云进行处理。不同于传统的三维点云，质心点云能够从城市3D Mesh模型中继承丰富的纹理和几何信息。Tutzauer等人（2019）探讨了Multi-Branch-1D-CNN方法（George等，2018）在城市3D Mesh模型解译方面的潜力。在Multi-Branch-1D-CNN方法中，作者首先计算每个三角面片的多尺度辐射和几何特征向量，构建质心点云，即每个三角面片由一个归一化的特征向量来表示，然后输入到1D CNN中进行特征提取。然而，Multi-Branch-1D-CNN方法只考虑全局的特征，而忽略了局部特征，导致局部细节丢失、特征鉴别能力不足等问题，影响解译结果的精度。为了衡量纹理信息在城市3D Mesh模型解译中的重要性，Laupheimer等人（2020b）将城市3D Mesh模型用质心点云来表示，而且质心点云附带的特征信息只包括三角面片所对应的纹理信息（如HSV中位数、均值等）与三角面片的法向量，最后利用面向点的深度神经网络对质心点云进行解译，从而实现对3D Mesh模型的解译。Laupheimer等人（2020b）在消融实验中比较了RF（Breiman，2001）、PointNet（Qi等，2017a）、PointNet++（Qi等，2017b）和Multi-Branch-1D-CNN（Tutzauer等，2019）的解译能力，结果表明PointNet++获得了最好的解译结果。这主要是因为PointNet++具有层次化的特征学习能力，能够更好地捕捉局部和全局信息。同时，消融实验也验证了纹理信息的重要性，其可以在全局范围内提高3D Mesh模型的解译精度。鉴于纹理信息在城市3D Mesh模型解译中的重要性，张荣庭等人（2023）在提出的复杂城市动态图卷积网络三维场景语义分割方法3DCity-Net中，利用3D Mesh模型固有的三维空间坐标信息和纹理信息构建的复合特征向量来表示3D Mesh模型中的三角面片。为降低纹理信息中噪声和冗余信息对城市3D Mesh模型解译精度的影响，3DCity-Net在骨干网络DGCNN（dynamic graph convolutional neural network）（Wang等，2019a）中嵌入了主成分分析模块。虽然3DCity-Net能够在一定程度上降低纹理信息中噪声和冗余信息对3D Mesh模型解译精度的影响，但是主成分分析模块是不可微的，难以实现端到端训练。为解决这一问题，Zhang等人（2023b）在3DCity-Net基础

上提出 Mesh-based DGCNN 城市 3D Mesh 模型解译架构。Mesh-based DGCNN 在网络结构中嵌入纹理信息融合模块 TIF，实现了对纹理信息的端到端处理，进而提取纹理高阶信息。虽然上述纹理信息融合模块能够抑制纹理影像噪声和冗余信息对深度神经网络性能的影响，但是纹理信息融合模块是“黑盒子”，难以解释其背后的机制，可解释性弱。由于稀疏建模能够容易地从影像块中发现有意义的数据结构，且具有学习可解释表征的能力和较强的理论保障，经典的稀疏建模在信号、影像恢复等任务中得到了广泛的应用。因此，针对上述问题，Zhang 和 Zhang（2023）在 Mesh-based DGCNN 的基础上通过融合稀疏建模理论和展开优化算法，提出稀疏先验引导的城市 3D Mesh 模型解译网络 MeshNet-SP。MeshNet-SP 基于 LISTA（learned iterative shrinkage and thresholding algorithm）（Evtimova 和 LeCun，2022；Gregor 和 LeCun，2010）框架实现了可微的稀疏编码模型（differentiable sparse coding module，DSC）。消融实验结果表明，DSC 模块能够让 MeshNet-SP 在不同水平的高斯噪声下仍具有较好鲁棒性，能够获得具有较高精度的城市 3D Mesh 解译结果。

与上述方法不同，为了增强质心点云特征向量的信息量，Laupheimer 等人（2020a）提出联合 LiDAR（light detection and ranging）点云与城市 3D Mesh 模型的多模态解译框架，通过点云—Mesh 联合（point cloud mesh association，PCMA）机制，由 LiDAR 点云生成的手工特征被转移到对应的三角面片（质心点云）特征向量中，而经过 PointNet++网络预测的 3D Mesh 模型标签也会转移到对应的 LiDAR 点云上，从而实现 LiDAR 点云数据的标注。通过使用 PCMA 机制，MultiModal- Net（Laupheimer 和 Haala，2022）在消融实验中，进一步验证了结合多模态特征在城市 3D Mesh 模型解译任务中的优越性。

与转换 3D Mesh 模型为质心点云的解译框架、联合 LiDAR 点云与 3D Mesh 模型的解译框架不同，Wilk 等人（2022）不直接处理 3D Mesh 模型，而是处理生成 3D Mesh 模型的源数据，如倾斜影像和 LiDAR 点云。Wilk 等人（2022）首先分别利用 PSP-Net（Zhao 等，2017）和 OPEGIEKA’s 方法（Dominik 等，2021）对倾斜影像和 LiDAR 点云进行解译，然后将解译的结果融合到 3D Mesh 模型的生成过程中，最终得到带有语义信息的 3D Mesh 模型。而 Grzeczkowicz 和 Vallet（2022）则是利用纹素采样（texel sampling）或泊松采样（Poisson disk sampling）方法直接从城市 3D Mesh 模型表面采样三维点云数据，然后利用先进的三维点云深度学习方法 KPConv（kernel point convolution）（Thomas 等，2019）对采样的三维点云进行解译，最后将三维点云解译结果反向映射到城市 3D Mesh 模型中。通过在城市 3D Mesh 模型表面进行采样来获取三维点云有以下几个优势：1）可通过控制采样点数量来平衡解译的性能与质量；2）三维点云的具有明确的法向量；3）生成的三维点云附带纹理信息；4）三维点云的标签可以容易地反向映射到 3D Mesh 的顶点或三角面片。Grzeczkowicz 和 Vallet（2022）通过消融实验验证了由泊松采样生成的三维点云能够得到更好的解译结果。

上述方法中，无论是将城市 3D Mesh 模型抽象为质心点云，还是从城市 3D Mesh 模型表面上采样三维点云，在处理这些点云数据时通常采用 KNN（K-nearest neighbor）等方法来获取邻居，并没有充分利用原始城市 3D Mesh 模型中元素（顶点、三角面片和边）间固有的拓扑关系。将几何拓扑中的图论概念引入到点云分析中的研究（Wang 等，2019a，b；Zhang 和 Zhang，2023），为城市 3D Mesh 模型解译提供了一种新的思路。GraphTransMesh（Tang 等，2022）将每个三角面片的质心点 COG（center-of-gravity）视为图（graph）的顶点，而图的边则由三角面片的相邻关系决定，如图 3 所示，其中蓝色点代表红色点的一环邻居。

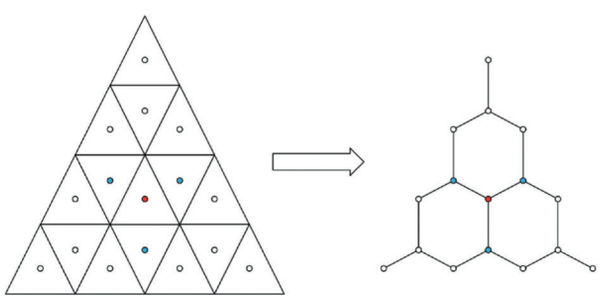


图 3　COG 图结构生成示意图（Tang 等，2022）

Fig. 3　Schematic diagram of COG graph structure generation (Tang et al. , 2022)

为了获取多尺度的特征，GraphTransMesh 基于 Transformer 和层次结构，设计了一个类似 U-Net（Ronneberger 等，2015）的网络结构，如图 4 所示，并提出用于 COG 图降采样和上采样的基于边距离的采样方法，如图 5 所示，实现了高效的层次特征提取。

在上述所有方法中，大多数方法在利用纹理信息时，只是简单地计算了每个三角面片所覆盖纹理的均值、方差等信息，并没有完全利用纹理图像的信息，难以从纹理图像中获取高阶的图像特征。为了解决这一问题，Yang等人（2023b）在GraphTrans-Mesh（Tang等，2022）基础上新增了纹理卷积模块TextureConv。TextureConv首先将三角纹理映射为矩形纹理，然后利用二维卷积神经网络从矩形纹理中提取高阶的纹理特征后赋给对应的三角面片质心点。

与面向多视图像表示的方法类似，面向质心点云表示的方法首先需要将城市3D Mesh模型转换为质心点云，然后利用面向点云的深度神经网络的特征学习能力，并结合城市3D Mesh模型的几何、纹理等相关信息来进行城市3D Mesh模型元素级别的解译。然而，通过质心点云数据格式对城市3D Mesh模型进行解译的间接方法，由于伴随采样、投影或质心抽象技术而导致的信息损失，在一定程度上影响了这些方法的性能。另一方面，城市3D Mesh模型自身的质量问题也会转移到质心点云上，例如城市3D Mesh模型中三角面片大小、数量的不均匀会影响地物样本的均衡，如较为平坦的地物（如墙面）的三角面片数量较非平坦的地物（如植被）少。此外，大多数面向质心点云表示的方法没有充分利用城市3D Mesh模型中元素间固有的拓扑关系，以及从纹理图像中提取高阶的语义特征。

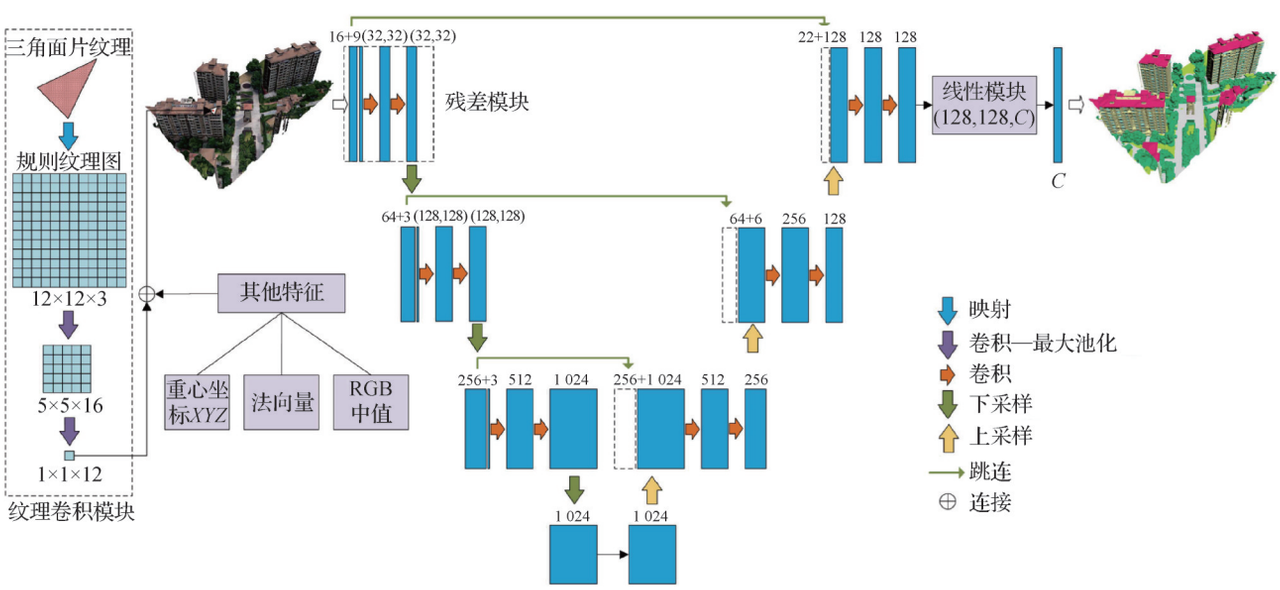


图4　面向城市3D Mesh模型解译的多尺度深度神经网络结构（Yang等，2023c）

Fig. 4　Multi-scale deep neural network architecture for urban 3D mesh model interpretation (Yang et al., 2023c)

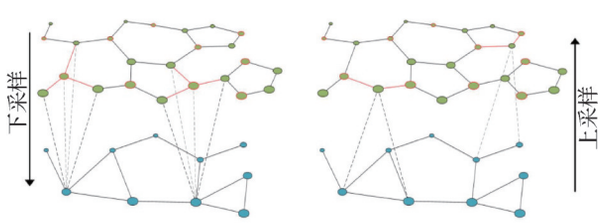


图5　COG图简化及表面连接邻居示意图（Yang等，2023b）

Fig. 5　Schematic diagram of COG graph simplification and surface connection neighbors (Yang et al., 2023b) ((a) the downsampling processes of the COG graph; (b) the upsampling processes of the COG graph)

1.2.3　面向3D Mesh模型元素的方法

城市3D Mesh模型同时表达了几何、拓扑以及高分辨率的纹理信息。面向3D Mesh模型元素的方法旨在直接在3D Mesh模型的元素（顶点、边和三角面片）上定义可学习的操作（如卷积、池化操作等），直接从城市3D Mesh模型中学习和提取丰富的高阶语义特征，以避免因多视图图像投影、质心点云抽象等预处理过程导致的信息损失。

在计算机图形学领域，MeshCNN（Hanocka等，2019）是直接在3D Mesh模型元素上定义可学习操作的典型代表之一。与经典CNN类似，MeshCNN针对3D Mesh模型元素定义了卷积、池化层和上池化层。MeshCNN的网络结构与U-Net网络结构类似，其输入的是5维的边特征，包括二面角、两个内角以

及两个边长比。MeshCNN 中的卷积操作主要应用于边及其关联的 4 个三角形的边上，而池化则是通过一种保持表面拓扑结构的边塌缩方法来实现。池化操作为后续的卷积生成了新低分辨率的 3D Mesh 模型。MeshCNN 通过学习来决定哪些边应该被塌缩，从而形成一个任务驱动的过程。在这个过程中，MeshCNN 网络将对重要的特征进行扩展，而丢弃掉冗余的特征。然而，MeshCNN 只在仿真的小尺度玩具 3D Mesh 数据集上得到了验证。为了将 MeshCNN 应用到真实的大尺度城市 3D Mesh 模型解译中，Knott 和 Groenendijk（2021）提出 RS-MeshCNN 方法，主要对 MeshCNN 进行了两方面的改进：1）使用广度优先搜索（breadth-first search，BFS）方法来生成易于管理的场景块，以实现高效处理；2）在现有的几何特征基础上增加光度特征，以提升网络的判别能力。

与 MeshCNN（Hanocka 等，2019）不同，Picasso（Lei 等，2021）和 PicassoNet-II（Lei 等，2024）针对 3D Mesh 模型的顶点和三角面片定义了卷积操作，主要包括 3 类：facet2vertex 卷积、vertex2facet 卷积和 facet2facet 卷积。其中，facet2vertex 卷积通过聚合与顶点相邻的三角面片的上下文信息作为对应顶点的特征，而不是从邻近的顶点进行聚合；vertex2facet 卷积则将三角面片顶点的特征进行聚合后赋给该三角面片；facet2facet 卷积则基于三角面片内部所有采样点的颜色来学习其纹理特征。为了能够在线进行池化操作，Picasso（Lei 等，2021）利用 GPU 实现了基于 QEM（quadric error metrics）（Garland 和 Heckbert，1997）的 3D Mesh 模型快速简化方法。这使得对大规模 3D Mesh 模型进行解译的网络能够实现端到端的训练。PicassoNet-II（Lei 等，2024）则使用该快速简化技术来生成多分辨率的 3D Mesh，并从中提取多尺度的特征。

不同于面向多视图像表示的方法和面向质心点云表示的方法，面向 3D Mesh 模型元素的方法无需对 3D Mesh 进行视图投影、点云抽象，而是直接在 3D Mesh 模型的顶点、边、三角面片上定义卷积和池化等操作，直接进行 3D Mesh 模型特征的学习和处理，有效地利用了 3D Mesh 模型内在的几何信息。然而，在处理大规模城市 3D Mesh 模型数据时，虽然通过 BFS 技术（Knott 和 Groenendijk，2021）、GPU 加速的 QEM 简化技术等能够实现对大规模城市 3D Mesh 模型解译网络的端到端训练，但是其过程较为复杂，仍需要有效的方法高效地进行大规模城市 3D Mesh 模型的解译。

2　国内外研究进展比较

随着城市数字化进程的加快，城市三维建模技术得到了迅猛发展，3D Mesh 模型成为城市三维信息的重要载体。城市 3D Mesh 模型具有高保真度和丰富的几何、纹理信息等特点，在城市规划、建筑设计和灾害应急响应等领域展现出巨大的应用潜力。然而，面对日益庞大的城市 3D Mesh 模型数据集，传统的机器学习方法（如 RF-MRF）逐渐暴露出处理效率低下、特征提取能力有限等问题，而深度学习凭借其强大的特征学习能力和对大数据的高效处理能力，已成为解决 3D Mesh 模型智能解译难题的关键技术。在城市 3D Mesh 模型的智能解译中，根据城市 3D Mesh 模型表示方式的不同，深度学习方法可以分为面向多视图表示的方法、面向质心点云表示的方法以及面向 3D Mesh 模型元素的方法。3 种方法的优缺点如表 1 所示。下面对这 3 种方法进行详细

表 1　3 种方法的优劣比较

Table 1　Comparison of advantages and disadvantages of three methods

类别	基本原理	优劣对比
面向多视图表示的方法	将城市 3D Mesh 模型投影到多个二维图像，利用二维图像深度学习方法	成熟的二维图像深度学习方法，灵活的视图选择；信息损失，忽略几何信息。
面向质心点云表示的方法	将城市 3D Mesh 模型中的每个面片表示为其质心点，利用三维点云深度学习方法	数据组织简单，局部特征捕捉，纹理信息利用；信息损失，拓扑关系未充分利用，样本分布不均衡。
面向 3D Mesh 模型元素的方法	直接在 3D Mesh 模型元素（顶点、边和三角面片）上定义卷积、池化等操作	直接处理 3D Mesh 数据，拓扑信息保留；计算复杂度高，训练难度增加。

的对比分析。

1)面向多视图表示的方法。通过将城市3D Mesh模型投影到多个二维图像上,然后利用成熟的二维图像深度学习技术来对这些投影图像进行特征提取,最后将学到的特征/标签反向映射到城市3D Mesh模型中。该方法的最大优势在于其能够充分利用已有二维图像深度学习技术,简化了特征工程的复杂性。然而,面向多视图表示的方法也存在一些固有的问题:(1)从城市3D Mesh模型到二维图像的投影过程会导致几何信息的损失,尤其是在处理遮挡和失真方面;(2)该方法忽略了3D Mesh模型本身的几何、拓扑等信息,仅依靠二维图像特征,导致解译结果的准确性受限。

2)面向质心点云表示的方法。通过将城市3D Mesh模型中的每个面片表示为其质心点,从而形成质心点云,然后利用先进的三维点云的深度学习技术对质心点云进行解译。由于质心点云与城市3D Mesh模型中的三角面片是一一对应的,因此质心点云的解译结果即为城市3D Mesh模型的解译结果。该方法在城市3D Mesh模型解译研究中占主导地位,其优势在于其数据组织相对简单,可通过KNN算法等或城市3D Mesh自身的拓扑关系来构建图模型,并能够从中有效地提取3D Mesh模型的局部特征、全局特征。此外,质心点云可以携带纹理信息,对提高城市3D Mesh模型解译精度有着重要作用。然而,面向质心点云表示的方法同样会导致几何信息的损失。特别地,当城市3D Mesh模型转换为质心点云后,利用KNN等方法根据质心点云构建图模型时忽略了城市3D Mesh模型元素间固有的拓扑关系。另外,城市3D Mesh模型的质量问题也会转移到质心点云上,如三角面片的数量和大小不均匀,导致样本不平衡。

3)面向3D Mesh模型元素的方法。直接处理3D Mesh模型的元素(如顶点、边和三角面片),定义了专门针对3D Mesh模型的卷积、池化等操作,可以直接从3D Mesh模型中提取丰富的高阶语义特征。这种方法避免了中间转换步骤带来的信息损失,保留了3D Mesh模型的拓扑结构,有助于从城市3D Mesh模型中提取高鉴别性的特征表示。然而,直接处理城市3D Mesh模型带来了更高的计算复杂度,并且要求特殊设计的网络架构,增加了训练的难度。特别地,在处理大规模城市3D Mesh模型数据时,尽管通过BFS技术和GPU加速的QEM简化技术等方法可以实现端到端的训练,但这一过程依然复杂,仍需有效的方法来实现高效的大规模城市3D Mesh模型解译任务。

3 数据集

随着传感技术、摄影测量技术的不断发展,城市3D Mesh模型重建变得越来越高效。城市3D Mesh数据标签的获取通常通过人工标注或从图像和点云数据中获得标签后投影到3D Mesh模型。由于城市3D Mesh模型标注困难,不仅需要具备专业知识,还需要耗费大量时间和人力资源,因此相比于二维图像、三维点云基准数据集,公开的城市3D Mesh基准数据集较少。针对城市3D Mesh模型解译任务,本文对主要的城市3D Mesh数据集的进行了简要概述,如表2所示。

表2 城市3D Mesh数据集

Table 2 Urban 3D mesh datasets

名称	类别数量	发布年份	范围/km^2	三角面数量/M	纹理
Hessigheim 3D(H3D)	11	2021	0.19	~37	√
SUM	6	2021	4.00	19	√
InstanceBuilding	12	2022	0.43	~8	√
Wuhan	7	2022	0.53	~20	√
Urban-BIS	10	2023	10.78	2 843	√
CUS3D	10	2024	2.85	~290	√

注:"√"表示包含。

1）H3D（Hessigheim 3D）数据集（Kölle 等，2021）。H3D Mesh 数据集由德国斯图加特大学（University of Stuttgart）和德国联邦水文研究所（German Federal Institute of Hydrology，BfG）在对德国 Hessigheim 村庄进行地面沉降检测时，通过 RIEGL VUX-1LR 扫描仪和两台 Sony Alpha 6000 倾斜相机获取倾斜影像和 LiDAR 点云后，再由 SURE（surface reconstruction）软件生成得到。H3D Mesh 数据集包括 2018 年 3 月、2018 年 11 月、2019 年 3 月和 2016 年 3 月共 4 期数据，如图 6 所示。在 H3D Mesh 数据集中，地物被划分为 11 个类别：低矮植被（low vegetation）、不透水表面（impervious surface）、车辆（vehicle）、城镇基础设施（urban furniture）、屋顶（roof）、房屋外立面（facade）、灌木（shrub）、树木（tree）、土壤/砾石（soil/gravel）、垂直表面（vertical surface）和烟囱（chimney）。H3D Mesh 以瓦片形式提供，以 2018 年 3 月时期的数据为例，其共包含 51 块 50 m × 50 m 的瓦片数据，其中训练集包含 43 块，验证集包含 8 块。数据链接：https://ifpwww.ifp.uni-stuttgart.de/benchmark/hessigheim/default.aspx。

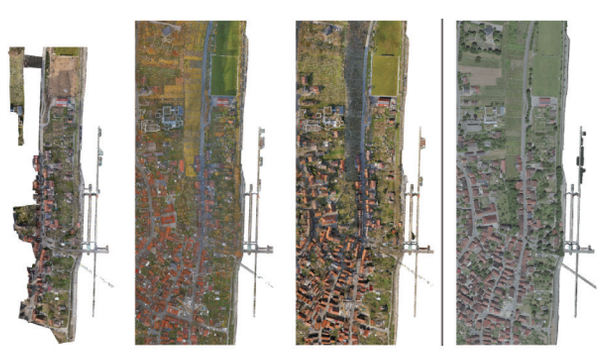
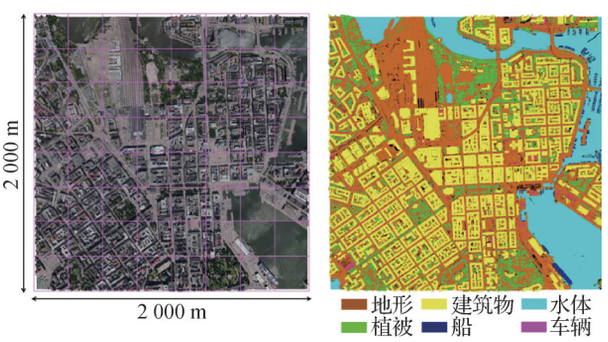


图 6　不同时期的 H3D Mesh 数据（Kölle 等，2021）

Fig. 6　H3D mesh data at different times（Kölle et al.，2021）

2）SUM 数据集（Gao 等，2021）。SUM Mesh 数据由商业软件 ContextCapture 从倾斜航空图像中重建而成，示意图如图 7 所示。原始倾斜影像的地面分辨率为 7.5 cm，覆盖芬兰赫尔辛基（Helsinki）市内约 12 km²，涵盖了不同类型的城市场景，如住宅区、商业区和公共空间，为训练和评估提供了丰富的样本。在进行表面重建时，为恢复不符合朗伯假设的 3D 水体，利用 2D 矢量图和正射影像进行贴图。此外，空中三角测量、密集图像匹配和网格曲面重建等处理均由 ContextCapture 完成。SUM 数据集包含约 1 900 万个三角形，覆盖约 4 km²，涵盖城市环境中常见的 6 类对象：地形、高植被、建筑、水体、车辆和船只。SUM 数据集被划分为 64 块，每块面积约为 250 m²，其中训练数据集 40 块，测试数据集和验证数据集分别为 12 块。数据链接：https://3d.bk.tudelft.nl/projects/meshannotation/。

3）InstanceBuilding 数据集（Chen 等，2022a）。该数据集是第 1 个专门用于城市场景中建筑实例分割的数据集，对 3D Mesh 城市场景和 2D 无人机图像中的建筑物进行了实例级标准，示意图如图 8 所示。该数据集共包含两种数据类型：1）608 幅无人机图像，共标注超过 16 000 个屋顶实例；2）4 个 3D Mesh 场景，共标注 892 个屋顶或建筑物，其中场景 3 和场景 4 包含纹理信息，而场景 1 和场景 2 不包含。数据集总共包含大约 374 万个顶点，747 万个三角面片。数据链接：https://californiachen.github.io/datasets/InstanceBuilding。

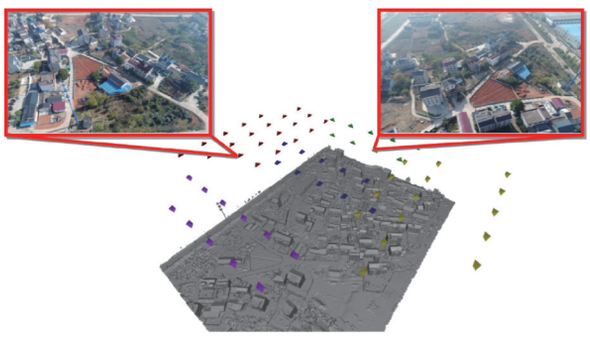


(a) 纹理SUM Mesh　(b) 标签真值

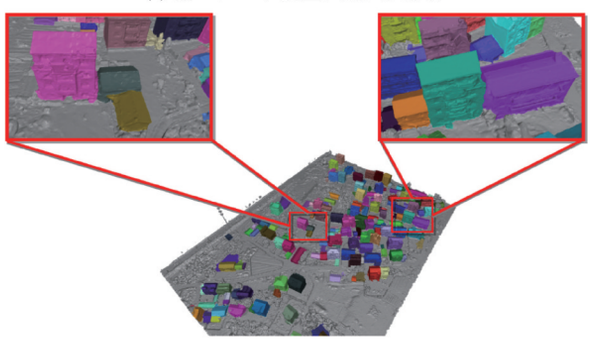
图 7　SUM 数据集示意图（Gao 等，2021）

Fig. 7　Schematic diagram of the SUM dataset（Gao et al.，2021）（(a) textured SUM mesh；(b) label ground truth）

4）Wuhan 数据集（Tang 等，2022）。该数据采集自武汉的一个住宅区，覆盖面积约为 0.53 km²，包含约 2 000 万个面，示意图如图 9 所示。Tang 等人（2022）使用 CloudCompare 软件手动标注数据集，并进行二次检查以确保质量。Wuhan Mesh 数据集中地物被划分为 7 个类别，分别是屋顶、外墙、窗户、不透水面、树木、车辆和低矮植被。不透水面主要包含道路，而外墙包括建筑外墙、阳台、屋顶细节以及空调外机空间。数据集被随机分为 54 个区块，训练集、验证集和测试集分别包含 24、8 和 22 个区块。数据链接：https://figshare.com/articles/dataset/Wuhan_data_and_code/16681849?file=35231683。



(a) 输入：三维模型和无人机影像



(b) 输出：三维建筑物实例

图8　InstanceBuilding数据集示意图(Chen等,2022a)

Fig. 8　Schematic diagram of the InstanceBuilding dataset (Chen et al., 2022a) ((a) input: 3-D model and UAV images; (b) output: 3-D building instances)

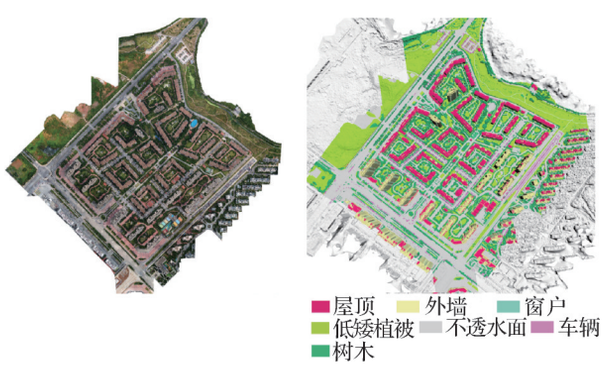


图9　Wuhan数据集示意图(Yang等,2023c)

Fig. 9　Schematic diagram of the Wuhan dataset (Yang et al., 2023c)

5) Urban-BIS(urban building instance segmentation)数据集(Yang等,2023a)。包括在不同城市的3个大场景(青岛、芜湖和龙华)、两个校园场景(粤海、丽湖)和一个小住宅区(银石),总面积达到10.78 km^2,包含约3 370栋建筑,是目前最大的三维真实场景数据。该数据集提供包括图像、点云以及3D Mesh在内的海量多模态数据和三维语义标注与建筑物实例标注。Urban-BIS提供了两类语义信息:城市级别语义和建筑级别语义。其中,城市级别共有7类,分别为:地面、水体、船舶、植被、桥梁、车辆和建筑物;建筑物级别共有7类,分别为:商业建筑、居住建筑、办公建筑、文化建筑、交通建筑、市政建筑和临时建筑。数据链接:https://vcc.tech/UrbanBIS。

6) CUS3D (comprehensive urban-scale semantic segmentation 3D)数据集(Gao等,2024)。包含精细标注的3D点云和3D Mesh数据类型,以及具有详细2D语义标签的高分辨率原始2D图像。CUS3D 3D Mesh由10 840幅UVA(unmanned aerial vehicle)航拍图像的重建得到,占地约2.85 km^2,涵盖了城市和农村的场景。在CUS3D 3D Mesh数据集中,地物被划分为10个类别:建筑物、草地、地面、农田、运动场、道路、车辆、植被、湖面和建筑工地。数据链接:https://github.com/CapyLin-G/CUS3D。

4　城市3D Mesh模型解译任务

城市3D Mesh模型解译的典型任务之一是对3D Mesh模型进行语义分割。假设$\boldsymbol{m}\in(V, F, E)$,其中,$\boldsymbol{m}=(m_1, m_2, \cdots, m_n)$为3D Mesh模型中元素的集合,$V$、$F$、$E$分别表示顶点、三角面片和边。3D Mesh语义分割指的是为通过深度神经网络从一组可能的语义标签中为元素m_i分配一个标签。深度神经网络模型的性能通常可通过以下指标进行衡量:交并比(intersection over union,IoU)、每类平均交并比(mean per-class intersection over union,mIoU)、精确率(precision)、召回率(recall)、准确率(accuracy,Acc)、整体准确率(overall accuracy,OA)、每类平均准确率(mean per-class accuracy,mAcc)以及F1分数。SUM数据集和Wuhan数据集语义分割定量结果比较分别如图10和图11所示。

从图10的结果来看,在SUM数据集的语义分割任务中,Grzeczkowicz和Vallet(2022)方法(Sample-PC-Mesh)在所有指标中表现更好,而PSSNet(planarity-sensible semantic segmentation network)(Gao等,2023)、GraphTransMesh(Tang等,2022)、Yang等人(2023a)、Yang等人(2023b)、Mesh-based DGCNN(Zhang等,2023b)以及MeshNet-SP(张荣庭等,2023)间的性能没有明显的差异。Sample-PC-Mesh之所以能够遥遥领先于其他方法,其最主要的原因是Sample-PC-Mesh通过泊松采样方法从3D Mesh表面进行了密集采样,然后利用点云处理中知

名的KPConv(Thomas等,2019)网络进行语义分割。这种方法能够让Sample-PC-Mesh充分发挥KPConv网络的微调能力,从而能够获得比其他基于质心点云表示的方法更优越的性能。

从图11的结果来看,在Wuhan数据集的语义分割任务中,GraphTransMesh(Tang等,2022)、Yang等人(2023a)以及Yang等人(2023b)在所有指标上的表现均无明显差异,Yang等人(2023b)的方法只是略高于其余两个方法。这其中主要的原因在于这3个方法的网络架构大体上类似,Yang等人(2023b)方法在GraphTransMesh基础上新增了纹理卷积模块TextureConv。TextureConv首先将三角纹理映射为矩形纹理,然后利用二维卷积神经网络从矩形纹理中提取高阶的纹理特征后赋给对应的三角面片质心点。然而,新增的纹理卷积模块TextureConv只是对单个三角面片内的纹理进行卷积,而没有顾及其邻域纹理信息。在城市3D Mesh模型中,单个三角面片所覆盖的面积很小,特别是空间变化剧烈的区域(如树叶),因此,单个三角面内所包含的纹理信息量十分有限,导致纹理卷积模块TextureConv难以学到高鉴别性的纹理特征。

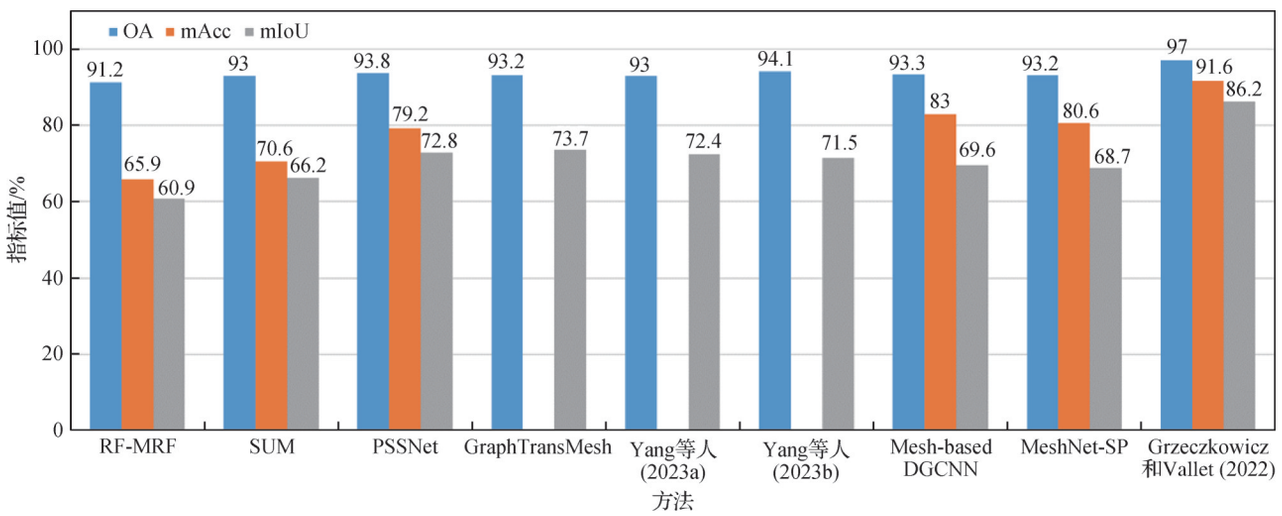


图10　SUM数据集上语义分割定量结果比较

Fig. 10　Quantitative comparison of semantic segmentation results on the SUM dataset

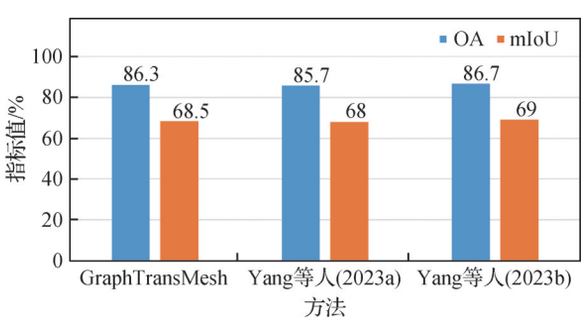


图11　Wuhan数据集上语义分割定量结果比较

Fig. 11　Quantitative comparison of semantic segmentation results on the Wuhan dataset

5　结语与展望

虽然面向城市3D Mesh模型解译的深度学习方法得到了迅猛发展,但是当前想要获得更精确的城市3D Mesh模型解译结果仍面临巨大挑战。在前文综合介绍各类面向城市3D Mesh模型解译的深度学习方法及其对比的基础上,以下提供一些具有研究潜力的未来研究方向,以及亟待解决的挑战性问题。

1)建立大规模高质量数据集。大规模数据集(如ImageNets数据集包含千万幅图像)是深度学习方法在图像、自然语言处理等领域取得突破性效果的必要条件之一。但是城市3D Mesh数据集的数量远不及图像、文本和视频等数据。例如SUM数据集只包含64块数据,覆盖总面积只有4 km^2,数据量远小于图像数据。而由小规模城市3D Mesh数据集训练得到的深度神经网络模型通常容易出现过拟合现象,导致深度神经网络模型难以泛化到其他数据集,制约了深度神经网络模型的性能发挥。高质量城市3D Mesh数据集的建立涉及数据采集、处理和标注等多个环节,是一项复杂且具有挑战性的任务,未来

研究可从以下两方面考虑：

（1）利用先进三维重建方法提高城市3D Mesh模型重建效率。基于摄影测量的城市3D Mesh模型重建涉及特征点提取与匹配、姿态估计、稠密点云生成、表面重建以及纹理映射等多道复杂工序，制约了城市3D Mesh模型重建效率。此外由于遮挡、阴影等影响，容易导致传统三维重建方法生成的3D Mesh模型产生空洞、纹理映射错误等问题。近年来，基于神经辐射场（neural radiance field，NeRF）、3D高斯溅射（Gaussian splatting，GS）等先进方法的三维重建在计算机视觉与图形学领域取得了重大突破，其时效性与重建质量得到显著提升，且可以渲染出任意视角下的图像，有利于解决遮挡、纹理映射等问题。NeRF的核心思想是将场景表示为一个连续的、可微的辐射场，隐式地表示场景的几何和外观信息。而3D GS则是在三维空间中生成高斯分布的点，通过优化这些高斯分布的参数（如位置、颜色和方差等），实现对场景的高效重建。将NeRF、3D GS应用于城市场景的三维重建中时，如何有效地将隐式辐射场、离散的高斯球转换为显式的、表面连续的3D Mesh模型；以及如何对大范围的城市场景进行快速并行处理都需要进一步讨论研究。

（2）扩充城市3D Mesh数据增强方法。数据增强是增加样本数量的有效途径之一。虽然目前存在一些3D Mesh数据增强的方法，如旋转、缩放、平移和翻转等，但是这些方法难以满足深度神经网络对3D Mesh数据多样性的需求。为此，可以借鉴图像处理中的混合数据增强技术（如MixUp、CutMix、CutOut）来增加数据的多样性和复杂性。此外，使用生成对抗网络（generative adversarial network，GAN）、Diffusion扩散模型等高级数据增强方法来生成合成的3D Mesh数据是将来的研究热点之一。GAN和Diffusion扩散模型分别通过对抗训练、逐步添加和移除噪声模型，学习数据的复杂分布，其可以控制生成的3D Mesh模型的特定属性，如形状、纹理等，可控性、稳定性较高。将GAN和Diffusion扩散模型等方法应用于城市3D Mesh数据增强时，面临生成速度慢、计算资源需求高等问题，如Diffusion扩散模型需要多次迭代来生成高质量的3D Mesh模型。因此，将来需研究高效的训练方法，如自适应学习率、梯度裁剪等；设计轻量级的生成模型减少参数量，提高生成速度，降低计算资源需求。

2）构建城市三维场景多模态大模型。虽然语言大模型（large language model，LLM）、视觉语言大模型（vision large language model，VLLM），如Llama（Touvron等，2023）、ChatGPT（Katz等，2024）和SAM（segment anything model）（Kirillov等，2023）等，在语言理解、生成和翻译，以及图像描述生成、图像语义分割等任务中表现出卓越能力，但它们并未基于三维物理世界。相较于自然语言与图像，三维物理世界涉及更丰富的概念，如空间关系、物理特性和布局等。LLM或VLLM难以适用于三维环境的理解，以及基于三维理解进行推理和规划等。城市三维场景多模态大模型的构建将有助于复杂场景的理解、模拟和分析，如自动驾驶、灾害模拟和城市规划等。因此，构建城市三维场景多模态大模型势在必行。

构建城市三维场景多模态大模型最主要的挑战之一是生成可用于各种三维相关任务（如三维描述生成、三维问答、三维锚定和三维目标识别等）的“三维—文本”配对数据集。与互联网上大量存在“图像—文本”配对数据相比，“三维—文本”配对数据的稀缺阻碍了城市三维场景多模态大模型的发展。鉴于语言大模型的卓越能力，可利用其生成“三维—文本”配对数据。例如，以研究区和地物在三维场景中的轴对齐边界框（AABB）作为输入，并提供有关场景的语义和空间位置信息。然后，向语言大模型提供具体的指令，以生成多样化的数据。或者，首先从三维场景的不同视角渲染若干幅图像，然后通过提示让ChatGPT提出关于图像的一系列有信息量的问题，再利用BLIP-2（Li等，2023）回答这些问题。最后，利用ChatGPT对所有这些描述进行总结，从而形成整个场景的全局三维描述。

构建城市三维场景多模态大模型的另一挑战在于如何提取能够与语言特征对齐的高鉴别性多尺度三维特征。一种方法是利用展开算法（unrolling algorithm）（Monga等，2021）对传统3D Mesh模型简化方法，如边塌缩算法、Graclus算法、QEM（quadric error metrics）算法等进行网络化，使其嵌入网络模型，以端到端提取多尺度特征，并从头开始训练三维编码器，使用类似对比学习等自监督范式来对齐二维图像和语言。或通过二维多视角图像构建三维特征。由于提取的三维特征被映射到与二维预训练特征相同的特征空间，因此可以无缝地使用二维VLLMs作为骨干网络，并输入三维特征，从而高效

地训练城市三维场景多模态大模型。

由于三维数据处理和大语言模型本身的复杂性，城市三维场景多模态大模型在计算效率方面同样面临巨大挑战。三维数据通常比二维数据更为复杂，涉及更多的几何信息和空间关系，这导致了更高的计算成本和存储需求。例如，三维点云、三维网格的处理，尤其是在进行实时处理或大规模数据集训练时，需要大量的计算资源。此外，城市三维场景多模态大模型通常具有庞大的参数量，这不仅增加了模型训练的时间和资源消耗，还可能导致过拟合和泛化能力下降。为提高计算效率，可研究神经网络压缩技术，如网络量化和结构剪枝等技术，去除冗余参数和简化模型结构，降低模型的复杂度和计算成本。此外，还需设计更加灵活和高效的大模型架构，根据输入数据的复杂度和任务需求动态调整其计算资源分配，从而在保证性能的同时降低计算成本。例如，通过分层和模块化的模型设计，让模型在不同层次上处理不同复杂度的数据，从而提高整体效率。提高计算效率的另一条研究途径是研究 Flash-Attention 等硬件优化策略，助力城市三维多模态大模型在移动边缘设备上的部署。

3）拓展语义化城市 3D Mesh 模型应用场景。语义化城市 3D Mesh 模型能够提供丰富的语义、空间和地物属性等信息，是对城市空间的精确映射，在城市规划、空间分析和环境分析等应用中起到至关重要的作用。然而，目前语义化城市 3D Mesh 模型在现实生活中的应用程度仍不是很高，需要进一步拓展语义化城市 3D Mesh 模型的应用场景。例如，将语义化城市 3D Mesh 模型应用于新兴领域，特别是低空经济领域。低空经济是以低空空域为依托，以通用航空产业为主导的经济活动，涉及无人机配送、空中观光、环境监测和农业植保等。这些经济活动对精确的三维地理信息和实时的空间感知能力都有很高的要求。语义化城市 3D Mesh 模型能够提供详细的建筑物、道路以及植被等元素的三维信息，帮助无人机和飞行器在复杂的低空环境中安全、高效地运行。例如，在无人机配送中，3D Mesh 模型可以用于路径规划和避障，确保无人机在高楼林立的城市中准确无误地完成任务。在空中观光中，3D Mesh 模型可以提供逼真的虚拟旅游体验，吸引更多的游客。

此外，语义化城市 3D Mesh 模型在低空经济中的应用还可以促进环境监测和灾害应对。通过高精度语义化城市 3D Mesh 模型，可以实时监测城市中的空气质量、温度分布等环境参数，及时发现和处理环境污染问题。在灾害应对中，语义化城市 3D Mesh 模型可以用于快速评估受灾区域的损毁情况，指导救援人员进行精准的救援行动。例如，在火灾或地震等自然灾害发生时，语义化城市 3D Mesh 模型可以提供详细的地形和建筑物信息，帮助救援队伍快速定位被困人员，制定最优的救援方案。

致谢：本文由中国图象图形学学会空间信息感知与决策专业委员会组织撰写，该专业委员会链接为 https://www.csig.org.cn/16/202309/51298.html。

参考文献（References）

Boulch A, Guerry J, Le Saux B and Audebert N. 2018. SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. Computers and Graphics, 71: 189-198 [DOI: 10.1016/j.cag.2017.11.010]

Breiman L. 1996. Bagging predictors. Machine Learning, 24(2): 123-140 [DOI: 10.1007/BF00058655]

Breiman L. 2001. Random forests. Machine Learning, 45(1): 5-32 [DOI: 10.1023/A:1010933404324]

Bronstein M M, Bruna J, LeCun Y, Szlam A and Vandergheynst P. 2017. Geometric deep learning: going beyond Euclidean data. IEEE Signal Processing Magazine, 34(4): 18-42 [DOI: 10.1109/MSP.2017.2693418]

Chatfield K, Simonyan K, Vedaldi A and Zisserman A. 2014. Return of the Devil in the details: delving deep into convolutional nets [EB/OL]. [2024-12-15]. https://arxiv.org/pdf/1405.3531.pdf

Chen D Y, Tian X P, Shen Y T and Ouhyoung M. 2003. On visual similarity based 3D model retrieval. Computer Graphics Forum, 22(3): 223-232 [DOI: 10.1111/1467-8659.00669]

Chen J Z, Xu Y H, Lu S F, Liang R H and Nan L L. 2022a. 3-D instance segmentation of MVS buildings. IEEE Transactions on Geoscience and Remote Sensing, 60: #5704014 [DOI: 10.1109/TGRS.2022.3183567]

Chen L C, Zhu Y K, Papandreou G, Schroff F and Adam H. 2018. Encoder-decoder with Atrous separable convolution for semantic image segmentation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 833-851 [DOI: 10.1007/978-3-030-01234-2_49]

Chen Y M and Feng M W. 2022b. Urban form simulation in 3D based on cellular automata and building objects generation. Building and Environment, 226: #109727 [DOI: 10.1016/j. buildenv. 2022.109727]

Chollet F. 2017. Xception: deep learning with depthwise separable con-

volutions//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1800-1807 [DOI: 10.1109/CVPR.2017.195]

Cohen-Steiner D, Alliez P and Desbrun M. 2004. Variational shape approximat //ACM SIGGRAPH 2004 Papers. Los Angeles, USA: ACM: 905-914 [DOI: 10.1145/1186562.1015817]

Cross G R and Jain A K. 1983. Markov random field texture models. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5(1): 25-39 [DOI: 10.1109/TPAMI.1983.4767341]

Dominik W, Bożyczko M and Tułacz-Maziarz K. 2021. Deep learning for automatic lidar point cloud processing. Archiwum Fotogrametrii, Kartografii i Teledetekcji, 33: 13-22 [DOI: 10.2478/apcrs-2021-0001]

Evtimova K and LeCun Y. 2022. Sparse coding with multi-layer decoders using variance regularizationv [EB/OL]. [2024-12-15]. https://arxiv.org/pdf/2112.09214.pdf

Gao L, Liu Y, Chen X, Liu Y X, Yan S and Zhang M J. 2024. CUS3D: a new comprehensive urban-scale semantic-segmentation 3D benchmark dataset. Remote Sensing, 16(6): #1079 [DOI: 10.3390/rs16061079]

Gao W X, Nan L L, Boom B and Ledoux H. 2021. SUM: a benchmark dataset of semantic urban meshes. ISPRS Journal of Photogrammetry and Remote Sensing, 179: 108-120 [DOI: 10.1016/j.isprsjprs.2021.07.008]

Gao W X, Nan L L, Boom B and Ledoux H. 2023. PSSNet: planarity-sensible semantic segmentation of large-scale urban meshes. ISPRS Journal of Photogrammetry and Remote Sensing, 196: 32-44 [DOI: 10.1016/j.isprsjprs.2022.12.020]

Garland M and Heckbert P S. 1997. Surface simplification using quadric error metrics//Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques. Los Angeles, USA: ACM Press: 209-216 [DOI: 10.1145/258734.258849]

George D, Xie X H and Tam G K. 2018. 3D mesh segmentation via multi-branch 1D convolutional neural networks. Graphical Models, 96: 1-10 [DOI: 10.1016/j.gmod.2018.01.001]

Gregor K and LeCun Y. 2010. Learning fast approximations of sparse coding//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel: Omnipress: 399-406

Grzeczkowicz G and Vallet B. 2022. Semantic segmentation of urban textured meshes through point sampling. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, V-2-2022: 177-184 [DOI: 10.5194/isprs-annals-V-2-2022-177-2022]

Guo Y L, Wang H Y, Hu Q Y, Liu H, Liu L and Bennamoun M. 2021. Deep learning for 3D point clouds: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(12): 4338-4364 [DOI: 10.1109/TPAMI.2020.3005434]

Hanocka R, Hertz A, Fish N, Giryes R, Fleishman S and Cohen-Or D. 2019. MeshCNN: a network with an edge. ACM Transactions on Graphics (TOG), 38(4): #90 [DOI: 10.1145/3306346.3322959]

Hao S J, Zhou Y and Guo Y R. 2020. A brief survey on semantic segmentation with deep learning. Neurocomputing, 406: 302-321 [DOI: 10.1016/j.neucom.2019.11.118]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hong Z H, Yang Y H, Liu J, Jiang S L, Pan H Y, Zhou R Y, Zhang Y, Han Y L, Wang J, Yang S H and Zhong C Y. 2022. Enhancing 3D reconstruction model by deep learning and its application in building damage assessment after earthquake. Applied Sciences, 12(19): #9790 [DOI: 10.3390/app12199790]

Hu W B, Zhao H S, Jiang L, Jia J Y and Wong T T. 2021. Bidirectional projection network for cross dimension scene understanding//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14368-14377 [DOI: 10.1109/cvpr46437.2021.01414]

Jaritz M, Gu J Y and Su H. 2019. Multi-view PointNet for 3D scene understanding//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea (South): IEEE: 3995-4003 [DOI: 10.1109/iccvw.2019.00494]

Katz D M, Bommarito M J, Gao S G and Arredondo P. 2024. GPT-4 passes the bar exam. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 382(2270): 1-17 [DOI: 10.1098/rsta.2023.0254]

Kazhdan M, Funkhouser T and Rusinkiewicz S. 2003. Rotation invariant spherical harmonic representation of 3D shape descriptors//Proceedings of 2003 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing. Aachen, Germany: Eurographics Association: 156-164

Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, Xiao T T, Whitehead S, Berg A C, Lo W Y, Dollár P and Girshick R. 2023. Segment anything//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3992-4003 [DOI: 10.1109/ICCV51070.2023.00371]

Knott M and Groenendijk R. 2021. Towards mesh-based deep learning for semantic segmentation in photogrammetry. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, V-2-2021: 59-66 [DOI: 10.5194/isprs-annals-V-2-2021-59-2021]

Kölle M, Laupheimer D, Schmohl S, Haala N, Rottensteiner F, Wegner J D and Ledoux H. 2021. The Hessigheim 3D (H3D) benchmark on semantic segmentation of high-resolution 3D point clouds and textured meshes from UAV LiDAR and multi-view-stereo. ISPRS Open Journal of Photogrammetry and Remote Sensing, 1: #100001 [DOI: 10.1016/j.ophoto.2021.100001]

Kundu A, Yin X Q, Fathi A, Ross D, Brewington B, Funkhouser T and Pantofaru C. 2020. Virtual multi-view fusion for 3D semantic

segmentation//Proceedings of the 16th European Conference on Computer Vision —— ECCV 2020. Glasgow, UK: Springer: 518-535 [DOI: 10.1007/978-3-030-58586-0_31]

Lafarge F and Mallet C. 2012. Creating large-scale city models from 3D-point clouds: a robust approach with hybrid representation. International Journal of Computer Vision, 99(1): 69-85 [DOI: 10.1007/s11263-012-0517-8]

Laupheimer D and Haala N. 2022. Multi-modal semantic mesh segmentation in urban scenes. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, V-2-2022: 267-274 [DOI: 10.5194/isprs-annals-V-2-2022-267-2022]

Laupheimer D, Shams Eddin M H and Haala N. 2020a. On the association of lidar point clouds and textured meshes for multi-modal semantic segmentation. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, V-2-2020: 509-516 [DOI: 10.5194/isprs-annals-V-2-2020-509-2020]

Laupheimer D, Shams Eddin M H and Haala N. 2020b. The importance of radiometric feature quality for semantic mesh segmentation//40. Wissenschaftlich-Technische Jahrestagung. Stuttgart: der DGPF, 29: 205-218

Lawin F J, Danelljan M, Tosteberg P, Bhat G, Khan F S and Felsberg M. 2017. Deep projective 3D semantic segmentation//Proceedings of the 17th International Conference on Computer Analysis of Images and Patterns. Ystad, Sweden: Springer: 95-107 [DOI: 10.1007/978-3-319-64689-3_8]

Lei H, Akhtar N and Mian A. 2021. Picasso: a CUDA-based library for deep learning over 3D meshes//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13849-13859 [DOI: 10.1109/cvpr46437.2021.01364]

Lei H, Akhtar N, Shah M and Mian A. 2024. Mesh convolution with continuous filters for 3-D surface parsing. IEEE Transactions on Neural Networks and Learning Systems, 35(10): 14863-14877 [DOI: 10.1109/TNNLS.2023.3281871]

Li J N, Li D X, Savarese S and Hoi S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 19730-19742

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin transformer: hierarchical vision transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/iccv48922.2021.00986]

Monga V, Li Y L and Eldar Y C. 2021. Algorithm unrolling: interpretable, efficient deep learning for signal and image processing. IEEE Signal Processing Magazine, 38(2): 18-44 [DOI: 10.1109/MSP.2020.3016905]

Mu T J, Shen M Y, Lai Y K and Hu S M. 2024. Learning virtual view selection for 3D scene semantic segmentation. IEEE Transactions on Image Processing, 33: 4159-4172 [DOI: 10.1109/TIP.2024.3421952]

Phong B T. 1998. Illumination for computer generated pictures//Seminal Graphics: Pioneering Efforts That Shaped the Field. New York, USA: ACM Press: 95-101 [DOI: 10.1145/280811.280980]

Qi C R, Su H, Mo K C and Guibas L J. 2017a. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 77-85 [DOI: 10.1109/CVPR.2017.16]

Qi C R, Yi L, Su H and Guibas L J. 2017b. PointNet++: deep hierarchical feature learning on point sets in a metric space [EB/OL]. [2024-12-15]. https://arxiv.org/pdf/1706.02413.pdf

Qiu Z J, Zhang L, Yao Y, Feng X Q and Gao J T. 2024. A survey on semantic segmentation in 3D point cloud scenes. Journal of Image and Graphics: 1-17 (仇志江, 张林, 姚垚, 冯小青, 高俊涛. 2024. 三维点云场景语义分割研究进展. 中国图象图形学报: 1-17) [DOI: 10.11834/jig.240650]

Riemenschneider H, Bódis-Szomorú A, Weissenberg J and Van Gool L. 2014. Learning where to classify in multi-view semantic segmentation//Proceedings of the 13th European Conference on Computer Vision-ECCV 2014. Zurich, Switzerland: Springer: 516-532 [DOI: 10.1007/978-3-319-10602-1_34]

Rong M Q and Shen S H. 2023. 3D semantic segmentation of aerial photogrammetry models based on orthographic projection. IEEE Transactions on Circuits and Systems for Video Technology, 33(12): 7425-7437 [DOI: 10.1109/TCSVT.2023.3273224]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Rouhani M, Lafarge F and Alliez P. 2017. Semantic segmentation of 3D textured meshes for urban scene analysis. ISPRS Journal of Photogrammetry and Remote Sensing, 123: 124-139 [DOI: 10.1016/j.isprsjprs.2016.12.001]

Sánchez J, Perronnin F, Mensink T and Verbeek J. 2013. Image classification with the fisher vector: theory and practice. International Journal of Computer Vision, 105(3): 222-245 [DOI: 10.1007/s11263-013-0636-x]

Skondras A, Karachaliou E, Tavantzis I, Tokas N, Valari E, Skalidi I, Bouvet G A and Stylianidis E. 2022. UAV mapping and 3D modeling as a tool for promotion and management of the urban space. Drones, 6(5): #115 [DOI: 10.3390/drones6050115]

Su H, Maji S, Kalogerakis E and Learned-Miller E. 2015. Multi-view convolutional neural networks for 3D shape recognition//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 945-953 [DOI: 10.1109/ICCV.2015.114]

Sutton C and McCallum A. 2012. An introduction to conditional random

fields. Foundations and Trends ® in Machine Learning, 4(4): 267-373 [DOI: 10.1561/2200000013]

Tang R K, Xia M J, Yang Y T and Zhang C. 2022. A deep-learning model for semantic segmentation of meshes from UAV oblique images. International Journal of Remote Sensing, 43(13): 4774-4792 [DOI: 10.1080/01431161.2022.2111665]

Thomas H, Qi C R, Deschaud J E, Marcotegui B, Goulette F and Guibas L. 2019. KPConv: flexible and deformable convolution for point clouds//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 6410-6419 [DOI: 10.1109/iccv.2019.00651]

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E and Lample G. 2023. Llama: open and efficient foundation language models [EB/OL]. [2024-12-15]. https://arxiv.org/pdf/2302.13971.pdf

Tutzauer P, Laupheimer D and Haala N. 2019. Semantic urban mesh enhancement utilizing a hybrid model. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-2/W7: 175-182 [DOI: 10.5194/isprs-annals-IV-2-W7-175-2019]

Ulku I and Akagündüz E. 2022. A survey on deep learning-based architectures for semantic segmentation on 2D images. Applied Artificial Intelligence, 36(1): #2032924 [DOI: 10.1080/08839514.2022.2032924]

Vedaldi A and Fulkerson B. 2010. Vlfeat: an open and portable library of computer vision algorithms//Proceedings of the 18th ACM International Conference on Multimedia. Firenze, Italy: ACM Press: 1469-1472 [DOI: 10.1145/1873951.1874249]

Wang H and Zhang J Y. 2022. A survey of deep learning-based mesh processing. Communications in Mathematics and Statistics, 10(1): 163-194 [DOI: 10.1007/s40304-021-00246-7]

Wang L, Huang Y C, Hou Y L, Zhang S M and Shan J. 2019a. Graph attention convolution for point cloud semantic segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 10288-10297 [DOI: 10.1109/CVPR.2019.01054]

Wang Y, Sun Y B, Liu Z W, Sarma S E, Bronstein M M and Solomon J M. 2019b. Dynamic graph CNN for learning on point clouds. ACM Transactions on Graphics (TOG), 38(5): #146 [DOI: 10.1145/3326362]

Wilk Ł, Mielczarek D, Ostrowski W, Dominik W and Krawczyk J. 2022. Semantic urban mesh segmentation based on aerial oblique images and point clouds using deep learning. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B2-2022: 485-491 [DOI: 10.5194/isprs-archives-XLIII-B2-2022-485-2022]

Wu Z R, Song S R, Khosla A, Yu F, Zhang L G, Tang X O and Xiao J X. 2015. 3D ShapeNets: a deep representation for volumetric shapes//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1912-1920 [DOI: 10.1109/CVPR.2015.7298801]

Xiao Y P, Lai Y K, Zhang F L, Li C P and Gao L. 2020. A survey on deep geometry learning: from a representation perspective. Computational Visual Media, 6(2): 113-133 [DOI: 10.48550/arXiv.2002.07995]

Yang G Q Xue F Y, Zhang Q, Xie K, Fu C W and Huang H. 2023a. UrbanBIS: a large-scale benchmark for fine-grained urban building instance segmentation//Proceedings of ACM SIGGRAPH 2023 Conference Proceedings. Los Angeles, USA: ACM Press: #16 [DOI: 10.1145/3588432.3591508]

Yang Y T, Tang R K, Xia M J and Zhang C. 2023b. A surface graph based deep learning framework for large-scale urban mesh semantic segmentation. International Journal of Applied Earth Observation and Geoinformation, 119: #103322 [DOI: 10.1016/j.jag.2023.103322]

Yang Y T, Tang R K, Xia M J and Zhang C. 2023c. A texture integrated deep neural network for semantic segmentation of urban meshes. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16: 4670- 4684 [DOI: 10.1109/JSTARS.2023.3276977]

Yu L and Wu X Q. 2024. Survey of texture optimization algorithms for 3D reconstructed scenes. Journal of Image and Graphics, 29(8): 2303-2318 (于柳, 吴晓群. 2024. 三维重建场景的纹理优化算法综述. 中国图象图形学报, 29(8): 2303-2318) [DOI: 10.11834/jig.230478]

Yu Y, Wang C P, Fu Q, Kou R K, Huang F Y, Yang B X, Yang T T and Gao M L. 2023. Techniques and challenges of image segmentation: a review. Electronics, 12(5): 1199 [DOI: 10.3390/electronics12051199]

Zhang G Y and Zhang R T. 2023. MeshNet-SP: a semantic urban 3D mesh segmentation network with sparse prior. Remote Sensing, 15(22): #5324 [DOI: 10.3390/rs15225324]

Zhang G Y, Zhang R T, Dai Q H, Chen J and Pan Y P. 2021. The direction of integration surveying and mapping geographic information and artificial intelligence 2.0. Acta Geodaetica et Cartographica Sinica, 50(8): 1096-1108 (张广运, 张荣庭, 戴琼海, 陈军, 潘云鹤. 2021. 测绘地理信息与人工智能2.0融合发展的方向. 测绘学报, 50(8): 1096-1108) [DOI: 10.11947/j.AGCS.2021.20210200]

Zhang J Y, Zhao X L, Chen Z and Lu Z J. 2019. A review of deep learning-based semantic segmentation for point cloud. IEEE Access, 7: 179118-179133 [DOI: 10.1109/ACCESS.2019.2958671]

Zhang L, Liu Y X, Sun Y J, Lan C Z, Ai H B and Fan Z L. 2022. A review of developments in the theory and technology of three-dimensional reconstruction in digital aerial photogrammetry. Acta Geodaetica et Cartographica Sinica, 51(7): 1437-1457 (张力, 刘

玉轩，孙洋杰，蓝朝桢，艾海滨，樊仲藜. 2022. 数字航空摄影三维重建理论与技术发展综述. 测绘学报，51(7)：1437-1457) [DOI：10.11947/j.AGCS.2022.20220130]

Zhang N，Pan Z Y，Li T H，Gao W and Li G. 2023a. Improving graph representation for point cloud segmentation via attentive filtering// Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：1244-1254 [DOI：10.1109/CVPR52729.2023.00126]

Zhang R T，Zhang G Y and Yin J H. 2023. Semantic segmentation method of 3D scenes using dynamic graph CNN for complex city. Acta Geodaetica et Cartographica Sinica，52(10)：1703-1713（张荣庭，张广运，尹继豪. 2023. 复杂城市动态图卷积网络三维场景语义分割法. 测绘学报，52(10)：1703-1713) [DOI：10.11947/j.AGCS.2023.20220466]

Zhang R T，Zhang G Y，Yin J H，Jia X P and Mian A. 2023b. Mesh-based DGCNN：semantic segmentation of textured 3-D urban scenes. IEEE Transactions on Geoscience and Remote Sensing，61：#4402812 [DOI：10.1109/TGRS.2023.3266273]

Zhao H H，Shi J P，Qi X J，Wang X G and Jia J Y. 2017. Pyramid scene parsing network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，USA：IEEE：2881-2890 [DOI：10.1109/CVPR.2017.660]

Zhu Q，Zhang L G，Ding Y L，Hu H，Ge X M，Liu M W and Wang W. 2022. From real 3D modeling to digital twin modeling. Acta Geodaetica et Cartographica Sinica，51(6)：1040-1049（朱庆，张利国，丁雨淋，胡翰，葛旭明，刘铭崴，王玮. 2022. 从实景三维建模到数字孪生建模. 测绘学报，51(6)：1040-1049) [DOI：10.11947/j.AGCS.2022.20210640]

作者简介

张广运，男，教授，主要研究方向为遥感三维智能生成与解译。E-mail：zgy@njtech.edu.cn

张荣庭，通信作者，男，讲师，主要研究方向为遥感三维智能生成与解译。E-mail：zrt@njtech.edu.cn

张余，男，副教授，主要研究方向为遥感图像处理。E-mail：useful@163.com

王麒雄，男，博士研究生，主要研究方向为遥感图像处理。E-mail：wqx1140@buaa.edu.cn

冯家齐，男，博士研究生，主要研究方向为遥感图像处理。E-mail：fengjiaqi@buaa.edu.cn

姜鸿翔，男，博士研究生，主要研究方向为遥感图像处理。E-mail：jianghongxiang@buaa.edu.cn