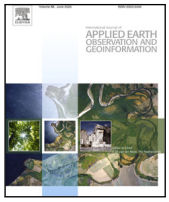
International Journal of Applied Earth Observation and Geoinformation 121 (2023) 103365

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Deep learning-based semantic segmentation of urban-scale 3D meshes in remote sensing: A survey

Jibril Muhammad Adam [a,b], Weiquan Liu [a,*], Zang Yu [a,*], Muhammad Kamran Afzal [a], Saifullahi Aminu Bello [a], Abdullahi Uwaisu Muhammad [b], Cheng Wang [a], Jonathan Li [a,c]

[a] Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen 361005, China
[b] Faculty of Computing, Department of Computer Science, Federal University Dutse, Nigeria
[c] Department of Geography and Environmental Management, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

ARTICLE INFO

Keywords:
Semantic segmentation
Deep learning
3D meshes
Urban-scale
Survey
Remote sensing

ABSTRACT

Semantic segmentation in 3D meshes is the classification of its constituent element(s) into specific classes or categories. Using the powerful feature extraction abilities of deep neural networks (DNNs), significant results have been obtained in the semantic segmentation of various remotely sensed data formats. With the increased utilization of DNNs to segment remotely sensed data, there have been commensurate in-depth reviews and surveys summarizing the various learning-based techniques and methodologies that entail these methods. However, most of these surveys focused on methods that involve popular data formats like LiDAR point clouds, synthetic aperture radar (SAR) images, and hyperspectral images (HSI) while 3D meshes hardly received any attention. In this paper, to our best knowledge, we present the first comprehensive and contemporary survey of recent advances in utilizing deep learning techniques for the semantic segmentation of urban-scale 3D meshes. We first describe the different approaches employed by mesh-based learning methods to generalize and implement learning techniques on the mesh surface, and then describe how the element-wise classification tasks are achieved through these methods. We also provide an in-depth discussion and comparative analysis of the surveyed methods followed by a summary of the benchmark large-scale mesh datasets accompanied with the evaluation metrics for assessing the segmentation performance of the methods. Finally, we summarize some of the contemporary problems of the field and provide future research directions that may help researchers in the community.

1. Introduction

The increasing availability of 3D remotely sensed data has proliferated research in several 3D vision tasks like autonomous driving (Kang et al., 2021), semantic segmentation (Weixiao et al., 2023), road extraction and city planning (Chen et al., 2022b) e.t.c. These data come in different formats such as point cloud, RGB-D images, and 3D meshes. While point cloud is the most popular format, 3D meshes are better at explicitly representing the geometry of scenes hence their rising popularity in depicting real-life scenes in remote sensing. The 3D meshes considered in this survey comprise both large- and city-scale reconstructed surfaces from a point cloud or other remotely sensed data.

A 3D mesh comprises three elements (Liu et al., 2023) namely: vertices, edges, and faces. Semantic segmentation in 3D meshes entails element-wise classification using different techniques and learning architectures. In recent years, deep neural network (DNN) architectures have replaced traditional (Liu et al., 2015) and machine learning (Rouhani et al., 2017) methods as the dominant technique for achieving semantic segmentation in both 2D (Ulku and Akagündüz, 2022) and 3D (Gao et al., 2021a) data formats. This is due to the powerful ability of DNNs to extract rich features from the considered data formats. With increasing processing power from advances in graphics processing units (GPUs) and the availability of specialized and labeled datasets, significant improvement in semantic segmentation results has been obtained by methods. These advances have also engendered progress in semantic segmentation-related tasks like instance segmentation (Chen et al., 2022a; Sharma et al., 2022) and object detection (Liang et al., 2021).

With increased interest in semantic segmentation using DNNs, survey papers have been published to detail the progress made in 2D (Ulku

* Corresponding authors.
E-mail addresses: jibrilmuhammadadam@gmail.com (J.M. Adam), wqliu@xmu.edu.cn (W. Liu), zangyu7@126.com (Y. Zang), m.kamran.afzal@live.com (M.K. Afzal), saifullahiabel@gmail.com (S.A. Bello), uwaisabdullahi87@gmail.com (A.U. Muhammad), cwang@xmu.edu.cn (C. Wang), junli@uwaterloo.ca (J. Li).

https://doi.org/10.1016/j.jag.2023.103365
Received 16 February 2023; Received in revised form 5 May 2023; Accepted 18 May 2023
Available online 2 June 2023

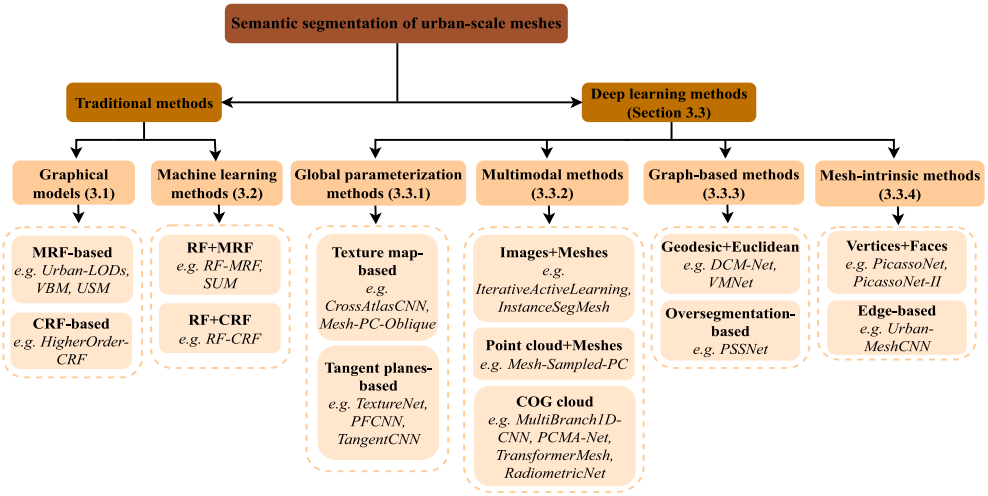


Fig. 1. A taxonomy of the different approaches of interpreting mesh surfaces for definition of deep learning techniques.

Fig. 2. A word cloud of paper titles surveyed by this paper.

and Akagündüz, 2022), 3D (Gao et al., 2021a), point cloud (Guo et al., 2021), unmanned aerial vehicle-images (UAV) (Osco et al., 2021), and precise boundary recovery of boundaries in segmented images and point cloud (Zhang et al., 2021). However, the successes recorded by DNNs in processing images and point clouds are not commensurate with their impact on mesh processing as evidenced by the almost nonexistence of survey papers that focus on DNN-based mesh processing. Wang and Zhang (2022) is the only existing mesh-based survey we found that addressed this challenge, albeit focusing on methods that consume small-scale or toy datasets and hence the need for a comprehensive survey of remote sensing-based mesh neural networks.

Even with the late implementation of DNNs to process 3D meshes in comparison to other data formats, research in the area has nevertheless blossomed. However, as mentioned earlier, methods that consume urban- and large-scale mesh scenes are mostly neglected, and with the significance of semantically processed 3D scenes in fields like autonomous driving (Kang et al., 2021; Mao et al., 2022; Peng et al., 2022) and city modeling (Lei et al., 2022), we believe there is a literature gap that relates to the semantic perception of **urban-scale 3D meshes** using **DNNs** in remote sensing. Specifically, the use of 3D meshes in urban scene perception and visualization is becoming more popular and their semantic segmentation is useful in many areas of our lives such as in mixed reality/augmented reality social platforms (Du et al., 2019), addressing challenges in smart cities (An, 2023) and, geospatial evaluation and visualization of sectors that affect our everyday lives like education (Wu et al., 2021). To our best knowledge, this survey is the first paper to focus on how various DNN architectures achieve the task of semantic segmentation in urban-scale meshes. To illustrate the significance of this work, we generated a word cloud (Fig. 2) from the titles of the papers we surveyed, and certain words like **semantic segmentation**, **mesh**, **3d**, **meshes**, **urban**, and **learning** can be seen to appear more frequently due to their larger sizes which subsequently points to their importance in remote sensing. Therefore, in this paper, we present the different methods that combine these terms to address the problem of semantic segmentation in remote sensing.

The contributions of this paper are outlined below:

- To the best of our knowledge, this is the first survey paper to comprehensively review the deep learning methods for semantic segmentation of urban-scale 3D meshes.
- This paper provides a classification of the learning methods for semantic segmentation of urban scenes based on the different techniques of generalizing deep learning operations on mesh surfaces.
- The paper provides descriptions of urban- and large-scale mesh datasets together with evaluation metrics for assessing the semantic segmentation task.

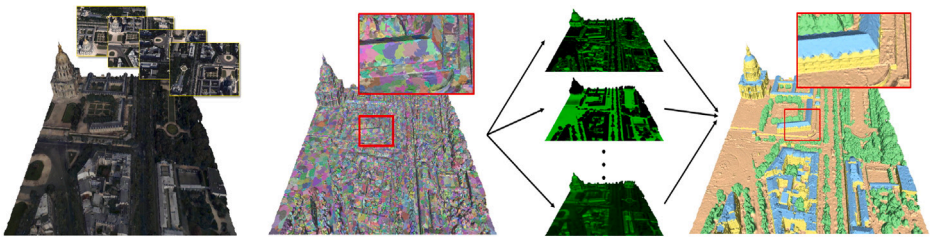


Fig. 3. An illustration of the classification of superfacets in RF-MRF (Rouhani et al., 2017) where urban scale meshes **(a)** are segmented into homogeneous superfacets **(b)** using a region growing algorithm (Cohen-Steiner et al., 2004). The classification and refinement of the superfacets into their appropriate classes is done using an RF-MRF **((d)** and **(c))** framework. A similar approach is followed by the modeling and reconstruction methods i.e. Urban-LODs (Verdie et al., 2015), VBM (Zhu et al., 2017) and USM (Zhu et al., 2018) with abstraction and reconstruction of urban artifacts from the superfacets replacing the RF **(d)** classification step in RF-MRF.
Source: Figure taken from RF-MRF (Rouhani et al., 2017).

- The paper presents the open challenges facing researchers in the field and proffer future research directions for researchers in the community.

This paper is organized as follows: We give a formulation for the semantic segmentation problem together with the inherent challenges of processing mesh scenes in Section 2. In Section 3, we present the different categorizations of mesh-based DNNs for semantic segmentation of urban scenes (shown in Fig. 1) followed by descriptions of benchmark datasets and evaluation metrics for assessing the methods in Section 4. Finally, contemporary challenges that still beset the field are presented in Section 5 followed by our concluding remarks in Section 6.

2. Overview and background concepts

In this section, we define the two terms *i.e.* 3D meshes and semantic segmentation that are most relevant to the theme of this paper.

Definition 1 (*3D Mesh*). Mathematically, a 3D mesh $\mathcal{M} = \{\mathcal{V}, \mathcal{F}, \mathcal{E}\}$, *is a set of unordered vertices* $\mathcal{V} = \{1, \ldots, n\}$ *where* $v_i \in \mathbb{R}^3$, *and a set of contiguous triangular faces* $\mathcal{F} \subset \mathcal{V} \times \mathcal{V} \times \mathcal{V}$ *that represent the topology and geometry of a remotely sensed scene.* $\mathcal{E}$ *is a set of undirected edges* $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ *s.t.* $(v_i, v_j) \in \mathcal{E}$ *iff* $(v_j, v_i) \in \mathcal{E}$ *that encodes information between adjacent vertices and subsequently faces.* Edges also delineate boundaries in scenes and objects on the 3D mesh. As mentioned earlier, the composition of these three elements i.e. *vertices, faces,* and *edges* enables 3D meshes to depict the geometry of real-life large-scale urban scenes.

Definition 2 (*Semantic Segmentation*).Given $e \in \{\mathcal{V}, \mathcal{F}, \mathcal{E}\}$ *with* $e = \{e_1, e_2, e_3 \ldots e_n\}$ *as a set of element(s) from an input 3D mesh* , $\mathcal{M}$ *that serves as input to a neural network* $\mathcal{N}$. *Semantic segmentation is achieved by assigning a label,* l_i *from a set of possible semantic labels,* $\mathcal{L} = \{l_1, l_2, l_3 \ldots l_m\}$ *to every element,* e_i *of the* n *elements of* $\mathcal{M}$ *via the neural network,* $\mathcal{N}$.

Put simply, the above definition denotes the element-wise classification of the chosen element of an input 3D mesh by a neural network. As opposed to the singularity of pixels in images and points in point clouds as processing units of choice, 3D meshes have three candidates to choose from hence the diversity in both the input and output of mesh-based DNNs. For instance, PSSNet (Weixiao et al., 2023) consumes facial features as input while vertices are chosen in DCM-Net (Schult et al., 2020).

Adding to the challenge of choosing an input element that best suits the classification task, mesh-based DNNs must devise a technique that can process the unstructured non-Euclidean format of the 3D mesh scene. This is in contrast to the use of standard convolutional neural networks (CNNs) in image processing due to the structured nature of 2D data formats. These networks must also implement learning operations like pooling and feature aggregation that are 3D mesh-compliant. Most importantly, the networks must be designed to process large-scale urban mesh scenes to address the computational cost incurred.

Notwithstanding the aforementioned challenges of processing mesh scenes using DNNs, methods are developed to achieve various tasks (Wang and Zhang, 2022). In this paper, we focus on the methods that consume large-scale urban meshes to achieve the task of semantic segmentation. We categorize these methods based on the different techniques they employed in exploiting the 3D mesh surface for learning-based processing. These categorizations are shown in Fig. 1.

3. Semantic segmentation of large-scale urban 3D meshes

As shown in Fig. 1, various techniques are utilized by mesh neural networks for semantic segmentation tasks. In this section, we provide a detailed review of these networks starting from a brief overview of traditional methods (graphical models and forest-based machine learning methods) (Sections 3.1 and 3.2) to the various approaches that are utilized by deep learning methods (Section 3.3). We also provide a summary and highlight the salient parts of the methods in this section in Table 1.

3.1. Graphical models

The earliest methods of labeling large-scale urban meshes adopted graphical methods specifically markov random field (MRF) and its variant conditional random fields (Lafferty et al., 2001; Yu and Fan, 2020) (CRFs). This is because semantic segmentation requires locality information to enhance fine-grained element-wise classification and neighborhood contextual information to encode the relationship between elements. This information is encoded in unary (per-element) and pairwise (per-edge: the relationship between elements) terms by MRFs to express the dependencies between objects and their constituent elements in urban scenes.

The main goal of these methods is automated reconstruction and modeling of urban scenes as opposed to element-wise classification (Section 2). Most of these methods follow a three-step process (Fig. 3) of modeling scenes by first oversegmenting elements into homogeneous segments like superfacets usually using unsupervised clustering-based algorithms like the region-growing method in Urban-LODs (Verdie et al., 2015), variational shape approximation (Cohen-Steiner et al., 2004) (VSA) to extract planar structures in VBM (Zhu et al., 2017) and orthogonal grids sampled from a 2D image projection of the input mesh in USM (Zhu et al., 2018). Following the oversegmentation step is an MRF-based classification of the extracted segments into different classes like roofs, trees, ground, and roads. The final step includes the abstraction and reconstruction of the labeled segments into regularized polygons i.e. levels of detail (LODs) representing various city artifacts that usually conform to a standard like CityGML (Kolbe et al., 2005).

CRFs (Lafferty et al., 2001) are a variant of MRFs but they do not suffer from the problem of label bias inherent in the latter and their ability to express conditional dependence between elements (represented as nodes of a graph) on the mesh scene can be exploited

Table 1
Overview of semantic segmentation methods for urban-scale 3D meshes. Important characteristics for the methods are shown, which comprise the year of publication, category, the structure of modules that make up a method, dataset(s) used, labeled element, strengths and weaknesses. ML, GM, GP, MM and, G denote machine learning, graphical model, global parameterization, multimodal and graph-based respectively.

Method	Type	Labeled element	Network structure	Dataset(s)	Highlight(s)/Strength(s)	Limitation(s)
RF-CRF (Riemenschneider et al., 2014)	ML	Face	RF (labeling) + CRF (enforcing spatial consistency and label smoothing)	ETHZ RueMonge 2014	The first use of IoU (PASCAL IOU) for evaluating segmentation task; Introduced the ETHZ RueMonge 2014 (Full428 and Sub28) dataset.	Labeling accuracy depends depends on the resolution of reconstructed meshes.
Urban-LODs (Verdie et al., 2015)	GM	Face	Superfacet clustering + MRF (classification) + LOD abstraction (planar proxies, iconization and LOD generation) + Reconstruction	Proprietary	Additional semantic rules were used to tackle errors that are not sufficiently addressed by the MRF labeling technique	More of a modeling method than a semantic segmentation one
HigherOrder-CRF (Liu et al., 2015)	GM	Face	Lower- + Higher-order CRF (structural labeling/segmentation)	Herz-Jesu-P8 and five reconstructed large-scale scenes	Expression of contextual regularities in urban scenes via higher-order potential of CRFs	Extraction of structural regularities using exact subgraph isomorphism method is computationally expensive for urban-scale meshes.
RF-MRF (Rouhani et al., 2017)	ML	Face	Superfacet clustering + Feature extraction (geometric and photometric) + RF (labeling) + MRF (label smoothing)	Proprietary (reconstructed urban scenes of Paris and Toulouse, France)	Landmark method in semantic segmentation of urban-scale meshes; Introduced joint labeling to handle the transition between neighboring regions	Non-differerentiable clustering algorithm that inhibits end-to-end training
VBM (Zhu et al., 2017)	GM	Face	Region (plane proxy) clustering (VSA) + Regularization and contour extraction (MRF) + Modeling and LOD generation	Proprietary (reconstructed buildings from urban images)	Incorporated prior shape knowledge i.e. height and direction and introduced a directional weighting mechanism to improve the segmentation process and reduce its sensitivity to noise.	More of a building modeling method than a semantic segmentation one
USM (Zhu et al., 2018)	GM	Face	Semantic segmentation (MRF-based classification of generated orthophotos)+ Building modeling (regularization and LOD generation)	Proprietary	Double use of MRF formulation in both segmentation and modeling steps; Similar to global parameterization methods (Section 3.3.1), the segmentation is done on orthophotos that are generated from the mesh scenes.	More of a building modeling method than a semantic segmentation one
TangentCNN (Tatarchenko et al., 2018)	GP	Vertex	Fully convolutional UNet (encoder–decoder) with skip connections	ScanNet v2	A generic method that can be used on any 3D data format that supports surface normal estimation; Pooling is implemented by hashing points/vertices onto a regular 3D grid.	Information loss due to parameterization of points/vertices to tangent images
TextureNet (Huang et al., 2019)	GP	Vertex	UNet with skip connections	ScanNet v2 and Matterport3D	Furthest point sampling and nearest neighbor interpolation used for downsampling and upsampling respectively.	Generation of 4-RoSy using QuadriFlow still induces distortion albeit less than other methods

(*continued on next page*)

Table 1 (*continued*).

Method	Type	Labeled element	Network structure	Dataset(s)	Highlight(s)/Strength(s)	Limitation(s)
CrossAtlasCNN (Li et al., 2019)	GP	Face	FCN with VGG19	ETHZ RueMonge2014	The only (in this survey) DNN-based method that used ETHZ RueMonge2014 dataset; Cross-atlas pooling that replaced pixel neighborhood (2D) with the geodesic neighborhood on the mesh surface.	Information loss due to the distortion induced by the parameterization process;
MultiBranch1D-CNN (George et al., 2018)	MM	Face	Multi-branch 1D CNN	H3D	Employed an MRF formulation for explicit label smoothing; Showed how CNN-based training and inference are faster than RF-based	Required MRF for explicit refinement of segmentation results
PFCNN (Yang et al., 2020)	GP	Vertex	UNet with skip connections	ScanNet v2	Best performing tangent plane-based method mainly due to its translation equivariant convolution; Pooling and unpooling operations adapted to map the N-directions of frames	Requires complex mechanisms like computation of frame fields and parallel transport
PCMA-Net (Laupheimer et al., 2020b)	MM	Face	PointNet++	H3D	The PCMA explicitly associates faces on meshes with points in a point cloud; The PCMA transfer mechanism served as a foundation technique for another COG cloud-based method i.e. MultiModal-Net	Discrepancy between the mesh and point cloud affects the association rate between the data formats and labeling of non-associated points.
RadiometricNet (Laupheimer et al., 2020a)	MM	Face	RF, PointNet and PointNet++	H3D	An experiment-heavy method to evaluate radiometric features using RF, PointNet, and PointNet++ methods; Hierarchical feature learning ability of PointNet++ gave it an edge over the other methods.	The classifiers that were used for the comparative analysis are point cloud-based methods.
DCM-Net (Schult et al., 2020)	G	Vertex	UNet (encoder–decoder) with skip connections	S3DIS, ScanNet v2 and Matterport3D	The first convolution is translation-invariant which enriches context information and reduces computation time; It skips recalculation of feature space neighborhood of DGCNN to enable deeper GCNs and shorter computational time	Mesh simplification method not GPU-compliant and hence inhibits end-to-end training
SUM (Gao et al., 2021b)	ML	Face	Superfacet clustering + RF	SUM	Introduced the SUM dataset; Similar to RF-MRF without the MRF component	Non-differentiable oversegmentation method
IterativeActive-Learning (Rong et al., 2021)	MM	Face	Finetuning 2D segmentation network (DeepLabv3+) + 2D-3D semantic fusion (back-projection and MRF)	Proprietary (two reconstructed urban scenesi.e. Urban1 and Urban2)	Introduced geometric constraints to ensure labeling consistency in the 3D mesh after the back-projection phase	Manual annotation of images for finetuning the segmentation network is limited, class-wise and quality-wise

(*continued on next page*)

Table 1 (*continued*).

Method	Type	Labeled element	Network structure	Dataset(s)	Highlight(s)/Strength(s)	Limitation(s)
VMNet (Hu et al., 2021)	G	Vertex	UNet with skip connections	ScanNet v2 and Matterport3D	Even with a significantly larger number of parameters, training VMNet is computationally more memory-efficient than DCM-Net	Mesh simplification method not GPU-compliant and hence inhibits end-to-end training
PicassoNet (Lei et al., 2021a)	M	Vertex	UNet	S3DIS	Introduced three mesh-intrinsic convolutions: vertex2facet, facet2facet, and facet2vertex; Presented Picasso library in Tensorflow; End-to-end trainable	Only consume vertex/point coordinates and colors as input features.
PicassoNet-II (Lei et al., 2021b)	M	Vertex	UNet with skip connections	S3DIS and ScanNet v2	Consume mesh-based geometric features as input; Pytorch implementation of Picasso library presented; End-to-end trainable	Not robust to meshes that are not edge-manifold
Urban-MeshCNN (Knott and Groenendijk, 2021)	M	Edge	MeshCNN	V3D	Added photometric (edge-based HSV color) features to the initial geometric features of MeshCNN; Screened Poisson surface reconstruction used to repair the non-manifold artifacts on the reconstructed mesh scenes	Not robust to non-manifold meshes; Uniformity of triangles in generated mesh from Poisson reconstruction method affects learning ability of the method
Mesh-PC-Oblique (Wilk et al., 2022)	GP	Face	PSP-Net (image) and a proprietary FCN (point cloud)	SUM and a proprietary Bordeaux (France) dataset	Involves three data formats i.e. images, point clouds and 3D meshes	Semantic segmentation done in point cloud and image data formats
Mesh-Sampled-PC (Grzeczkowicz and Vallet, 2022)	MM	Face	KPConv	SUM	Obtained the best result on the SUM dataset; Sampled the mesh using two methods: texel and Poisson disk sampling	Information (especially textural) loss due to point sampling of the mesh; Used a point-specific method to compute semantic segmentation results
InstanceSegMesh (Chen et al., 2022a)	MM	Face	2D roof instance segmentation (Swin transformer) + Clustering of instance masks + Back-projection of clustered masks to 3D and subsequent segmentation (MRF)	InstanceBuilding (InstanceSegMesh)	Introduced the first MVS-based instance segmentation dataset i.e. InstanceBuilding; Adapted 2D instance segmentation metrics AP, AP50, and AP75 to assess the performance of instance segmentation in 3D meshes.	Final 3D instance segmentation performance heavily relies on the quality of the 2D roof instance segmentation
MultiModal-Net (Laupheimer and Haala, 2022)	MM	Face	RF	H3D and V3D	An experiment-heavy method to evaluate multimodal features generated using the PCMA technique; Can also be classified as a machine learning method (Section 3.2)	Discrepancy between the mesh and point cloud affects the association rate between the data formats and labeling of non-associated points.

(*continued on next page*)

Table 1 (*continued*).

Method	Type	Labeled element	Network structure	Dataset(s)	Highlight(s)/Strength(s)	Limitation(s)
TransformerMesh (Tang et al., 2022)	MM	Face	Hierarchical network in the spirit of PointNet++ with transformers, neighbor embedding, and feature propagation modules	SUM and Wuhan (proprietary)	Transformers are used to express contextual information between points in the COG cloud.	COG-based representation does not represent textural features of the mesh adequately.
PSSNet (Weixiao et al., 2023)	G	Face	Superfacet clustering (MRF and RF) + Classification of superfacets (PointNet and Gated graph sequence)	SUM and H3D	Introduced mesh-based evaluation metrics for assessing oversegmentation task; Follows similar point-based methods SPG (Landrieu and Simonovsky, 2018) and SSP (Landrieu and Boussaha, 2019).	Non-differentiable oversegmentation method

to extract and subsequently merge similar elements iteratively. This technique is implemented by HigherOrder-CRF (Liu et al., 2015) to initially extract photometric and geometric features using lower-order (unary and pairwise terms) potentials and higher-order potentials to express structural regularities constraints in urban scenes. Using subgraph matching, the scene is iteratively segmented into its structural components.

Notwithstanding the ability of graphical models to express contextual information, the accompanying challenges of adopting them for semantic segmentation have made researchers look for other alternatives. Some of these challenges include the computational inefficiencies involved in executing these models like the discovery of all possible graphs for the graph matching method in HigherOrder-CRF (Liu et al., 2015). The density of elements in urban scenes also exacerbates this problem. Most graphical models also use handcrafted features which inhibits their discriminative ability and hence the pivot of researchers to machine and deep learning methods.

3.2. Machine learning methods

In an attempt to leverage graphical models' ability to express contextual information and to counteract some of their weaknesses like imprecise predictions due to a large number of possible semantic classes during inference, inefficient methods of fine-tuning parameters, and the low discriminative power of their handcrafted features, some methods use random forests (RFs) within a graphical model framework for labeling and boundary refinement tasks respectively.

The major work in this category is RF-MRF (Rouhani et al., 2017), wherein the authors used a supervised random forest classifier to predict the labels of superfacets that were generated using a region-growing clustering algorithm (Cohen-Steiner et al., 2004) (Fig. 3). Improving on existing graphical models (Verdie et al., 2015), geometric and photometric features are extracted and concatenated per superfacet as opposed to the prevalent use of the former in most graphical models. Using a joint label space denoting the classes of superfacets and their adjacent neighbors, randomized decision trees are trained to predict the joint labels. The method used an MRF framework (energy minimization) to refine the predicted probabilities of labels thereby enforcing spatial coherence between superfacets and contextual clarity between class boundaries. Superfacet-based classification using randomized decision trees is also used in labeling mesh faces in the development of the SUM (Gao et al., 2021b) (Fig. 9) benchmark dataset of urban-scale meshes. The superfacets were generated using a region-growing algorithm (Lafarge and Mallet, 2012) to group similar faces into homogeneous clusters. A manual refinement method is chosen by the authors to refine the predicted labels as opposed to the MRF formulation in RF-MRF (Rouhani et al., 2017). Moving away from superfacet-based classification, face-based classification is used for semantic segmentation of reconstructed multi-view stereo (MVS) 3D meshes of urban scenes in RF-CRF (Riemenschneider et al., 2014). An RF-CRF formulation is used to label faces and enforce spatial connectivity between adjacent faces. The authors also presented the ETHZ RueMonge (Riemenschneider et al., 2014) dataset that can be used for image and mesh labeling tasks for urban scene understanding.

The advantages of using RF classifiers for labeling urban meshes over graphical models include the ability to support multiple classes, more tractable computational time due to a manageable number of parameters, and the use of more discriminative and non-linear features. Another significant improvement brought about by RF-based methods is the development and utilization of benchmark datasets for urban meshes like SUM (Gao et al., 2021b) and ETHZ RueMonge (Riemenschneider et al., 2014) which subsequently facilitated the use of supervised learning (Riemenschneider et al., 2014; Gao et al., 2021b; Rouhani et al., 2017) approaches and standard evaluation metrics of evaluating semantic segmentation tasks like PASCAL intersection over union (IoU) (Riemenschneider et al., 2014), mean class Accuracy (mAcc) and Overall Accuracy (OA) (Gao et al., 2021b).

Notwithstanding the aforementioned advantages of RF-based methods, they are still beset by challenges such as the use of handcrafted features and non-differentiable clustering algorithms e.g. the region-growing algorithm that was used to generate superfacets in RF-MRF (Rouhani et al., 2017) (hard association between faces and superfacets) which prohibits end-to-end learning end-to-end. With the increasing availability of labeled benchmark datasets and the need for end-to-end semantic segmentation methods, researchers in the field are increasingly developing deep learning methods to address the aforementioned challenges of both graphical and RF-based methods and the subsequent semantic segmentation of large-scale urban meshes.

3.3. Deep learning methods

Deep learning is a sub-field of machine learning that has recently facilitated a lot of progress in high-level vision tasks like semantic segmentation in various data formats. This is due to the task-driven feature extraction abilities of deep learning methods that are trained in an end-to-end manner e.g. learning facial features from 3D meshes for face-wise classification (face-based semantic segmentation). Deep learning methods are getting better at processing small- and large-scale datasets for semantic segmentation tasks (Ulku and Akagündüz, 2022; Lateef and Ruichek, 2019; Zhang et al., 2021) as methods continue to obtain more accurate results.

In this section, we review the different approaches employed by deep learning methods for the semantic segmentation of large-scale urban meshes. We classify these methods into appropriate categories (Fig. 1) with detailed descriptions of how they achieve the element-wise classification of urban meshes.

3.3.1. Global parameterization methods

Early deep learning methods leveraged existing 2D semantic segmentation networks by establishing a one-to-one correspondence between a 3D mesh surface and a 2D domain. This process is called parameterization and the generated 2D domain is the parameterized space (Ray et al., 2006). By one-to-one (bijectivity), the parameterization process should be able to project the geometric, textural, angular, and geodesic neighborhood properties of the 3D mesh surface to the 2D domain as accurately as possible. By parameterizing 3D scenes to a corresponding 2D parameter space, 2D-based semantic segmentation architectures are used by methods for element-wise classification. This is because 2D architectures are more mature (Ulku and Akagündüz, 2022; Lateef and Ruichek, 2019) in computer vision due to the success of 2D convolution on the regular and structured grids of 2D data images.

The first category of methods maps elements from 3D mesh scenes to points (pixels) on **2D texture maps**. The texture maps are the output of the parameterization (UV mapping) process which entails repeatedly cutting the mesh surface to a 2D domain while minimizing the distortions that are induced by the process. In CrossAtlasCNN (Li et al., 2019), faces from 3D mesh scenes are segmented and projected into atlases which are subsequently packed into 2D texture maps using a bin-packing method (Korf, 2002). Due to the discontinuities between atlases in a texture map, the authors developed a convolution operator i.e. *cross-atlas convolution* that jumps across the discontinuities to extract features. Similarly, features are upsampled and pooled using cross-atlas-based deconvolution and pooling operations respectively. Using these operations, a fully convolutional network (FCN) (Long et al., 2015) is used to classify the pixels (faces) of the texture maps. In a comparison between results obtained from semantic segmentation of urban scenes in a point cloud and aerial oblique images, the authors of Mesh-PC-Oblique (Wilk et al., 2022) rendered urban scenes of the SUM (Gao et al., 2021b) dataset using Pyrender (Matl, 2019) to oblique images with their corresponding labels. Semantic segmentation of the images is performed using Pyramid Scene Parsing Network (PSPNet) (Zhao et al., 2017) leveraging its ability to encode contextual information by concatenating extracted features from hierarchical image pyramids. Results of the segmentation are projected to texture maps of corresponding urban scenes on which the evaluation of the segmentation task is performed.

The second category of methods parameterizes elements or points of 3D scenes to **tangent planes or fields** from which features are extracted using convolution operations that are purposefully built for the parameterized domain. For instance, 3D mesh scenes are parameterized to four-way rotationally symmetric (4-RoSy) fields using QuadriFlow (Huang et al., 2018) in TextureNet (Huang et al., 2019) to generate a uniformly distributed orientation field of sampled points. A trade-off of the parameterization process is the inducement of directional ambiguity in the parameterized field. To extract features from the field, the authors developed a purposefully-built convolution operation i.e. *TextureConv* that gets rid of the directional ambiguity. Using the TextureConv operation, a UNet (Ronneberger et al., 2015) (Fig. 4) architecture is used to extract per-point features that are subsequently used for the downstream task of semantic segmentation. Borrowing from the concept of connections from differential geometry (Ray et al., 2009), local geodesic patches are sampled from the mesh scene and projected to regular tangent planes i.e. $\mathcal{N}$*-direction parallel frame fields* that are connected together to form flat Euclidean surfaces that support 2D convolution. By using frame fields, the authors of PFCNN (Yang et al., 2020) were able to connect and align the tangent planes to the geometric attributes on the mesh surface via parallel transport technique and also to aggregate learned features in $\mathcal{N}$*-directions*. A purposefully-built convolution operation i.e. *PFConv* combined with mesh simplification techniques (Garland and Heckbert, 1997; Hoppe, 1996) (Fig. 6) (pooling) are then used in a UNet-based (Ronneberger et al., 2015) (Fig. 4) network to extract per-vertex features that are used for classifying the mesh scenes. Similarly, tangent-based convolutions (continuous kernel convolution) are used in TangentCNN (Tatarchenko et al., 2018) to extract features from 2D-based tangent images that are generated from per-point tangent planes sampled from point cloud and mesh surfaces.

Even though the global nature of the parameterization process has enabled DNNs to process large-scale urban scenes, methods in this category still face a lot of challenges. These problems include distortions, discretizations, and occlusions that are usually induced in the parameterized domain as a result of the parameterization process. This inevitably affects semantic segmentation results, negatively.

3.3.2. Multimodal methods

The methods under this category are hybrid models that utilize information from other data formats like point clouds and images in conjunction with 3D meshes for the segmentation task. These methods leverage the more matured image- and point cloud-based DNN architectures for element-wise feature extraction and label prediction from images and point clouds that are usually generated from related 3D meshes. Evaluations are usually mesh-based in these methods.

The first category of methods combines **image and mesh** processing to achieve semantic segmentation. For facet-wise classification of reconstructed urban meshes from calibrated unlabeled images, a set of the images are used to fine-tune a 2D segmentation network i.e. DeepLabv3+(Chen et al., 2018) and predict the probabilities of pixels' labels in IterativeActiveLearning (Rong et al., 2021). These probabilities are back-projected onto the 3D reconstructed scenes followed by an MRF framework that assigns labels to the faces of the mesh scenes. This process is repeated iteratively using the quality of the labeled mesh scenes as criteria to evaluate the 2D segmentation process and to choose suitable images that will be returned to the training set as labeled for subsequent iterations. In InstanceSegMesh (Chen et al., 2022a), 2D–3D fusion is used for instance segmentation of 3D mesh buildings in urban scenes. To enable the hybrid segmentation process, the authors developed the first mesh-based instance segmentation benchmark dataset i.e. *InstanceBuilding* that comprises labeled UAV images and their corresponding 3D mesh scenes. The method first uses the Swin transformer network (Liu et al., 2021) to segment and compute instance masks for roofs in the UAV images followed by a back-projection of the masks onto the 3D mesh buildings to generate corresponding 3D instances. Finally, an MRF formulation is used to segment the remaining parts of the 3D mesh buildings. To our best knowledge, InstanceSegMesh (Chen et al., 2022a) is the first mesh-based method for instance segmentation.

The second category of methods is a hybrid of **point clouds and 3D meshes**. In Mesh-Sampled-PC (Grzeczkowicz and Vallet, 2022), point clouds are generated by sampling urban scenes from the SUM (Gao et al., 2021b) dataset. Using the well-known point cloud semantic segmentation network KPConv (Thomas et al., 2019), points are semantically labeled and converted to face-wise classification by adding the predicted probabilities of points that comprise each face.

The third category of methods transforms the 3D mesh scene to a **centre of gravity (COG)** point cloud representation and subsequently uses point-based DNNs to semantically segment the COG cloud. The COG cloud is an abstraction of the mesh surface that represents faces with their centers of gravity and their associated features e.g. geometric and radiometric face-wise features. Generated COG clouds usually have the same number of points with the faces of the mesh. The advantage of COG clouds over point clouds is the fusion of mesh-intrinsic features e.g. geometric, radiometric, and adjacency information to the former. In MultiBranch1D-CNN (Tutzauer et al., 2019), multi-scale radiometric and geometric feature vectors are computed per face to generate a COG cloud that is fed into a 1D CNN (George et al., 2018) that is extended to classify faces of the mesh scene. To evaluate their Point Cloud-Mesh-Association (PCMA) mechanism, a technique of transferring point cloud

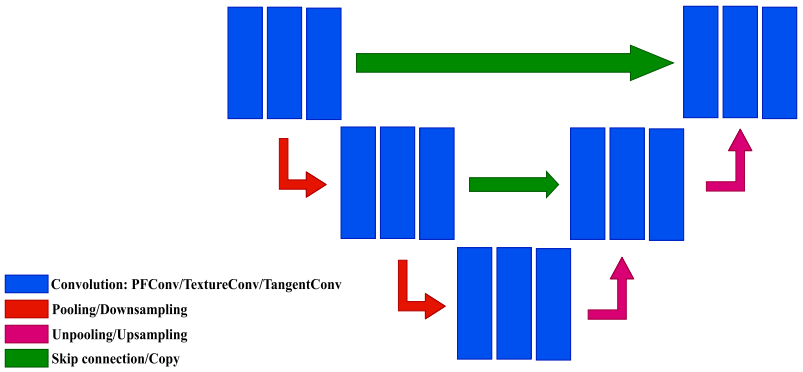


Fig. 4. A template of the UNet architecture for vertex-wise (ScanNet (Dai et al., 2017) dataset) classification using N locally flat cover sheets, 4-RoSy tangent vectors and 2D tangent images as inputs and *PFConv*, *TextureConv* and *Tangent convolutions* as convolutions in PFCNN (Yang et al., 2020), TextureNet (Huang et al., 2019) and TangentCNN (Tatarchenko et al., 2018) respectively. While for classification of vertices of the ScanNet (Dai et al., 2017) dataset.

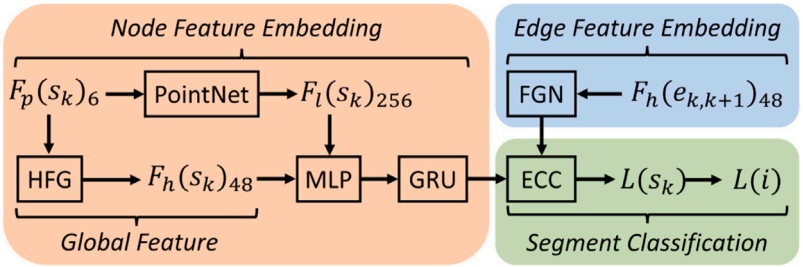


Fig. 5. An illustration of the three modules involved in the classification of superfacets in PSSNet (Weixiao et al., 2023). The ***Node Feature Embedding*** module learns and refines per-superfacet features, $F(S_k)_{256}$ from face centroids and RGB values $F_p(S_k)_6$ using PointNet (Qi et al., 2017a) while handcrafted features (HFG), $F_h(S_k)_{48}$ are embedded via Gated Recurrent Units (Cho et al., 2014) (GRUs). ***Edge Feature Embedding*** module uses a filter generating network (FGN) to learn features ($F_h(e_{k,k+1})_{48}$) of edges between superfacets through which they exchange information for the refinement process. Finally, Gated Graph Neural Networks (Li et al., 2016) (GGNN) and Edge-Conditions Convolutions (Simonovsky and Komodakis, 2017) (ECC) take as input both embeddings and predict classes of superfacets ($L(S_k)$) which are transferred to its constituent faces ($L(i)$) in the ***Segment Classification*** module.
Source: Figure taken from PSSNet (Weixiao et al., 2023).

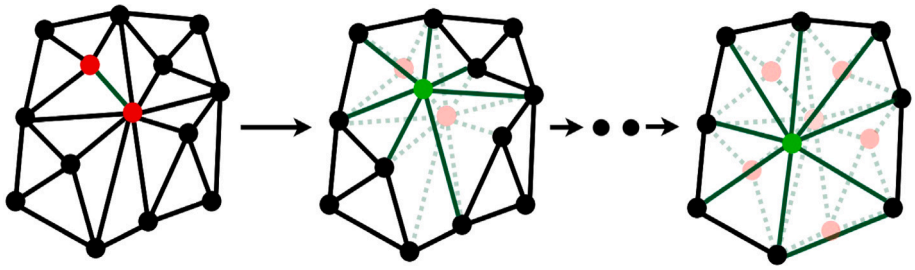


Fig. 6. An illustration of QEM-based (Garland and Heckbert, 1997) simplification operation where edges between adjacent vertices are repeatedly contracted to generate a new vertex. The feature of the new vertex is computed as the average of the contracted vertices. It is used to generate multi-resolution hierarchies of 3D meshes in DCM-Net (Schult et al., 2020) and as a pooling operation in PFCNN (Yang et al., 2020), PicassoNet and PicassoNet-II (Lei et al., 2021a,b) (GPU-accelerated version).

features and labels to faces of 3D mesh scenes, the authors of PCMA-Net (Laupheimer et al., 2020b) constructed per-face COG clouds and employed PointNet++ (Qi et al., 2017b) to classify the cloud points (faces of the mesh) to urban classes like roofs, buildings, vegetation and vehicles of the H3D (Cramer et al., 2018; Kölle M. Laupheimer et al., 2021) dataset. Using the PCMA (Laupheimer et al., 2020b) mechanism, ablation experiments were conducted in MultiModal-Net (Laupheimer and Haala, 2022) which showed the superiority of using multimodal features over mesh-only features. Similarly, ablation studies were carried out on COG clouds with enhanced radiometric information to investigate the effect of the feature on semantic segmentation results in RadiometricNet (Laupheimer et al., 2020a). PointNet++ (Qi et al., 2017b) was shown to outperform PointNet (Qi et al., 2017a) and RF classifiers in the semantic segmentation task due to its multi-scale learning ability from subsampled COG clouds. Finally, the ability of transformer (Vaswani et al., 2017) networks to learn and handle long dependencies between entities is leveraged in TransformerMesh (Tang et al., 2022) to encode contextual information between points of COG clouds that are abstracted from mesh scenes of the SUM (Gao et al., 2021b) and Wuhan (Tang et al., 2022) datasets. By representing the COG cloud as a graph, the authors of TransformerMesh (Tang et al., 2022) introduced a sampling method (based on edge distances) that is used to downsample and upsample the COG graph for efficient hierarchical feature extraction. Using a PointNet++-inspired (Qi et al., 2017b) architecture, the extracted features are used in labeling the faces (points on the COG cloud) of the mesh urban scenes.

Similar to parameterization-based methods 3.3.1, methods in this category leverage the feature learning abilities of well-known 2D and/or point-based DNNs in conjunction with relevant information from 3D meshes for element-wise classification. On the other side, the detour of segmenting mesh scenes via other data formats affects the performance of these methods due to the loss of information that

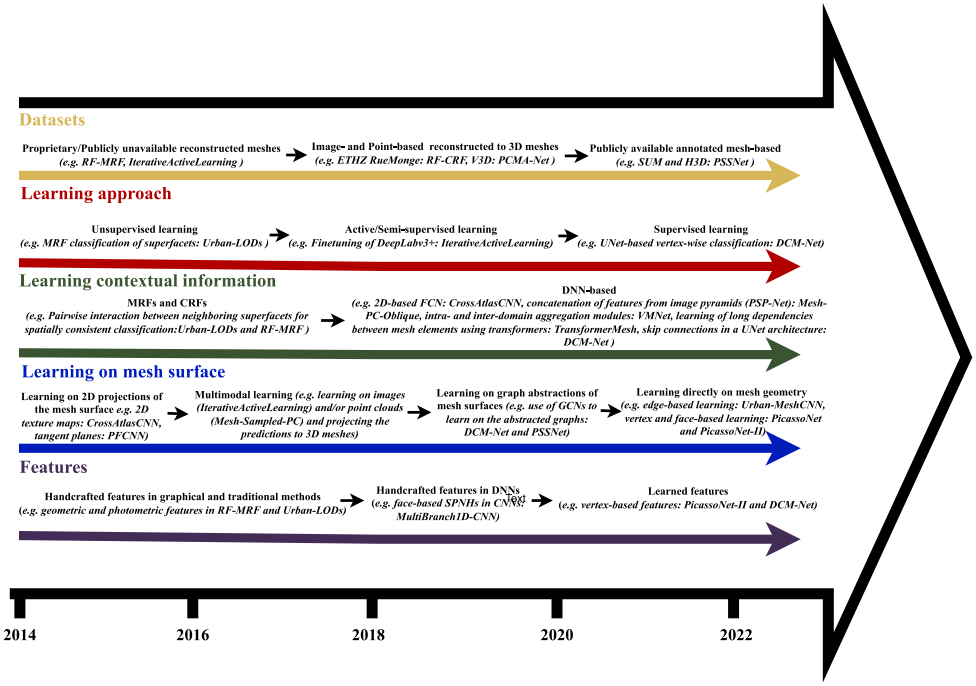


Fig. 7. An illustration of themes and trends of the progress made by semantic segmentation methods in various aspects of deep learning.

accompanies the sampling, projection, or COG abstraction techniques. Mesh-intrinsic challenges are sometimes carried over to the abstracted surface such as the non-uniformity of face densities on mesh surfaces affecting class labels in abstracted COG clouds by assigning more points and labels to less planar surfaces like vegetation than more planar surfaces like walls. This point and class imbalance have to be addressed by transfer techniques like PCMA (Laupheimer et al., 2020b). Also, the final segmentation results significantly depends on the performance of the sub-networks that process the other data formats.

3.3.3. Graph-based methods

The recent successes of geometric deep learning (Bronstein et al., 2017, 2021) methods in areas like social sciences and computer graphics have enabled researchers to employ different learning techniques e.g. convolution, pooling, and feature aggregation on non-Euclidean data formats like social networks and 3D mesh surfaces respectively. In mesh-based semantic scene analysis, the mesh is represented as a graph with the chosen elements (mostly vertices and faces) denoting nodes of the graph while adjacency of elements is represented as edges connecting adjacent nodes on the graph. Using this graph-based representation, methods in this category design or leverage (Wu et al., 2020; Bronstein et al., 2017, 2021) the feature processing abilities of graph neural networks (GNNs) to construct architectures for semantic segmentation of urban-scale 3D mesh scenes.

The first category of methods in this category extract and fuse learned features from the **geodesic and Euclidean domains** of the mesh graph and use them for the semantic segmentation task. In DCM-Net (Schult et al., 2020), geodesic and Euclidean graph-based convolutions (Wang et al., 2019) are simultaneously operated on geodesic and Euclidean neighborhoods of mesh-induced graphs in a UNet-based (Ronneberger et al., 2015) architecture to compute vertex-based predictions for labeling mesh scenes in three (Dai et al., 2017; Chang et al., 2017; Armeni et al., 2016) large scale indoor scenes. To extract multi-scale features from different resolutions of the mesh scenes, the authors adapted two mesh simplification techniques i.e. quadric error metrics (QEM) (Garland and Heckbert, 1997) (Fig. 6) and vertex clustering (VC) (Rossignac and Borrel, 1993) to generate downsampled versions of the scenes at each level of the architecture. To enhance context information learning, the method concatenates learned features from the geodesic and Euclidean convolutions in each layer thereby fusing local objects features with geodesically disconnected inter-object features respectively. A major addition of DCM-Net (Schult et al., 2020) is the development of an edge-based sampling method i.e. *random edge sampling (RES)* on the Euclidean neighborhood of the mesh-graph that enhances the generalizability of the network and at the same time, decreases the computational cost of training the network. A similar fashion of dual feature learning from Euclidean and geodesic domains is employed in VMNet (Hu et al., 2021), where multi-scale contextual features that are extracted (Tang et al., 2020) from downsampled (Garland and Heckbert, 1997; Rossignac and Borrel, 1993) hierarchies of voxelized versions of the mesh in Euclidean domain are projected back onto vertices in the geodesic domain. Afterward, the projected features are refined and aggregated in the geodesic domain followed by fusing them with the learned Euclidean-based features. Similar to DCM-Net (Schult et al., 2020), a UNet-based (Ronneberger et al., 2015) architecture is employed to extract the vertex-based features that are used for the downstream task of semantic segmentation.

The second category of methods is **overegmentation-based**. The methods cluster similar elements into homogeneous segments and use the segments as processing units for semantic segmentation DNNs. This way, the computational load of processing large-scale urban scenes is reduced to a tractable one. Also, the intra-relationship between elements in segments and inter-relationship between segments can be used to encode local and global contextual information between elements and semantic entities in scenes. Following this, planarity-preserving segments are generated from urban scenes by clustering planar and non-planar faces of the SUM (Gao et al., 2021b) and H3D (Kölle M. Laupheimer et al., 2021) datasets into homogeneous segments using a combination of RF-MRF framework and graph cut (Boykov et al., 2001) algorithm in PSSNet (Weixiao et al., 2023). Inspired by large-scale point-based oversegmentation methods like SPG (Landrieu and Simonovsky, 2018) and SSP (Landrieu and Boussaha, 2019), graphs that have the segments as nodes are constructed with edges connecting adjacent segments and also representing contextual dependencies between semantic entities of urban scenes. Using PointNet (Qi et al., 2017a), nodal embeddings are generated and subsequently used in

classifying the segments using a well-known GNN (Li et al., 2016). The final per-segment label predictions are then transferred to their constituent faces to evaluate the per-face labeling accuracy of the method (Figs. 5 and 10).

The methods in this category are much closer to processing meshes directly because graphs are data formats that represent the underlying meshes as closely as possible. Improvements and new methods are increasingly introduced in the field of geometric deep learning which opens new opportunities for researchers to adapt them for semantic segmentation of urban 3D meshes. However, methods in this category face challenges such as non-differentiability of clustering or simplification algorithms which prevents end-to-end learning. Also, graph-based methods tend to store complete graphs in memory during training which makes it computationally prohibitive to process large-scale urban scenes.

3.3.4. Mesh-intrinsic methods

The methods in this category define learning operations directly on the intrinsic geometric entities i.e. faces, vertices, and/or edges thereby circumventing all pre-processing techniques like parameterization or transformation to graphs or COG clouds. This way, these methods do not suffer from the information loss induced by the pre-processing techniques and also they benefit from directly learning and extracting rich features from the mesh scenes.

The first category of methods learns features directly from the **vertices and faces** of the mesh scenes. In both PicassoNet (Lei et al., 2021a) and PicassoNet-II (Lei et al., 2021b), vertex- and face-based convolution operations together with a GPU-accelerated QEM-based (Garland and Heckbert, 1997) (Fig. 6) simplification method that is used for pooling and unpooling operations were developed as part of the *Picasso* library. By enabling parallel GPU processing of the QEM (Garland and Heckbert, 1997) (Fig. 6) technique, large-scale mesh segmentation networks i.e. *PicassoNet* and *PicassoNet-II* that process batches on the fly are trained in an end-to-end manner. The simplification technique is also used in PicassoNet-II (Lei et al., 2021b) to generate the multi-resolution meshes from which multi-layered features are extracted for vertex-based labeling in an encoder–decoder architecture.

The second category of methods is **edge-based**. In Urban-MeshCNN (Knott and Groenendijk, 2021), urban mesh scenes were classified into four classes i.e. *building, ground, low vegetation, and high vegetation* using an extended version of MeshCNN (Hanocka et al., 2019). MeshCNN (Hanocka et al., 2019) is a well-known mesh-based DNN that utilizes edge-specific convolutions, pooling, and unpooling operations for processing mostly synthetic small-scale toy datasets. To enable its utilization on photogrammetric meshes, the authors of Urban-MeshCNN (Knott and Groenendijk, 2021) made two significant additions: breadth-first search (BFS) partitioning technique to generate manageable chunks of the scenes for efficient processing and the addition of photometric features to the existing geometric features to improve the discriminative abilities of the network. Using these enhancements, per-edge predictions are computed.

Methods under this category are the end goal of mesh DNNs because they leverage the intrinsic geometries of the mesh scenes thereby enabling direct feature learning and processing. Still, techniques that will alleviate challenges of end-to-end large-scale learning e.g. QEM-based (Garland and Heckbert, 1997) simplification (Fig. 6) in PicassoNet, PicassoNet-II (Lei et al., 2021a,b) and the BFS chunking technique in Urban-MeshCNN (Knott and Groenendijk, 2021) are needed for tractable and efficient processing of urban-scale mesh scenes.

Fig. 7 illustrates the different themes and trends in deep learning-based semantic segmentation of 3D mesh scenes. In terms of datasets, it can be observed that mesh-specific datasets i.e. SUM (Gao et al., 2021b) and H3D (Kölle M. Laupheimer et al., 2021) are beginning to replace image- and point-based (Cramer, 2010) datasets. This is due to the increasing interest in directly processing and classifying the mesh surface instead of obtaining results from other data formats and projecting them back to the mesh for evaluation. Semantic segmentation results obtained from direct processing of mesh scenes faithfully represent the perceived scenes more than results from other data formats. This is the same justification for the *"Learning on mesh surface"* theme where the progression of defining learning-based operations on 2D projections (Li et al., 2019; Huang et al., 2019; Yang et al., 2020; Tatarchenko et al., 2018), sampled point clouds (Grzeczkowicz and Vallet, 2022) and graph abstractions (Schult et al., 2020; Weixiao et al., 2023) to intrinsic geometry (Lei et al., 2021a,b) of the mesh surface. It is a progression from indirect to direct classification of mesh scenes. As for the *"Learning approach"* theme, initial methods especially the modeling-based (Zhu et al., 2018; Verdie et al., 2015) approaches used unsupervised clustering techniques to segment mesh scenes into homogeneous segments (superfacets) which are subsequently classified into urban classes using (mostly) graphical models. The increasing availability of annotated datasets (Gao et al., 2021b; Kölle M. Laupheimer et al., 2021; Dai et al., 2017) has enabled methods to use supervised approaches like RF (Gao et al., 2021b; Rouhani et al., 2017) and DNNs (Schult et al., 2020; Yang et al., 2020; Wilk et al., 2022) to classify elements of 3D meshes.

Another theme that is not shown in Fig. 7 is the progression of the various techniques of reducing urban meshes to tractable units to handle the computational load of training methods. Early methods used unsupervised clustering algorithms like the region-growing technique (Lafarge and Mallet, 2012) in RF-MRF (Rouhani et al., 2017) to cluster similar faces into superfacets that are later used as processing units in the downstream task of semantic segmentation. These techniques are usually time-consuming and therefore, later methods parameterized whole mesh scenes into 2D (Yang et al., 2020; Huang et al., 2019; Li et al., 2019) and point-based (Grzeczkowicz and Vallet, 2022; Tang et al., 2022) data formats that are suitable and computationally-efficient to process using standard 2D CNNs (Chen et al., 2018; Liu et al., 2021) and point-based (Qi et al., 2017a,b; Thomas et al., 2019) methods respectively. However, these techniques are not mesh-intrinsic and differentiable which leads to a loss of information and the inability to use them in collaboration with other modules in an end-to-end trainable method. As a solution, recent methods used mesh-intrinsic simplification methods such as the use of QEM (Garland and Heckbert, 1997) and learning-based edge collapse (Hanocka et al., 2019) techniques which are topology-preserving edge collapse methods in DCM-Net (Schult et al., 2020) PicassoNet-II (Lei et al., 2021b) and Urban-MeshCNN (Knott and Groenendijk, 2021) respectively. Furthermore, the QEM6 simplification method in PicassoNet (Lei et al., 2021a) and PicassoNet-II (Lei et al., 2021b) is GPU-accelerated which makes it seamless to use it in an end-to-end DNN.

4. Datasets and evaluation metrics

Here, we give brief descriptions of the benchmark urban- and large-scale mesh datasets for semantic segmentation tasks. We also provide the evaluation metrics that are used in assessing the performance of segmentation methods.

4.1. Datasets

Urban-scale mesh scenes are usually generated from natural scenes of point clouds or multi-view images using MVS and structure-from-motion (SfM) reconstruction methods (Lafarge et al., 2012; Vu et al., 2011) or off-the-shelf commercial products like SURE Aerial,[1] PIX4Dmapper[2] (PIX4D), ContextCapture[3] (Bentley Solutions) and MeshLab (open source) (Cignoni et al., 2008). The reconstructed mesh scenes are then processed and labeled manually or by transferring pixel or point labels from the images or point clouds respectively to the mesh.

[1] https://www.nframes.com/products/sure-aerial/

[2] https://www.pix4d.com/product/pix4dmapper-photogrammetry-software/

[3] https://www.bentley.com/software/contextcapture/

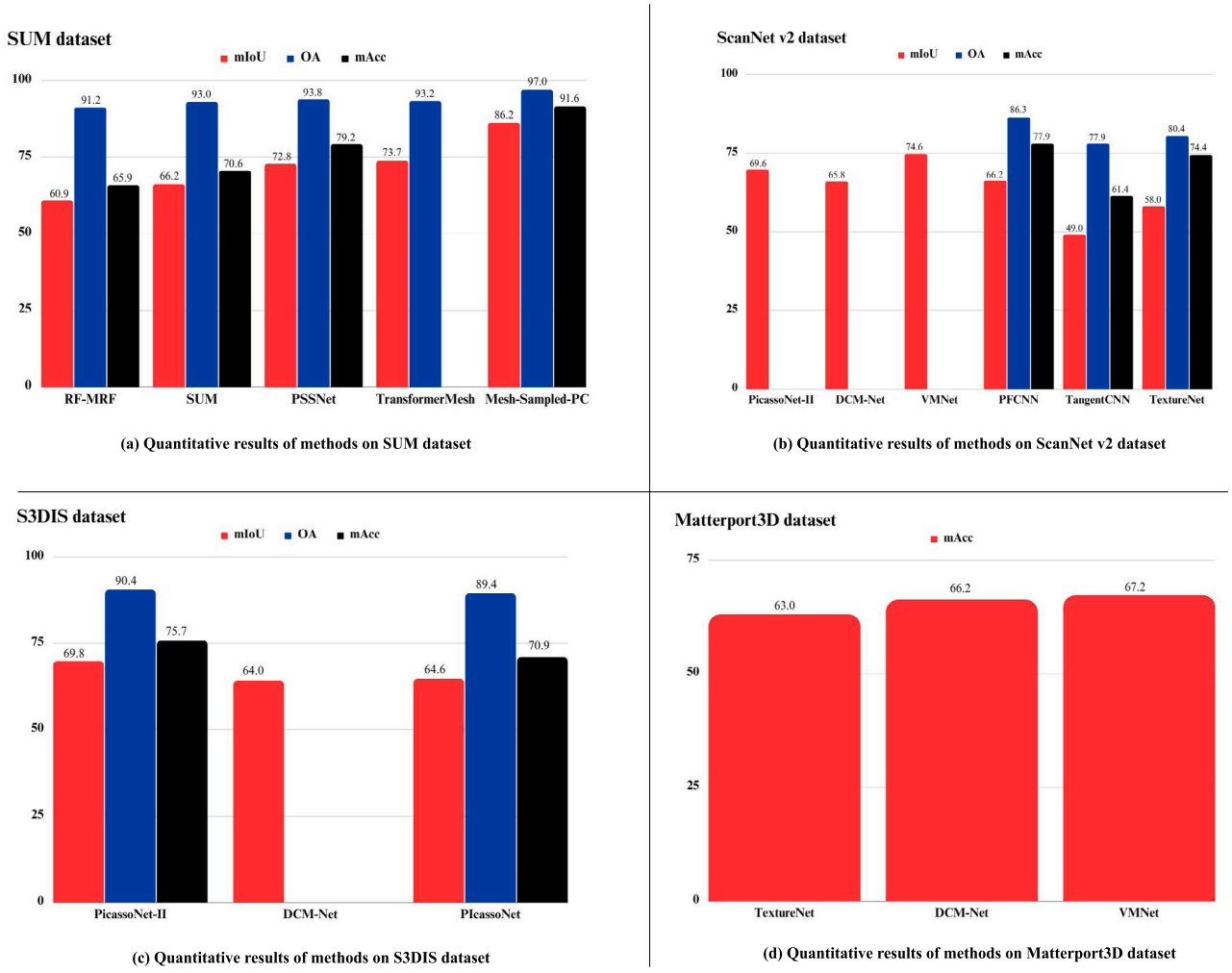


Fig. 8. Quantitative results based on the three most frequently used semantic segmentation-based metrics i.e. *mIoU, mAcc and OA* obtained from experiments on four popular large-scale mesh datasets i.e. SUM (Gao et al., 2021b), ScanNet v2 (Dai et al., 2017), S3DIS (Armeni et al., 2016) and, Matterport3D (Chang et al., 2017).

There are different categories of semantic segmentation datasets for urban meshes. We give brief discussions of the prominent mesh-based semantic segmentation datasets below:

- **SUM** (Gao et al., 2021b): contains 64 tiles with each covering an area of 250 × 250 m^2. 40, 12, and 10 tiles are used for training, testing, and validation of methods. The mesh scenes are reconstructions of Helsinki, Finland covering 4 km^2 of its urban area. The dataset is labeled into six categories i.e. *terrain, building, high vegetation, water, vehicle, and boat,* and an *unclassified* class that contains distorted and incomplete objects. The labels are carried on the faces of the mesh.
- **Hessigheim 3D (H3D)** (Kölle M. Laupheimer et al., 2021): covers 0.19km^2 of Hessigheim, a village in Germany. The dataset has both 3D mesh and point cloud modes with faces carrying the labels in the mesh data mode. 40% of the faces are unlabeled while the remaining faces are labeled into 11 categories i.e. *facade, roof, shrub, tree, soil/gravel, low vegetation, impervious surface, vehicle, urban furniture, vertical surface, and chimney*.
- Real-world indoor 3D mesh reconstructions: **ScanNet v2** (Dai et al., 2017) contains 1513 3D mesh reconstructions of indoor scenes with its vertices labeled into 20 categories including classes such as bathroom, closet, kitchen, gym, hallway e.t.c. and a miscellaneous class. **Matterport3D** (Chang et al., 2017) contains 90 3D mesh building instances reconstructed from RGB-D scenes using Poisson (Chuang and Kazhdan, 2011) surface reconstruction technique. It used the same labeling categories of ScanNet v2 (Dai et al., 2017) dataset with faces carrying the labels. Similarly, indoor mesh scenes are reconstructed from RGB-D scenes in **SceneNN** (Hua et al., 2016) with vertex-wise labeling. **Stanford Large-scale 3D Indoor Spaces (S3DIS)** (Armeni et al., 2016) is a 2D-3D-S that contains registered 3D annotated 3D meshes and point clouds reconstructed from RGB-D images. The labels in the 3D mesh modality are face-wise.
- **ETHZ RueMonge 2014** (Riemenschneider et al., 2014): contains 700m long pixel-level annotated street scenes of Rue Monge, Paris, France along with indexes to their 3D format equivalents i.e. mesh and point cloud. The dataset can be used for image, mesh, and point cloud labeling tasks.
- InstanceBuilding (Chen et al., 2022a): 3D instance segmentation dataset from InstanceSegMesh (Chen et al., 2022a) containing annotations for UAV building images and their 3D instances. Out of the 892 building scenes, images have both their 2D and 3D components annotated and is the first dataset specifically developed for instance segmentation of buildings in urban scenes.

Other relevant datasets are **Vaihingen 3D (V3D)** (Cramer, 2010) (point cloud) and **SUN RGB-D** (Song et al., 2015). For a systematic and detailed analysis of remotely sensed benchmark datasets for earth observation, EarthNets (Xiong et al., 2022) is a good resource.

There are also proprietary datasets that are not publicly made available for researchers in the community. These datasets are used

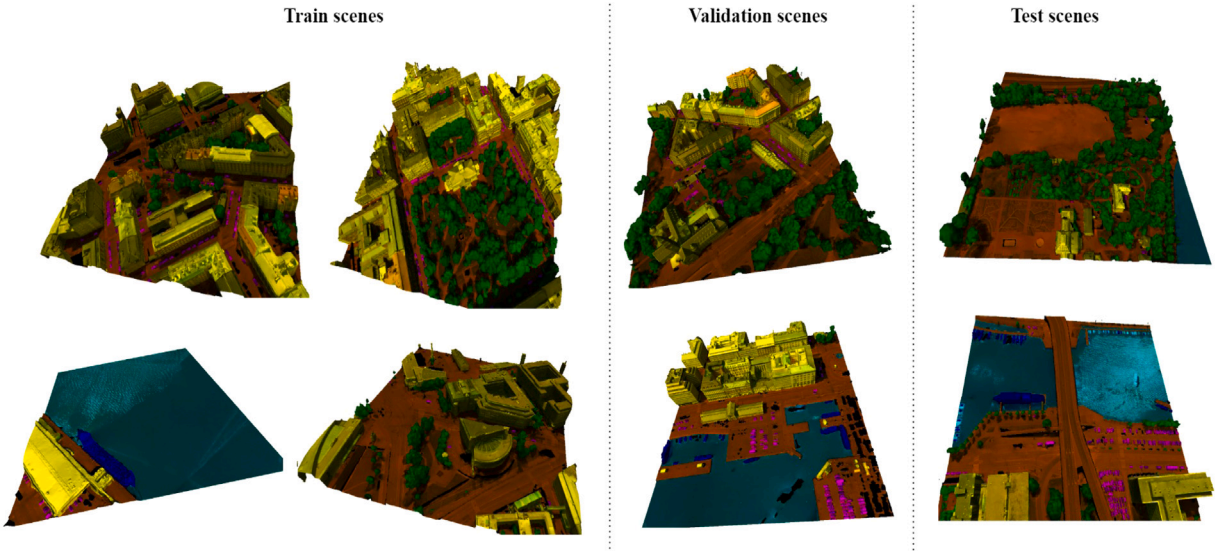


Fig. 9. Exhibition of 3D mesh scenes from the SUM (Gao et al., 2021b) urban dataset.

by authors to evaluate the performance of their methods. Examples include the Wuhan dataset in TransformerMesh (Tang et al., 2022) and the reconstructed urban scenes of Paris and Toulouse, France in RF-MRF (Rouhani et al., 2017).

4.2. Evaluation metrics

Several evaluation metrics have been proposed to evaluate the performance of semantic segmentation methods and other related tasks such as instance segmentation and oversegmentation. For semantic segmentation (Gao et al., 2021b; Grzeczkowicz and Vallet, 2022), the most frequently used metrics are *Intersection over Union (IoU), mean per-class Intersection over Union (mIoU), precision, recall, Accuracy (Acc), Overall Accuracy (OA), mean per-class Accuracy (mAcc), and F1 score*. For instance segmentation (Chen et al., 2022a), the well-known instance-level metric (Lin et al., 2014) of *Average Precision (AP)* fixed at *IoU* values of *0.5 (AP50)* and *0.75 (AP75)* are used in InstanceSegMesh (Chen et al., 2022a). The authors used area of faces to compute the *IoUs*. For oversegmentation, the authors of PSSNet (Weixiao et al., 2023) developed mesh-adapted evaluation metrics to assess the oversegmentation performance of their method. Similar to the point-based (Landrieu and Simonovsky, 2018; Landrieu and Boussaha, 2019) counterparts of PSSNet (Weixiao et al., 2023), the authors used *object purity (OP), boundary precision (BP) and boundary recall (BR)* to measure the quality of generated superfacets. Face areas are also used for most of the computations of these metrics.

From the results shown in Fig. 8(a), Mesh-Sampled-PC (Grzeczkowicz and Vallet, 2022) performs better in all the reported metrics. This is mainly due to the utilization of KPConv (Thomas et al., 2019), a well-known and matured point cloud semantic segmentation method, to classify sampled point clouds from mesh scenes of the SUM (Gao et al., 2021b) dataset. It enabled the method to leverage the fine-tuned ability of the KPConv (Thomas et al., 2019) model to classify the point clouds hence the superior performance over other mesh-based methods. Even though Mesh-Sampled-PC (Grzeczkowicz and Vallet, 2022) is +12.5 and +13.4 points (*mIoU*) (Fig. 8(a)) ahead of TransformerMesh (Tang et al., 2022) and PSSNet (Weixiao et al., 2023) respectively, we believe the approaches employed by the latter methods to process the mesh scenes will continue to garner interest from researchers in the field. This is because these methods process representations e.g. superfacet-based graphs in PSSNet (Weixiao et al., 2023) that are much closer to representing the mesh surface than the sampled point clouds in Mesh-Sampled-PC (Grzeczkowicz and Vallet, 2022) (Fig. 7). Consequently, the results (albeit lower) obtained by these methods are more useful in understanding labeled mesh scenes because of the proximity of the representation.

Also, the results (Figs. 8(a) and 8(b)) of PSSNet (Weixiao et al., 2023) and TransformerMesh (Tang et al., 2022) are very similar which points to the similarity with which both methods employed in classifying the mesh scenes. Both methods simplify the input meshes into manageable processing units i.e. superfacets in PSSNet (Weixiao et al., 2023) and COG clouds in Tang et al. (2022) and subsequently used well-known models in GCNs and transformer networks respectively for the semantic segmentation task. It will be interesting to see how end-to-end mesh-intrinsic methods such as PicassoNet-II (Lei et al., 2021b) will perform on the urban scenes of the SUM (Gao et al., 2021b) dataset.

From Fig. 8(b), we can observe the superiority of VMNet (Hu et al., 2021) in terms of the *mIoU* metric over other methods. This can be explained by the method's utilization of a well-matured sparse voxel-based (Tang et al., 2020) convolution operation in UNet (Ronneberger et al., 2015) architecture on voxelized hierarchies of the mesh scenes. Combined with its intra- and inter-domain aggregation modules, we believe these are the reasons for the superior performance (+8.8 *mIoU*) of VMNet (Hu et al., 2021) over its graph-based, multidomain (Euclidean and geodesic) counterpart, DCM-Net (Schult et al., 2020). Of note are the performances of the mesh-intrinsic end-to-end methods, PicassoNet (Lei et al., 2021a) and PicassoNet-II (Lei et al., 2021b) in Figs. 8(b) and 8(c). This is mainly due to the utilization of the different modules of the *Picasso* library such as the GPU-accelerated QEM (Garland and Heckbert, 1997) (Fig. 6) decimation method that is 30 times faster than its non-GPU counterpart used in methods like DCM-Net (Schult et al., 2020) and VMNet (Hu et al., 2021).

For global parameterization methods (Section 3.3.1 and Fig. 8(b)), the best performing method is PFCNN (Yang et al., 2020) with +8.2 (*mIoU*), +5.9 (*OA*) and +3.5 (*mACC*) point increase on the second best performing method, TextureNet (Huang et al., 2019) based on experiments on the ScanNet v2 dataset. Even though these methods benefit from using matured and fine-tuned 2D (Li et al., 2019) and point-based (Wilk et al., 2022) on parameterized domains, research interest in the area is waning as shown in Fig. 7 (Learning on mesh surface) and Table 1. This is partly due to the information loss inherent in the parameterization method and the growing theme of designing methods that directly process the mesh geometry. This trend is showing promise as illustrated by an increase of +3.4 points (*mIoU*) (Fig. 8(b)) of PicassoNet-II (Lei et al., 2021b) over PFCNN (Yang et al., 2020).

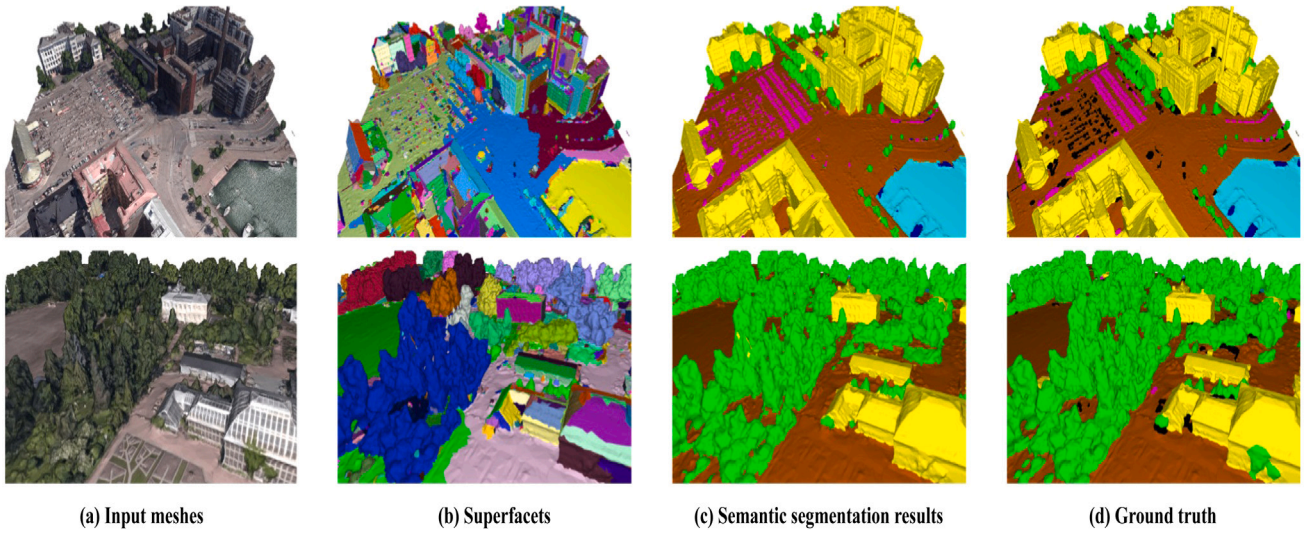


Fig. 10. Semantic segmentation results obtained by PSSNet (Weixiao et al., 2023) on the SUM (Gao et al., 2021b) dataset.

5. Open challenges and future research directions

Notwithstanding the successes of DNNs in the semantic segmentation of urban scenes, there are contemporary challenges that need further attention from researchers in this area.

(1) Annotated datasets: It can be observed from Section 4.1 that there is an urgent need for annotated urban-scale datasets. There are only two (Gao et al., 2021b; Kölle M. Laupheimer et al., 2021) annotated urban-scale datasets purposefully built for semantic segmentation of 3D scenes while the remaining (Dai et al., 2017; Chang et al., 2017; Armeni et al., 2016; Cramer, 2010; Riemenschneider et al., 2014) datasets are either for indoor scenes or built for point- and image-based segmentation tasks. There is also a need for datasets that enable other segmentation-related tasks like instance segmentation and object detection as evidenced by only one (Chen et al., 2022a) dataset for the former task and none for the latter. Therefore, more mesh-specific publicly available benchmark datasets are needed.

(2) Segmentation-related tasks: As 3D meshes are becoming the preferred format of visualization in computer vision (Gao et al., 2021b), more methods for analyzing urban scenes are required for perceiving and discriminating the various objects and artifacts that make up the real world. The glaring absence of object detection methods and the almost non-existence of instance-level detection methods (Chen et al., 2022a) in the methods we reviewed in this work points to a dire need for these types of methods in the research area.

(3) End-to-end trainability: Of all the methods we surveyed, only a few (Lei et al., 2021a,b) are end-to-end trainable. This is mainly due to the utilization of non-differentiable algorithms like the region-growing algorithms in SUM (Gao et al., 2021b) and PSSNet (Weixiao et al., 2023) or the GPU non-compliant simplification methods in DCM-Net (Schult et al., 2020) and VMNet (Hu et al., 2021). Although there are attempts to address this challenge such as the GPU-accelerated QEM (Garland and Heckbert, 1997) (Fig. 6) in PicassoNet (Lei et al., 2021a) and PicassoNet-II (Lei et al., 2021b), more similar solutions are needed to help researchers train mesh DNNs efficiently.

(4) Computational tractability: Processing urban-scale mesh scenes involve dealing with thousands or sometimes millions of vertices and faces. Existing methods try to strike a balance between techniques that transforms the scenes into more tractable units such as superfacets (Weixiao et al., 2023), COG clouds (Tang et al., 2022), or 2D texture maps (Li et al., 2019) and the information loss that usually accompanies these techniques. Therefore, more mesh-specific techniques that preserve the geometric and topological features of scenes as much as possible while at the same time reducing computational load are required. One of the steps in this direction is the random edge sampling (RES) method for sampling the mesh-graph neighborhood in DCM-Net (Schult et al., 2020).

(5) Defects of urban-scale mesh scenes: Most of the existing semantic segmentation DNNs for small-scale datasets e.g. MeshCNN (Hanocka et al., 2019) are designed to process cleaned, manifold, and genus 0 mesh surfaces. However, reconstructed urban scenes of 3D meshes contain a lot of geometric and topological defects, noise, distortions, and non-manifold edges and vertices. Therefore, methods that process urban-scale scenes must have high robustness and resilience toward these defects. Mesh-specific convolution and pooling operations must also be able to generalize on these types of scenes.

6. Conclusion

To our best knowledge, this is the first review paper that focuses on semantic segmentation of urban-scale mesh scenes using deep learning. We gave a formulation for the semantic segmentation problem and pointed out the challenges that are encountered by mesh-based DNNs in achieving this task. Based on the different techniques employed by methods to address these challenges, we classified these methods and described the various ways they achieve the element-wise classification of mesh scenes. We then presented a comparative analysis of these methods highlighting their strengths and weaknesses followed by discussions on the datasets and evaluation metrics involved in executing and assessing methods. Finally, we discussed the contemporary challenges faced by researchers in the field and provide insights on future research avenues for researchers in the community.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61971363), China Postdoctoral Science Foundation (No. 2021M690094), and FuXiaQuan National Independent Innovation Demonstration Zone Collaborative Innovation Platform (No. 3502ZCQXT2021003).

References

An, A., 2023. Adopting metaverse-related mixed reality technologies to tackle urban development challenges: An empirical study of an australian municipal government. IET Smart Cities.

Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S., 2016. 3D semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1534–1543.

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 122, 2–1239.

Bronstein, M.M., Bruna, J., Cohen, T., Velickovic, P., 2021. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478.

Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P., 2017. Geometric deep learning: going beyond euclidean data. IEEE Signal Process. Mag. 34, 18–42.

Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., Savva, M., Song, S., Zeng, A., Zhang, Y., 2017. Matterport3d: Learning from rgb-d data in indoor environments. In: 2017 International Conference on 3D Vision (3DV). IEEE Computer Society, pp. 667–676.

Chen, Z., Deng, L., Luo, Y., Li, D., Junior, J.M., Gonçalves, W.N., Nurunnabi, A.A.M., Li, J., Wang, C., Li, D., 2022b. Road extraction in remote sensing data: A survey. Int. J. Appl. Earth Obs. Geoinf. 112, 102833.

Chen, J., Xu, Y., Lu, S., Liang, R., Nan, L., 2022a. 3-d instance segmentation of mvs buildings. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder–decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818.

Cho, K., Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP.

Chuang, M., Kazhdan, M., 2011. Interactive and anisotropic geometry processing using the screened Poisson equation. ACM Trans. Graph. 1–10.

Cignoni, P., Callieri, M., Corsini, M., Dellepiane, M., Ganovelli, F., Ranzuglia, G., et al., 2008. Meshlab: an open-source mesh processing tool. In: Eurographics Italian Chapter Conference, Salerno, Italy. pp. 129–136.

Cohen-Steiner, D., Alliez, P., Desbrun, M., 2004. Variational shape approximation. In: ACM SIGGRAPH 2004 Papers. pp. 905–914.

Cramer, M., 2010. The dgpf-test on digital airborne camera evaluation overview and test design. Photogrammetrie-Fernerkundung-Geoinformation 73–82.

Cramer, M., Haala, N., Laupheimer, D., Mandlburger, G., Havel, P., 2018. Ultra-high precision uav-based lidar and dense image matching. ISPRS-Int. Arch. Photogram., Remote Sens. Spatial Inf. Sci. 621, 115–120.

Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, pp. 2432–2443.

Du, R., Li, D., Varshney, A., 2019. Geollery: A mixed reality social media platform. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

Gao, W., Nan, L., Boom, B., Ledoux, H., 2021b. Sum: A benchmark dataset of semantic urban meshes. ISPRS J. Photogramm. Remote Sens. 179, 108–120.

Gao, B., Pan, Y., Li, C., Geng, S., Zhao, H., 2021a. Are we hungry for 3d lidar data for semantic segmentation? a survey of datasets and methods. IEEE Trans. Intell. Transp. Syst. 23, 6063–6081.

Garland, M., Heckbert, P.S., 1997. Surface simplification using quadric error metrics. In: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques. pp. 209–216.

George, D., Xie, X., Tam, G.K., 2018. 3D mesh segmentation via multi-branch 1d convolutional neural networks. Graph. Models 96, 1–10.

Grzeczkowicz, G., Vallet, B., 2022. Semantic segmentation of urban textured meshes through point sampling. ISPRS Ann. Photogram., Remote Sens. Spatial Inf. Sci. 2, 177–184.

Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M., 2021. Deep learning for 3d point clouds: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 43, 4338–4364.

Hanocka, R., Hertz, A., Fish, N., Giryes, R., Fleishman, S., Cohen-Or, D., 2019. Meshcnn: a network with an edge. ACM Trans. Graph. 1–12.

Hoppe, H., 1996. Progressive meshes. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. pp. 99–108.

Hu, Z., Bai, X., Shang, J., Zhang, R., Dong, J., Wang, X., Sun, G., Fu, H., Tai, C.L., 2021. Vmnet: Voxel-mesh network for geodesic-aware 3d semantic segmentation. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, IEEE Computer Society, pp. 15468–15478.

Hua, B.S., Pham, Q.H., Nguyen, D.T., Tran, M.K., Yu, L.F., Yeung, S.K., 2016. Scenenn: A scene meshes dataset with annotations. In: Int. Conf. on 3D Vis. (3DV). IEEE, pp. 92–101.

Huang, J., Zhang, H., Yi, L., Funkhouser, T., Nießner, M., Guibas, L.J., 2019. Texturenet: Consistent local parametrizations for learning from high-resolution signals on meshes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, pp. 4435–4444.

Huang, J., Zhou, Y., Niessner, M., Shewchuk, J.R., Guibas, L.J., 2018. Quadriflow: A scalable and robust method for quadrangulation. Comput. Graph. Forum 37, 147–160.

Kang, D., Wong, A., Lee, B., Kim, J., 2021. Real-time semantic segmentation of 3d point cloud for autonomous driving. Electronics 10, 19–60.

Knott, M., Groenendijk, R., 2021. Towards mesh-based deep learning for semantic segmentation in photogrammetry. ISPRS Ann. Photogram., Remote Sens. Spatial Inf. Sci. 5, 9–66.

Kolbe, T.H., Gröger, G., Plümer, L., 2005. Citygml: Interoperable Access To 3d City Models. Springer, pp. 883–899.

Kölle M. Laupheimer, D., Schmohl, S., Haala, N., Rottensteiner, F., Wegner, J.D., 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 J. Photogram. Remote Sens. 100001.

Korf, R.E., 2002. A new algorithm for optimal bin packing. In: Eighteenth National Conference on Artificial Intelligence. American Association for Artificial Intelligence, pp. 731–736.

Lafarge, F., Keriven, R., Brédif, M., Vu, H.H., 2012. A hybrid multiview stereo algorithm for modeling urban scenes. IEEE Trans. Pattern Anal. Mach. Intell. 5–17.

Lafarge, F., Mallet, C., 2012. Creating large-scale city models from 3d-point clouds: A robust approach with hybrid representation. Int. J. Comput. Vis. 99, 69–85.

Lafferty, J.D., McCallum, A., Pereira, F.C., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289.

Landrieu, L., Boussaha, M., 2019. Point cloud oversegmentation with graph-structured deep metric learning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, pp. 7432–7441.

Landrieu, L., Simonovsky, M., 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 4558–4567.

Lateef, F., Ruichek, Y., 2019. Survey on semantic segmentation using deep learning techniques. Neurocomputing 338, 321–348.

Laupheimer, D., Haala, N., 2022. Multi-modal semantic mesh segmentation in urban scenes. ISPRS Ann. Photogram., Remote Sens. Spatial Inf. Sci. 26, 7–274.

Laupheimer, D., Shams Eddin, M., Haala, N., 2020a. The importance of radiometric feature quality for semantic mesh segmentation. In: DGPF Annual Conference, Stuttgart, Germany. Publikationen der DGPF.

Laupheimer, D., Shams Eddin, M., Haala, N., 2020b. On the association of lidar point clouds and textured meshes for multi-modal semantic segmentation. ISPRS Ann. Photogram., Remote Sens. Spatial Inf. Sci. 2, 509–516.

Lei, H., Akhtar, N., Mian, A., 2021a. Picasso: A cuda-based library for deep learning over 3d meshes. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, pp. 13849–13859.

Lei, H., Akhtar, N., Shah, M., Mian, A., 2021b. Geometric feature learning for 3d meshes. arXiv preprint arXiv:2112.01801.

Lei, B., Stouffs, R., Biljecki, F., 2022. Assessing and benchmarking 3d city models. Int. J. Geogr. Inf. Sci. 1–22.

Li, S., Luo, Z., Zhen, M., Yao, Y., Shen, T., Fang, T., Quan, L., 2019. Cross-atlas convolution for parameterization invariant learning on textured mesh surface. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE Computer Society, pp. 6136–6145.

Li, Y., Zemel, R., Brockschmidt, M., Tarlow, D., 2016. Gated graph sequence neural networks. In: Proceedings of ICLR'16.

Liang, W., Xu, P., Guo, L., Bai, H., Zhou, Y., Chen, F., 2021. A survey of 3d object detection. Multimedia Tools Appl. 29617–29641.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, J., Wang, J., Fang, T., Tai, C.L., Quan, L., 2015. Higher-order crf structural segmentation of 3d reconstructed surfaces. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2093–2101.

Liu, W., Zang, Y., Xiong, Z., Bian, X., Wen, C., Lu, X., Wang, C., Junior, J.M., Gonçalves, W.N., Li, J., 2023. 3D building model generation from mls point cloud and 3d mesh using multi-source data fusion. Int. J. Appl. Earth Obs. Geoinf. 116, 103171.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3431–3440.

Mao, J., Shi, S., Wang, X., Li, H., 2022. 3D object detection for autonomous driving: a review and new outlooks. arXiv preprint arXiv:2206.09474.

Matl, M., 2019. Pyrender. https://github.com/mmatl/pyrender.

Osco, L.P., Junior, J.M., Ramos, A.P.M., de Castro Jorge, L.A., Fatholahi, S.N., de Andrade Silva, J., Matsubara, E.T., Pistori, H., Gonçalves, W.N., Li, J., 2021. A review on deep learning in uav remote sensing. Int. J. Appl. Earth Obs. Geoinf. 102, 102456.

Peng, Y., Qin, Y., Tang, X., Zhang, Z., Deng, L., 2022. Survey on image and point-cloud fusion-based object detection in autonomous vehicles. IEEE Trans. Intell. Transp. Syst. 1–18.

Qi, C.R., Su, H., Kaichun, M., Guibas, L.J., 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, pp. 77–85.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. pp. 5105–5114.
Ray, N., Li, W.C., Lévy, B., Sheffer, A., Alliez, P., 2006. Periodic global parameterization. ACM Trans. Graph. 25, 1460–1485.
Ray, N., Vallet, B., Alonso, L., Levy, B., 2009. Geometry-aware direction field processing. ACM Trans. Graph. 1–11.
Riemenschneider, H., Bódis-Szomorú, A., Weissenberg, J., Gool, L.V., 2014. Learning where to classify in multi-view semantic segmentation. In: European Conference on Computer Vision. Springer, pp. 516–532.
Rong, M., Shen, S., Hu, Z., 2021. 3D semantic labeling of photogrammetry meshes based on active learning. In: 2020 25th International Conference on Pattern Recognition. ICPR, IEEE, pp. 3550–3557.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October (2015) 5-9, Proceedings, Part III 18. Springer, pp. 234–241.
Rossignac, J., Borrel, P., 1993. Multi-resolution 3d approximations for rendering complex scenes. In: Modeling in Computer Graphics. Springer, pp. 455–465.
Rouhani, M., Lafarge, F., Alliez, P., 2017. Semantic segmentation of 3d textured meshes for urban scene analysis. ISPRS J. Photogramm. Remote Sens. 124–139.
Schult, J., Engelmann, F., Kontogianni, T., Leibe, B., 2020. DualConvMesh-net: Joint geodesic and euclidean convolutions on 3D meshes. In: Conf. on Comp. Vis. and Patt. Recog.. CVPR.
Sharma, R., Saqib, M., Lin, C., Blumenstein, M., 2022. A survey on object instance segmentation. SN Comput. Sci. 1–23.
Simonovsky, M., Komodakis, N., 2017. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3693–3702.
Song, S., Lichtenberg, S.P., Xiao, J., 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE Computer Society, pp. 567–576.
Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., Han, S., 2020. Searching efficient 3d architectures with sparse point-voxel convolution. In: European Conference on Computer Vision. Springer, pp. 685–702.
Tang, R., Xia, M., Yang, Y., Zhang, C., 2022. A deep-learning model for semantic segmentation of meshes from uav oblique images. Int. J. Remote Sens. 4774–4792.
Tatarchenko, M., Park, J., Koltun, V., Zhou, Q.Y., 2018. Tangent convolutions for dense prediction in 3d. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3887–3896.
Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L., 2019. Kpconv: Flexible and deformable convolution for point clouds. In: 2019 IEEE/CVF International Conference on Computer Vision. ICCV, IEEE Computer Society, pp. 6410–6419.
Tutzauer, P., Laupheimer, D., Haala, N., 2019. Semantic urban mesh enhancement utilizing a hybrid model. ISPRS Ann. Photogram., Remote Sens. Spatial Inf. Sci. 4, 175–182.
Ulku, I., Akagündüz, E., 2022. A survey on deep learning-based architectures for semantic segmentation on 2d images. Appl. Artif. Intell. 1–45.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems. Vol. 30.
Verdie, Y., Lafarge, F., Alliez, P., 2015. Lod generation for urban scenes. ACM Trans. Graph. 34, 1–14.
Vu, H.H., Labatut, P., Pons, J.P., Keriven, R., 2011. High accuracy and visibility-consistent dense multiview stereo. IEEE Trans. Pattern Anal. Mach. Intell. 88, 9–901.
Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M., 2019. Dynamic graph cnn for learning on point clouds. Acm Trans. Graph. (Tog) 38, 1–12.
Wang, H., Zhang, J., 2022. A survey of deep learning-based mesh processing. Commun. Math. Stat. 10, 163–194.
Weixiao, G., Nan, L., Boom, B., Ledoux, H., 2023. Pssnet: Planarity-sensible semantic segmentation of large-scale urban meshes. ISPRS J. Photogramm. Remote Sens. 196, 32–44.
Wilk, Ł., Mielczarek, D., Ostrowski, W., Dominik, W., Krawczyk, J., 2022. Semantic urban mesh segmentation based on aerial oblique images and point clouds using deep learning. Int. Arch. Photogram., Remote Sens. Spatial Inf. Sci..
Wu, C., Chen, M., Wu, D., Ma, J., Xu, J., Ma, B., 2021. Work-in-progress- design method of a real-time monitoring system for ict evaluation process in education based on cesiumjs 3d visualization. In: 2021 7th International Conference of the Immersive Learning Research Network. ILRN, IEEE, pp. 1–3.
Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y., 2020. A comprehensive survey on graph neural networks. IEEE Trans. Neural Netw. Learn. Syst. 32, 4–24.
Xiong, Z., Zhang, F., Wang, Y., Shi, Y., Zhu, X.X., 2022. Earthnets: Empowering ai in earth observation. arXiv preprint arXiv:2210.04936.
Yang, Y., Liu, S., Pan, H., Liu, Y., Tong, X., 2020. Pfcnn: Convolutional neural networks on 3d surfaces using parallel frames. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13578–13587.
Yu, B., Fan, Z., 2020. A comprehensive review of conditional random fields: variants, hybrids and applications. Artif. Intell. Rev. 53, 4289–4333.
Zhang, R., Li, G., Wunderlich, T., Wang, L., 2021. A survey on deep learning-based precise boundary recovery of semantic segmentation for images and point clouds. Int. J. Appl. Earth Obs. Geoinf. 102, 102411.
Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, pp. 6230–6239.
Zhu, L., Shen, S., Gao, X., Hu, Z., 2018. Large scale urban scene modeling from mvs meshes. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 614–629.
Zhu, L., Shen, S., Hu, L., Hu, Z., 2017. Variational building modeling from urban mvs meshes. In: 2017 International Conference on 3D Vision (3DV). IEEE, pp. 318–326.