

JSENet: Joint Semantic Segmentation and Edge Detection Network for 3D Point Clouds

Zeyu HU¹[0000-0003-3585-7381], Mingmin Zhen¹[0000-0002-8180-1023], Xuyang BAI¹[0000-0002-7414-0319], Hongbo Fu²[0000-0002-0284-726X], and Chiew-lan Tai¹

¹ Hong Kong University of Science and Technology

{zhuam,mzhen,xbaiad,taicl}@cse.ust.hk

² City University of Hong Kong

hongbofu@cityu.edu.hk

Abstract. Semantic segmentation and semantic edge detection can be seen as two dual problems with close relationships in computer vision. Despite the fast evolution of learning-based 3D semantic segmentation methods, little attention has been drawn to the learning of 3D semantic edge detectors, even less to a joint learning method for the two tasks. In this paper, we tackle the 3D semantic edge detection task for the first time and present a new two-stream fully-convolutional network that jointly performs the two tasks. In particular, we design a joint refinement module that explicitly wires region information and edge information to improve the performances of both tasks. Further, we propose a novel loss function that encourages the network to produce semantic segmentation results with better boundaries. Extensive evaluations on S3DIS and ScanNet datasets show that our method achieves on par or better performance than the state-of-the-art methods for semantic segmentation and outperforms the baseline methods for semantic edge detection.

Keywords: Semantic Segmentation, Semantic Edge Detection, 3D Point Clouds, 3D Scene Understanding

1 Introduction

Semantic segmentation (SS) and semantic edge detection (SED) are two fundamental problems for scene understanding. The former aims to parse a scene and assign a class label to each pixel in images or each point in 3D point clouds. The latter focuses on detecting edge pixels or edge points and classifying each of them to one or more classes. Interestingly, the SS and the SED tasks can be seen as two dual problems with even interchangeable outputs in an ideal case (see Fig. 1). While SS has been extensively studied in both 2D and 3D [1-7], SED has only been explored in 2D [8-11], to our best knowledge.

There are strong motivations to address both problems in a joint learning framework. On one hand, previous SS models tend to struggle in edge areas since these areas constitute only a small part of the whole scene and thus have little

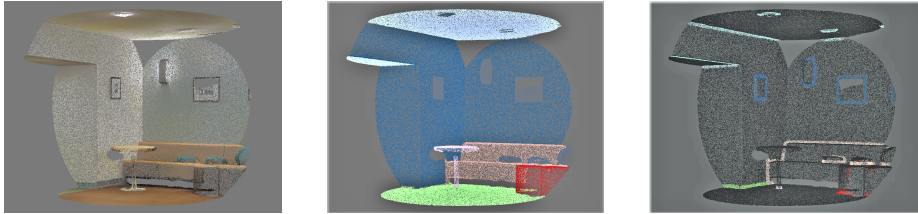


Fig. 1: (Left) Point cloud of a real-world scene from S3DIS [12]; (Middle) Semantic segmentation point (SSP) mask; (Right) Semantic edge point (SEP) map. For visualization, we paint an edge point to the color of one of its class labels.

effect on the supervision during training [13, 14]. Performing the complementary SED task simultaneously may help the network sharpen the boundaries of the predicted SS results [15]. On the other hand, existing SED models are easily affected by non-semantic edges in the scene [8, 11], while a trained SS model is less sensitive to those edges [16]. Information from the SS model thus may help SED models suppress network activations on the non-semantic edges. Despite the close relationships of the two tasks, there is no existing work tackling them jointly as far as we know.

Although not focusing on the SS and the SED tasks, in 2D, several existing works have already made fruitful attempts at proposing joint learning methods for complementary segmentation and edge detection tasks [2, 17–20, 14, 15]. These works often treat the two tasks naïvely by sharing parts of their networks and limit the interactions between them to the feature domain. Predicted segmentation masks and edge maps are used only for loss calculation and do not contribute to each other. Strong links between the outputs of the two tasks are not fully exploited.

In this work, we introduce the task of SED into the 3D field and propose *JSENet*, a new 3D fully-convolutional network (FCN) for joint SS and SED. The proposed network consists of two streams and a joint refinement module on top of them. Specifically, we use a classical encoder-decoder FCN for SS in one stream, which outputs semantic segmentation point (SSP) masks, and add an SED stream outputting semantic edge point (SEP) maps in parallel. The key to our architecture is the lightweight joint refinement module. It takes the output SSP masks and SEP maps as inputs and jointly refines them by explicitly exploiting the duality between them. Moreover, we further propose a novel loss function to encourage the predicted SSP masks to better align with the ground truth semantic edges.

To summarize, our contributions are threefold:

1. We introduce the task of SED into the 3D field and design an FCN-based architecture with enhanced feature extraction and hierarchical supervision to generate precise semantic edges.

2. We propose a novel framework for joint learning of SS and SED, named JSENet, with an effective lightweight joint refinement module that brings improvements by explicitly exploiting the duality between the two tasks.
3. We propose a dual semantic edge loss, which encourages the network to produce SS results with finer boundaries.

To build our FCN network in 3D, we resort to KPConv [21], a recently proposed kernel-based convolution method for point clouds, for its state-of-the-art performance and ease of implementation. We conduct extensive experiments to demonstrate the effectiveness of our method. Since there is no existing 3D SED dataset, we construct a new 3D SED benchmark using S3DIS [12] and ScanNet [7] datasets. We achieve the state-of-the-art performance for the SS task with IoU scores of 67.7% on S3DIS Area-5 and 69.9% on ScanNet test set. For the SED task, our method outperforms the baseline methods by a large margin.

2 Related Work

2.1 3D Semantic Segmentation

Based on different data representations, 3D SS methods can be roughly divided into three categories: multiview image-based, voxel-based, and point-based. Our method falls into the point-based category.

Although the *multiview image-based* methods easily benefit from the success of 2D CNN [22, 23], for SS, they suffer from occluded surfaces and density variations, and rely heavily on viewpoint selection. Based on powerful 3D CNN [24–30], *voxel-based* methods achieve the best performance on several 3D SS datasets, but they need intensive computation power.

Compared with the previously mentioned methods, *point-based* methods suffer less from information loss and thus achieve high point-level accuracy with less computation power consumption [5, 6]. They can be generally classified into four categories: neighboring feature pooling [31–34], graph construction [35–38], attention-based aggregation [39], and kernel-based convolution [40–46, 21]. Among all the point-based methods, the recently proposed kernel-based method KPConv [21] achieves the best performance for efficient 3D convolution. Thus, we adopt KPConv to build our backbone and refinement network.

2.2 2D Semantic Edge Detection

Learning-based SED dates back to the early work of Prasad et al. [47]. Later, Hariharan et al. [48] introduced the first Semantic Boundaries Dataset. After the dawn of deep learning, the HFL method [49] builds a two-stage prediction process by using two deep CNNs for edge localization and classification, respectively. More recently, CASENet [8] extended the CNN-based class-agnostic edge detector HED [50] to a class-aware semantic edge detector by combining low- and high-level features with a multi-label loss function for supervision. Later,

several follow-up works [9–11, 51] improved CASENet by adding diverse deep supervision and reducing annotation noises.

In 2D images, semantic edges of different classes are weakly related since they are essentially occlusion boundaries of projected objects. Based on this observation, 2D SED methods treat SED of different classes as independent binary classification problems and utilize structures that limit the interaction between different classes like group convolution modules. Unlike the ones in 2D, semantic edges in 3D are physical boundaries of objects and thus are highly related to each other. In this work, we study the problem of 3D SED for the first time. We adopt from 2D methods the idea of extracting enhanced features and construct a network that does not limit the interaction between different classes.

2.3 Joint Learning of Segmentation and Edge Detection

For 2D images, several works have explored the idea of combining networks for complementary segmentation and edge detection tasks to improve the learning efficiency, prediction accuracy, and generalization ability [17–20, 14, 15, 2].

To be more specific, for salient object detection, researchers have exploited the duality between the binary segmentation and *class-agnostic* edge detection tasks [20, 15]. As for SS, such *class-agnostic* edges are used to build semantic segmentation masks with finer boundaries [2, 19, 14, 18, 17]. In contrast, we tackle the problem of joint learning for SS and *class-aware* SED. Furthermore, unlike previous works, which limit the interactions between segmentation and edge detection to the sharing of features and network structures, our method exploits the close relationship between SSP masks and SEP maps.

3 JSENet

In this section, we present our JSENet architecture for the joint learning of SS and class-aware SED. As depicted in Fig. 2, our architecture consists of two streams of networks with a shared feature encoder and followed by a joint refinement module. The first stream of the network, the SS stream, is a standard encoder-decoder FCN for SS using the same structure as presented in KPConv [21]. The second stream, the SED stream, is another FCN with enhanced feature extraction and hierarchical supervision. We then fuse the outputs from the two streams using our carefully designed joint refinement module to produce refined SSP masks and refined SEP maps. Next, we will describe each of the modules in detail and then explain the supervisions used for joint learning.

3.1 Semantic Segmentation Stream

We denote the SS stream as $\mathcal{S}_\theta(\mathcal{P})$ with parameters θ , taking a point cloud $\mathcal{P} \in \mathbb{R}^{N \times 6}(x, y, z, r, g, b)$ with N points as input and outputting an SSP mask. More specifically, for a segmentation prediction of K semantic classes, it outputs a categorical distribution $s \in \mathbb{R}^{N \times K}$: $s(p|\mathcal{P}, \theta)$ representing the probability of

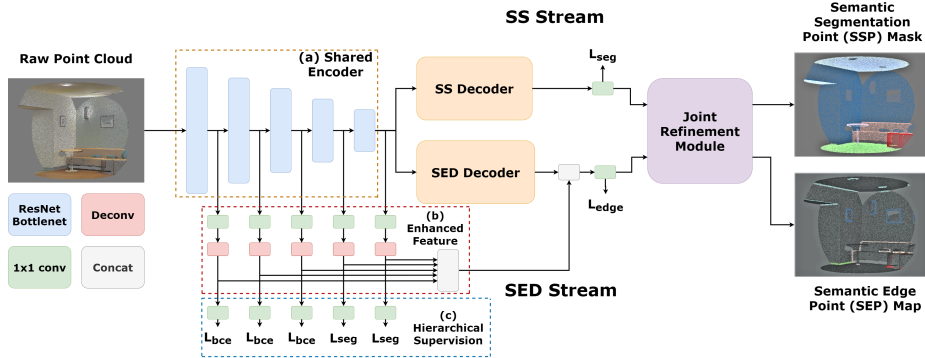


Fig. 2: JSENet architecture. Our architecture consists of two main streams. The SS stream can be any fully-convolutional network for SS. The SED stream extracts enhanced features through a skip-layer structure and is supervised by multiple loss functions. A joint refinement module later combines the information from the two streams and outputs refined SSP masks and SEP maps.

point p belonging to each of the K classes. We supervise this stream with the standard multi-class cross-entropy loss (L_{seg} in Fig. 2). The SS stream can be any feedforward 3D fully-convolutional SS network such as InterpCNNs [46] and SSCNs [29]. In this work, we adopt KPConv [21] and use it as our backbone network for its efficient convolution operation and ability to build complex network structures. There are two types of KPConv: rigid and deformable. We use the rigid version in this work for its better convergence performance. In order to demonstrate the improvements introduced by joint learning, we use the same structure as described in KPConv.

3.2 Semantic Edge Detection Stream

We denote the SED stream as $\mathcal{E}_\phi(\mathcal{P})$ with parameters ϕ , taking a point cloud \mathcal{P} with N points as input and outputting SEP maps. For K defined semantic categories, this stream outputs K SEP maps $\{e_1, \dots, e_K\}$, each having the same size as \mathcal{P} with one channel. We denote $e_k(p|\mathcal{P}, \phi)$ as the network output, which indicates the computed edge probability of the k -th semantic category at point p . Note that one point may belong to multiple categories.

The SED stream shares the same feature encoder with the SS stream to force the network to prefer representations with a better generalization ability for both tasks. Our backbone is an FCN-based structure. However, one major drawback of an FCN-based structure is the loss in spatial information of the output after propagating through several alternated convolutional and pooling layers. This drawback would harm the localization performance, which is essential for our SED task. Besides, according to the findings of CASENet [8], the low-level features of CNN-based encoder are not suitable for semantic classification (due to the limited receptive fields) but are able to help augment top classifications by

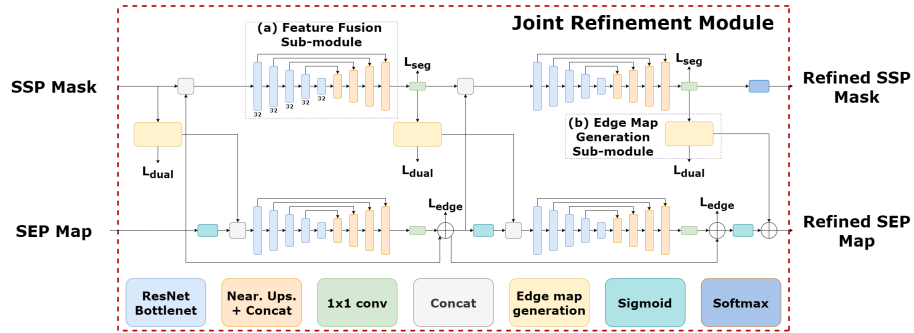


Fig. 3: Illustration of the joint refinement module, which consists of two branches.

suppressing non-edge activations and providing detailed edge localization information. Based on these observations, we propose to extract enhanced features with hierarchical supervisions to alleviate the problem of spatial information loss and offer richer semantic cues for final classification.

Enhanced feature extraction. In detail, we extract the feature maps generated by the layers of the shared encoder (Fig. 2(a)). We then reduce the numbers of their feature channels and deconvolve them to the size of the input point cloud. The features of different layers participate and augment the final SED through a skip-layer architecture, as shown in Fig. 2(b).

Hierarchical supervision. As shown in Fig. 2(c), from the extracted feature maps of the first three layers, we generate binary edge point maps indicating the probability of points belonging to the semantic edges of any classes. From the last two layers, we generate two SSP masks. All binary edge point maps are supervised by the weighted binary cross-entropy loss (L_{bce}) using ground-truth (GT) binary edges obtained from the GT SSP masks. The two SSP masks are supervised by the standard multi-class cross-entropy loss (L_{seg}). The output SEP maps of the SED stream are supervised by a weighted multi-label loss (L_{edge}) following the idea from CASENet [8]. We will give details about different loss functions in Section 3.4.

3.3 Joint Refinement Module

We denote the joint refinement module as $\mathcal{R}_\gamma(s, e_1, \dots, e_K)$ with parameters γ , taking as input the SSP mask s coming from the SS stream and the SEP maps $\{e_1, \dots, e_K\}$ generated by the SED stream. As shown in Fig. 3, we construct a two-branch structure with simple feature fusion sub-modules (Fig. 3(a)) and novel edge map generation sub-modules (Fig. 3(b)). The upper branch is responsible for segmentation refinement, and the lower one is for edge refinement. We feed different joint features (described in detail below) to the feature fusion sub-modules of the two branches and generate refined SSP masks and SEP maps.

Feature fusion sub-module. To fuse the region and edge features, we construct two simple U-Net [52] like feature fusion sub-modules for each refinement branch. In detail, each feature fusion sub-module consists of five encoding layers with channel sizes of 32 for all layers. For segmentation refinement, the feature fusion sub-module takes the concatenated SSP mask $s \in \mathbb{R}^{N \times K}$ and SEP maps $\{e_1, \dots, e_K\}$ ($e_i \in \mathbb{R}^N$) as input and outputs a refined SSP mask directly. As for edge refinement, we find that adjusting the activation values of the SEP maps is more effective than asking the neural network to output refined SEP maps naively. Thus, we put the unrefined SEP maps through a sigmoid operation and concatenate them with the edge activation point maps $\{a_1, \dots, a_K\}$ ($a_i \in \mathbb{R}^N$) generated by the edge map generation sub-module. The concatenated features are then fed to the feature fusion sub-module to generate auxiliary point maps, which are added to the unrefined SEP maps to adjust the edge activation values.

Edge map generation sub-module. In order to exploit the duality between two tasks, we design an edge map generation sub-module to convert an SSP mask to edge activation point maps. More formally, we denote the edge map generation sub-module as $\mathcal{G}(s)$, which takes a categorical distribution $s \in \mathbb{R}^{N \times K}$ as input and outputs edge activation point maps $\{a_1, \dots, a_K\}$ ($a_i \in \mathbb{R}^N$), where $a_i(p)$ is a potential that represents whether a particular point p belongs to the semantic boundary of the i -th class. It is computed by taking a spatial derivative on the segmentation output as follows:

$$a_i = col_i(|M * Softmax(s) - Softmax(s)|), \quad (1)$$

where col_i denotes the i -th column and M denotes the Mean filter, which takes neighboring points within a small radius. As illustrated in Fig. 4, the i -th column of the output tensor of a softmax operation represents an activation point mask for class i , where a higher value indicates a higher probability of belonging to class i . On this mask, points near to the boundaries of class i will have neighbors with activation values of significant differences, and points far from the boundaries will have neighbors with similar activation values. Thus, after the mean filtering and subtraction, points nearer to the predicted boundaries will have larger activation values. The converted edge activation point maps are fed to the edge refinement branch and utilized for loss calculation as well.

Supervisions. For supervision, the output SSP masks are supervised by the multi-class cross-entropy loss (L_{seg}), and the output SEP maps are supervised by the weighted multi-label loss (L_{edge}). Additionally, we supervise the generated edge activation point maps with our proposed dual semantic edge loss (L_{dual}) to encourage the predicted SS results to align with the GT semantic edges correctly (see Section 3.4). After refinement, the final output SSP mask is normalized by a softmax operation. The final output SEP maps are normalized by a sigmoid operation and added element-wise with the final edge activation point maps.

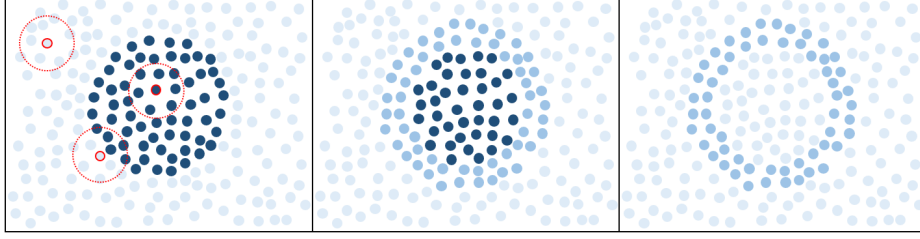


Fig. 4: Illustration of the edge map generation process on 2D points. (Left) An activation point mask, with dark colors representing high activation values. Three red points represent three different situations: far from the activated region, near the boundary, and within the activated region. (Middle) The activation point mask after the mean filtering; (Right) The generated edge activation point map.

3.4 Joint Multi-task Learning

The key to joint learning of the SS and the SED tasks is to design a proper supervision signal. The total loss is formulated as:

$$L_{total} = \lambda_0 L_{seg} + \lambda_1 L_{edge} + \lambda_2 L_{bce} + \lambda_3 L_{dual}. \quad (2)$$

During training, λ_0 is set to the number of semantic classes to balance the influences of the two tasks. The other weights are set to 1. We describe in detail all the loss functions used for supervision in our framework below.

Multi-class cross-entropy loss. We use a standard multi-class cross-entropy loss, which denoted as L_{seg} , on a predicted SSP mask s :

$$L_{seg}(\hat{s}, s) = - \sum_k \sum_p \hat{s}_k(p) \log(s_k(p)), \quad (3)$$

where $\hat{s} \in \mathbb{R}^{N \times K}$ denotes GT semantic labels in a one-hot form.

Weighted multi-label loss. Following the idea proposed in CASENet [8], we address the SED problem for a point cloud by a multi-label learning framework and implement a point-wise weighted multi-label loss L_{edge} . Suppose an input point cloud \mathcal{P} has K SEP maps $\{e_1, \dots, e_K\}$ ($e_i \in \mathbb{R}^N$) predicted by the network and K label point maps $\{\hat{e}_1, \dots, \hat{e}_K\}$ ($\hat{e}_i \in \mathbb{R}^N$), where \hat{e}_k is a binary point map indicating the ground truth of the k -th class semantic edges. The point-wise weighted multi-label loss L_{edge} is formulated as:

$$L_{edge}(\{\hat{e}_1, \dots, \hat{e}_K\}, \{e_1, \dots, e_K\}) = \sum_k \sum_p \{-\beta_k \hat{e}_k(p) \log(e_k(p)) - (1 - \beta_k)(1 - \hat{e}_k(p)) \log(1 - e_k(p))\}, \quad (4)$$

where β_k is the percentage of non-edge points in the point cloud of the k -th class to account for the skewness of sample numbers.

Weighted binary cross-entropy loss. To supervise the generated binary edge point maps in the SED stream, we implement a point-wise weighted binary cross-entropy loss L_{bce} . Let b denote the predicted binary edge point map, and \hat{b} the GT binary edge point map converted from the GT SEP maps. The point-wise weighted cross-entropy loss is defined as:

$$L_{bce}(\hat{b}, b) = \sum_p \{-\beta \hat{b}(p) \log(b(p)) - (1 - \beta)(1 - \hat{b}(p)) \log(1 - b(p))\}, \quad (5)$$

where β is the percentage of non-edge points among all classes.

Dual semantic edge loss. As mentioned above, inspired by the duality between SS and SED, we design an edge map generation sub-module to convert the predicted SSP mask $s \in \mathbb{R}^{N \times K}$ to edge activation point maps $\{a_1, \dots, a_K\}$ ($a_i \in \mathbb{R}^N$) (c.f., Equation 1). In a similar way, we can compute GT edge activation point maps $\{\hat{a}_1, \dots, \hat{a}_K\}$ from the GT semantic labels \hat{s} :

$$\hat{a}_i = \text{col}_i(|M * \text{One_hot}(\hat{s}) - \text{One_hot}(\hat{s})|). \quad (6)$$

Note that the softmax operation for predicted SSP mask s is changed to the one-hot encoding operation for GT semantic labels \hat{s} . Taking the converted GT edge activation point maps, we can define the loss function as follows:

$$L_{dual}(\{\hat{a}_1, \dots, \hat{a}_K\}, \{a_1, \dots, a_K\}) = \sum_k \sum_p \beta (|\hat{a}_k(p) - a_k(p)|), \quad (7)$$

where β is the same weight as above. Intuitively, the network will get penalized when there are mismatches on edge points. It is worth noting that the loss function will not be dominated by the non-edge points since the calculated loss values on these points are zeros or very small numbers. The above dual loss is naturally differentiable and exploits the duality between SS and SED.

4 Experiments

To demonstrate the effectiveness of our proposed method, we now present various experiments conducted on the S3DIS [12] and ScanNet[7] datasets, for which GT SSP masks are available and we can generate GT SEP maps from them easily. We first introduce the dataset preparation and evaluation metrics in Section 4.1, and then present the implementation details for reproduction in Section 4.2. We report the results of our ablation studies in Section 4.3, and the results on the S3DIS and ScanNet datasets in Sections 4.4.

4.1 Datasets and Metrics

We use S3DIS [12] and ScanNet [7] datasets for our experiments. There are two reasons for choosing these datasets: 1) they are both of high quality and 2) semantic edges are better defined on indoor data: compared to existing 3D

outdoor datasets, in indoor scenes, more detailed semantic labels are defined and objects are more densely connected. The S3DIS dataset consists of 3D point clouds for six large-scale indoor areas captured from three different buildings. It has around 273 million points annotated with 13 semantic classes. The ScanNet dataset includes 1513 training scenes and 100 test scenes in a mesh format, all annotated with 20 semantic classes, for online benchmarking.

To generate the GT SEP maps, for S3DIS, we directly check the neighbors within a $2cm$ radius of each point in the dataset. For a specific point, if it has neighbors with different semantic labels, we label it as an edge point of all the semantic classes that appear in its neighborhood. As for the ScanNet dataset, following the work of KPConv [21], we first rasterize the training meshes by uniformly sampling points on the faces of meshes and then downsample the generated point clouds with $1cm$ grids. We use these point clouds for training and generate the GT SEP maps using the same way as described for the S3DIS dataset. During testing, we project the semantic edge labels to the vertices of the original meshes and test directly on meshes.

To evaluate the performance of SS and SED, we adopt the standard mean intersection over union (mIoU) for SS and use the mean maximum F-measure (MF) at the optimal dataset scale (ODS) for SED following the works in 2D [8–11]. We generate thicker edges for point clouds than for images since a point cloud is much sparser than an image. Since we have thicker edges, the localization tolerance used in the 2D case is not introduced to our evaluation.

4.2 Implementation Details

In this section, we discuss the implementation details for our experiments. JSENet is coded in Python and TensorFlow. All the experiments are conducted on a PC with 8 Intel(R) i7-7700 CPUs and a single GeForce GTX 1080Ti GPU.

Training. Since the 3D scenes in both datasets are of huge size, we randomly sample spheres with $2m$ radius in the training set and augment them with Gaussian noise, random scaling, and random rotation. Following the settings in KPConv, the input point clouds are downsampled with a grid size of $4cm$. In all our experiments, unless explicitly stated otherwise, we use a Momentum gradient descent optimizer with a momentum of 0.98 and an initial learning rate of 0.01. The learning rate is scheduled to decrease exponentially. In particular, it is divided by 10 for every 100 epochs. Although the framework is end-to-end trainable, in order to clearly demonstrate the efficacy of the proposed joint refinement module, we first train our network without the joint refinement module for 350 epochs and then optimize the joint refinement module alone with the other parts fixed for 150 epochs.

Testing. Similar to the training process, during testing, we sample spheres with $2m$ radius from the testing set regularly and ensure each point to be sampled for multiple times. The predicted probabilities for each point are averaged through a voting scheme [21]. All predicted values are projected to the original point clouds (S3DIS) or meshes (ScanNet) for evaluation.

Table 1: Ablation experiments of network structures on S3DIS Area-5. **SEDS**: semantic edge detection stream; **EFE**: enhanced feature extraction; **HS**: hierarchical supervision; **SSS**: semantic segmentation stream; **JRM**: joint refinement module. The results in some cells (with ‘-’) are not available, since the corresponding models perform either SS or SED.

0	SEDS	EFE	HS	SSS	JRM	mIoU (%)	mMF (ODS)(%)
1	✓	✓	✓	✓	✓	67.7	31.0
2	✓	✓	✓	✓		66.2	30.5
3				✓		64.7	-
4	✓	✓	✓			-	30.2
5	✓	✓				-	29.9
6	✓					-	29.4

4.3 Ablation Study

In this section, we compare the performances of JSENet under different settings on the S3DIS dataset since it is originally presented in a point cloud format and all semantic labels are available. Following the common setting [5, 53, 21, 3, 37, 54, 30], we use Area-5 as a test scene and train our network on the other scenes. All experiments are conducted keeping all hyperparameters the same.

Network structures. In Table 1, we evaluate the effectiveness of each component of our method. For the SS task, as shown in the table (Row 3), the performance of training our SS stream alone is 64.7% in terms of mIoU. Our SS stream shares the same architecture with KPConv and the reported score of KPConv is 65.4% in their paper. By naïvely combining the SS stream and the SED stream, we can improve the SS task by 1.5% (Row 2). From the joint refinement module, we further gain about 1.5% (Row 1) improvement in performance. We achieve about 3% improvement comparing to training our SS stream alone and still more than 2% improvement comparing to the result reported in KPConv.

For the SED task, it can be seen from the table that the performance of training the SED stream alone without the enhanced feature extraction and hierarchical supervision is 29.4% (Row 6) in terms of mMF (ODS). We gain about 0.5% and 0.3% improvements in performance from the enhanced feature extraction (Row 5) and hierarchical supervision (Row 4), respectively. Naïvely combining the SS stream and the SED stream brings a further improvement of 0.3% (Row 2) in terms of mMF. By adding the joint refinement module, we can further improve the SED task by 0.5% (Row 1).

Choice of loss functions for hierarchical supervision. To justify our choices of the loss functions in the SED stream for hierarchical supervision, we test the performances of SED using different settings of supervision. As shown in Table 2a, if all hierarchical supervisions are removed, the performance of our SED model will decrease by 0.3%. Among all the choices listed in the table, the one used in our network achieves the best result.

Table 2: (a) Comparison of different supervision choices for SED. (b) Effects of the dual semantic edge loss in terms of boundary quality (F-score).

(a)		(b)	
Method	mMF (ODS) (%)	Method	F-score (%)
L_{bce} for all five layers	30.1	JSENet w/o dual loss	22.7
L_{seg} for all five layers	30.1	JSENet	23.1
No hierarchical supervision	29.9		
L_{bce} for first three, L_{seg} for last two	30.2		

Efficacy of dual semantic edge loss. We further showcase the effects of the dual semantic edge loss in terms of F-score for edge alignment of the predicted SSP masks in Table 2b. We train our network without the dual semantic edge losses for the edge map generation sub-modules as the baseline. It is shown that the dual semantic edge losses bring an improvement of 0.4% in terms of F-score.

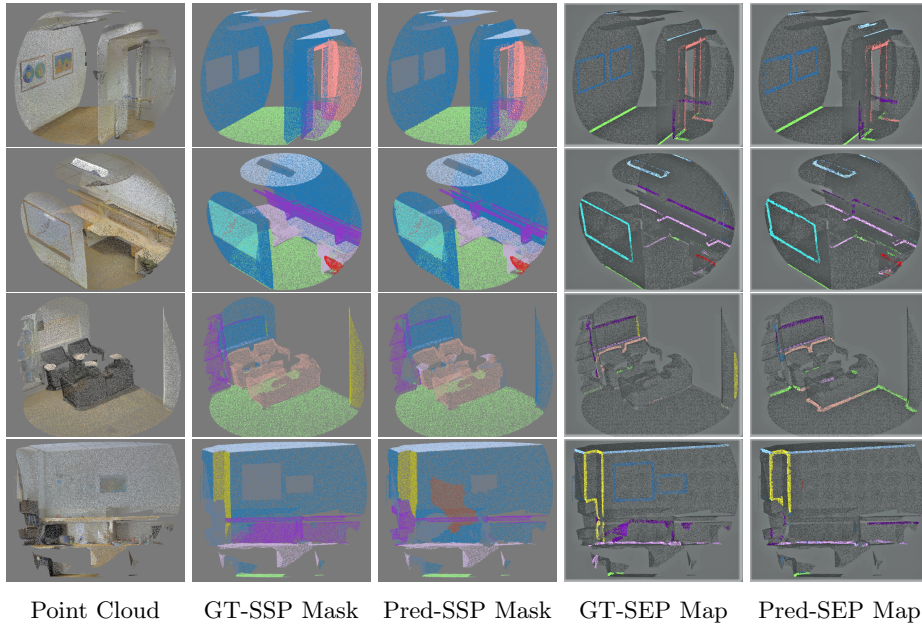


Fig. 5: Qualitative results on S3DIS Area-5.

4.4 Results on S3DIS & ScanNet Datasets

To compare JSENet with the state-of-the-arts, for SS, we choose the latest methods [3-5, 37, 30, 53-64, 42, 29] with reported results on the S3DIS or the ScanNet

Table 3: mIoU scores (%) of semantic segmentation on S3DIS Area-5 and ScanNet test set.

Method	S3DIS	ScanNet
TangentConv [54]	52.6	43.8
RNN Fusion [55]	53.4	-
SPGraph [56]	58.0	-
FCPN [57]	-	44.7
PointCNN [3]	57.3	45.8
ParamConv [58]	58.3	-
PanopticFusion [59]	-	52.9
TextureNet [60]	-	56.6
SPH3D-GCN [61]	59.5	61.0
HPEIN [37]	61.9	61.8
MCCNN [62]	-	63.3
MVPNet [4]	62.4	64.1
PointConv [42]	-	66.6
KPConv rigid [21]	65.4	68.6
KPConv deform [21]	67.1	68.4
SparseConvNet [29]	-	72.5
MinkowskiNet [30]	65.4	73.6
JSENet (ours)	67.7	69.9

datasets as our baselines. For SED, since we cannot find any existing solutions in 3D, we extend CASENet [8], which is the state-of-the-art method for 2D SED to 3D for comparison. Besides, we build another baseline network using the same structure as the one presented in KPConv with a changed output layer.

For the S3DIS dataset, we use the same *train-test* splitting setting as in Section 4.3. For the ScanNet dataset, since it is built for online benchmarking, the GT semantic labels for its test set are not available. Thus, for the SS task, we train our network using the 1513 training scenes and report the testing results on the 100 test scenes following the common setting. To evaluate the task of SED, as explained in Section 4.1, we obtain semantic edge labels from the training set and divide the 1513 training scenes into a new training set with 1201 scenes and a new test set with 312 scenes following the *train-val* splitting file offered by ScanNet. All the compared networks for SED are trained on the new training set and tested on the new test set. Qualitative results are shown in Fig. 5.

Results of SS task. The results for the SS task are reported in Table 3. The detailed IoU scores of each class for the S3DIS dataset can be found in the supplementary. The details for the ScanNet dataset can be found on the ScanNet benchmarking website³. For the S3DIS dataset, JSENet achieves a 67.7% mIoU score and outperforms all the baseline methods. For the ScanNet dataset, JSENet ranks third with a 69.9% mIoU score. The first two (i.e., SparseConvNet and MinkowskiNet) are two voxel-based methods that require intense computation power. Compared to other point-based methods, JSENet achieves the best performance and consistently outperforms the results of KPConv.

Results of SED task. We present the results for the SED task in Tables 4 and 5. It can be seen that our method outperforms the two baseline methods

³ http://kaldir.vc.in.tum.de/scannet_benchmark/semantic_label_3d

Table 4: MF (ODS) scores (%) of semantic edge detection on S3DIS Area-5.

Method	mean	ceil.	floor	wall	beam	col.	wind.	door	chair	table	book.	sofa	board	clut.
CASENet [8]	27.1	46.5	49.0	33.3	0.2	21.9	12.6	22.6	36.9	33.6	21.8	25.1	22.6	26.1
KPConv [21]	29.4	43.7	41.8	36.4	0.2	23.6	13.4	29.7	39.8	37.3	26.6	29.4	29.2	31.3
JSENet (ours)	31.0	44.5	43.2	38.8	0.2	24.1	13.2	36.7	37.7	36.3	29.1	34.0	33.3	32.4

on both datasets. We find that the extended CASENet architecture performs worse than KPConv with the changed output layer. This result supports our previous argument (Section 2.2) that 3D semantic edges have stronger relevance between different classes (since they are physical boundaries of objects and the 2D semantic edges are occlusion boundaries of projected objects), and thus the network designs that enforce limitations between interactions of different classes would harm the edge detection performance.

Table 5: MF (ODS) scores (%) of semantic edge detection on ScanNet val set.

Method	mean	bath	bed	bksf	cab	chair	cntr	curt	desk	door	floor	othr	pic	ref	show	sink	sofa	tab	toil	wall	wind
CASENet [8]	32.3	38.2	55.9	29.9	36.0	36.0	36.8	28.1	28.5	19.1	26.6	25.1	32.2	28.6	26.6	23.9	19.0	45.7	27.1	54.6	27.4
KPConv [21]	34.8	40.5	55.9	33.6	38.5	39.3	38.0	32.9	31.2	22.0	25.1	28.5	36.2	30.8	30.9	22.7	22.8	46.5	33.3	55.2	32.3
JSENet (ours)	37.3	43.8	55.8	35.9	38.2	41.0	40.8	34.5	35.9	25.5	28.7	29.5	37.3	36.2	31.7	28.1	28.3	48.5	35.6	53.2	37.8

5 Conclusions

In this paper, we proposed a joint semantic segmentation and semantic edge detection network (JSENet), which is a new two-stream fully-convolutional architecture with a lightweight joint refinement module that explicitly wires region information and edge information for output refinement. We constructed a two-branch structure with simple feature fusion sub-modules and novel edge map generation sub-modules for the joint refinement module and designed the dual semantic edge loss that encourages the network to produce sharper predictions around object boundaries. Our experiments show that JSENet is an effective architecture that produces SSP masks and SEP maps of high qualities. JSENet achieves the state-of-the-art results on the challenging S3DIS and ScanNet datasets, significantly improving over strong baselines. For future works, one straightforward direction is to explore the potential of joint improvement of these two tasks in the 2D field. Moreover, we notice that our SED method is easily affected by the noises in the GT labels introduced by human annotation. We believe that future works with special treatments on the misaligned GT semantic edges will further improve the performance of the SED task.

Acknowledgements. This work is supported by Hong Kong RGC GRF 16206819, 16203518 and Centre for Applied Computing and Interactive Media (ACIM) of School of Creative Media, City University of Hong Kong.

References

1. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 3431–3440
2. Takikawa, T., Acuna, D., Jampani, V., Fidler, S.: Gated-scnn: Gated shape cnns for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 5229–5238
3. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B.: Pointcnn: Convolution on x-transformed points. In: Advances in neural information processing systems. (2018) 820–830
4. Jaritz, M., Gu, J., Su, H.: Multi-view pointnet for 3d scene understanding. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2019) 0–0
5. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 652–660
6. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in neural information processing systems. (2017) 5099–5108
7. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE. (2017)
8. Yu, Z., Feng, C., Liu, M.Y., Ramalingam, S.: Casenet: Deep category-aware semantic edge detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 5964–5973
9. Liu, Y., Cheng, M.M., Fan, D.P., Zhang, L., Bian, J., Tao, D.: Semantic edge detection with diverse deep supervision (2018)
10. Yu, Z., Liu, W., Zou, Y., Feng, C., Ramalingam, S., Vijaya Kumar, B., Kautz, J.: Simultaneous edge alignment and learning. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 388–404
11. Acuna, D., Kar, A., Fidler, S.: Devil is in the edges: Learning semantic boundaries from noisy annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 11075–11083
12. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. (2016)
13. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 2881–2890
14. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 1925–1934
15. Wu, Z., Su, L., Huang, Q.: Stacked cross refinement network for edge-aware salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 7264–7273
16. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Learning a discriminative feature network for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 1857–1866

17. Cheng, D., Meng, G., Xiang, S., Pan, C.: Fusionnet: Edge aware deep convolutional networks for semantic segmentation of remote sensing harbor images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **10**(12) (2017) 5769–5783
18. Bertasius, G., Shi, J., Torresani, L.: Semantic segmentation with boundary neural fields. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2016) 3602–3610
19. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters—improve semantic segmentation by global convolutional network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2017) 4353–4361
20. Su, J., Li, J., Zhang, Y., Xia, C., Tian, Y.: Selectivity or invariance: Boundary-aware salient object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*. (2019) 3799–3808
21. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: *Proceedings of the IEEE International Conference on Computer Vision*. (2019) 6411–6420
22. Boulch, A., Le Saux, B., Audebert, N.: Unstructured point cloud semantic labeling using deep segmentation networks. *3DOR* **2** (2017) 7
23. Lawin, F.J., Danelljan, M., Tosteberg, P., Bhat, G., Khan, F.S., Felsberg, M.: Deep projective 3d semantic segmentation. In: *International Conference on Computer Analysis of Images and Patterns*, Springer (2017) 95–107
24. Roynard, X., Deschaud, J.E., Goulette, F.: Classification of point cloud scenes with multiscale voxel deep network. *arXiv preprint arXiv:1804.03583* (2018)
25. Ben-Shabat, Y., Lindenbaum, M., Fischer, A.: 3dmfv: Three-dimensional point cloud classification in real-time using convolutional neural networks. *IEEE Robotics and Automation Letters* **3**(4) (2018) 3145–3152
26. Le, T., Duan, Y.: Pointgrid: A deep network for 3d shape understanding. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2018) 9204–9214
27. Meng, H.Y., Gao, L., Lai, Y., Manocha, D.: Vv-net: Voxel vae net with group convolutions for point cloud segmentation (2018)
28. Riegler, G., Osman Ulusoy, A., Geiger, A.: Octnet: Learning deep 3d representations at high resolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2017) 3577–3586
29. Graham, B., Engelcke, M., van der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2018) 9224–9232
30. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Jun 2019)
31. Li, J., Chen, B.M., Hee Lee, G.: So-net: Self-organizing network for point cloud analysis. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2018) 9397–9406
32. Huang, Q., Wang, W., Neumann, U.: Recurrent slice networks for 3d segmentation of point clouds. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (Jun 2018)*
33. Zhao, H., Jiang, L., Fu, C.W., Jia, J.: Pointweb: Enhancing local neighborhood features for point cloud processing. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2019) 5565–5573

34. Zhang, Z., Hua, B.S., Yeung, S.K.: Shellnet: Efficient point cloud convolutional neural networks using concentric shells statistics. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 1607–1616
35. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)* **38**(5) (2019) 1–12
36. Wang, L., Huang, Y., Hou, Y., Zhang, S., Shan, J.: Graph attention convolution for point cloud semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 10296–10305
37. Jiang, L., Zhao, H., Liu, S., Shen, X., Fu, C.W., Jia, J.: Hierarchical point-edge interaction network for point cloud semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 10433–10441
38. Liu, J., Ni, B., Li, C., Yang, J., Tian, Q.: Dynamic points agglomeration for hierarchical point sets learning. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 7546–7555
39. Xie, S., Liu, S., Chen, Z., Tu, Z.: Attentional shapecontextnet for point cloud recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 4606–4615
40. Su, H., Jampani, V., Sun, D., Maji, S., Kalogerakis, E., Yang, M.H., Kautz, J.: Splatnet: Sparse lattice networks for point cloud processing. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (Jun 2018)
41. Hua, B.S., Tran, M.K., Yeung, S.K.: Pointwise convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 984–993
42. Wu, W., Qi, Z., Fuxin, L.: Pointconv: Deep convolutional networks on 3d point clouds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 9621–9630
43. Lei, H., Akhtar, N., Mian, A.: Octree guided cnn with spherical kernels for 3d point clouds. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2019)
44. Komarichev, A., Zhong, Z., Hua, J.: A-cnn: Annularly convolutional neural networks on point clouds. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2019)
45. Lan, S., Yu, R., Yu, G., Davis, L.S.: Modeling local geometric structure of 3d point clouds using geo-cnn. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2019)
46. Mao, J., Wang, X., Li, H.: Interpolated convolutional networks for 3d point cloud understanding. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 1578–1587
47. Prasad, M., Zisserman, A., Fitzgibbon, A., Kumar, M.P., Torr, P.H.: Learning class-specific edges for object detection and segmentation. In: *Computer Vision, Graphics and Image Processing*. Springer (2006) 94–105
48. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: 2011 International Conference on Computer Vision, IEEE (2011) 991–998
49. Bertasius, G., Shi, J., Torresani, L.: High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision. 2015 IEEE International Conference on Computer Vision (ICCV) (Dec 2015)
50. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE international conference on computer vision. (2015) 1395–1403

51. Hu, Y., Zou, Y., Feng, J.: Panoptic edge detection (2019)
52. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (2015) 234–241
53. Tchapmi, L., Choy, C., Armeni, I., Gwak, J., Savarese, S.: Segcloud: Semantic segmentation of 3d point clouds. In: *2017 international conference on 3D vision (3DV)*, IEEE (2017) 537–547
54. Tatarchenko, M., Park, J., Koltun, V., Zhou, Q.Y.: Tangent convolutions for dense prediction in 3d. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2018) 3887–3896
55. Ye, X., Li, J., Huang, H., Du, L., Zhang, X.: 3d recurrent neural networks with context fusion for point cloud semantic segmentation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. (2018) 403–417
56. Landrieu, L., Simonovsky, M.: Large-scale point cloud semantic segmentation with superpoint graphs. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2018) 4558–4567
57. Rethage, D., Wald, J., Sturm, J., Navab, N., Tombari, F.: Fully-convolutional point networks for large-scale point clouds. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. (2018) 596–611
58. Wang, S., Suo, S., Ma, W.C., Pokrovsky, A., Urtasun, R.: Deep parametric continuous convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2018) 2589–2597
59. Narita, G., Seno, T., Ishikawa, T., Kaji, Y.: Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. *arXiv preprint arXiv:1903.01177* (2019)
60. Huang, J., Zhang, H., Yi, L., Funkhouser, T., Nießner, M., Guibas, L.J.: Texturenet: Consistent local parametrizations for learning from high-resolution signals on meshes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2019) 4440–4449
61. Lei, H., Akhtar, N., Mian, A.: Spherical kernel for efficient graph convolution on 3d point clouds. *arXiv preprint arXiv:1909.09287* (2019)
62. Hermosilla, P., Ritschel, T., Vázquez, P.P., Vinacua, À., Ropinski, T.: Monte carlo convolution for learning on non-uniformly sampled point clouds. *ACM Transactions on Graphics (TOG)* **37**(6) (2018) 1–12
63. Su, H., Jampani, V., Sun, D., Maji, S., Kalogerakis, E., Yang, M.H., Kautz, J.: Splatnet: Sparse lattice networks for point cloud processing. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2018) 2530–2539
64. Zhang, C., Luo, W., Urtasun, R.: Efficient convolutions for real-time semantic segmentation of 3d point clouds. In: *2018 International Conference on 3D Vision (3DV)*, IEEE (2018) 399–408