Multimedia Tools and Applications (2024) 83:60949–60964
https://doi.org/10.1007/s11042-023-17814-4

Semantic segmentation of large-scale point clouds with neighborhood uncertainty

Yong Bao[1] · Haibiao Wen[2] · Baoqing Zhang[1]

Received: 9 September 2023 / Revised: 24 October 2023 / Accepted: 2 December 2023 /
Published online: 28 December 2023

Abstract

Large-scale point cloud segmentation is one of the important research directions in the field of computer vision, aiming at segmenting 3D point cloud data into parts with semantic meaning, which is widely used in the fields of robot perception, automated driving, and virtual reality. In practical applications, intelligences often face various uncertainties such as sensor noise, missing data, and uncertain model parameter estimation. However, many current research works do not consider the effects of these uncertainties, which can cause the model to overfit the noisy data and thus affect the model performance. In this paper, we propose a point cloud segmentation method with domain uncertainty that can greatly improve the robustness of the model to noise. Specifically, we first compute the neighborhood uncertainty, which is more reflective of the semantics of a local region than the prediction of a single point, which will reduce the impact of noise. Next, we fuse the uncertainty into the objective function, which allows the model to focus more on relatively deterministic data. Finally, we validate on the large-scale datasets S3DIS and Toronto3D, and the segmentation performance is substantially improved in both cases.

Keywords Point cloud · Segmentation · Uncertainty

1 Introduction

Point Cloud Segmentation (PCS) is one of the important tasks in the field of computer vision and machine learning, which aims to divide 3D point cloud data into semantic parts or

✉ Haibiao Wen
845488427@qq.com

Yong Bao
864549591@qq.com

Baoqing Zhang
359073191@qq.com

[1] Nanning University, No. 8, Longting Road, Yongning District, Nanning 541699, China

[2] Guangzhou College of Technology and Business, 166 Sanhua Road, Leping Town, Sanshui District, Foshan 510850, China

objects, and has been applied in the fields of autonomous driving [1], virtual reality [2, 3]. The important challenge for the point cloud segmentation task is that its raw data is usually irregular, unstructured and disordered. Although many methods excel in 2D computer vision, they are not able to process such data directly.

So there are some works proposed to deal with point clouds directly, such as the pioneering work PointNet [4] is a very effective method to deal with 3D point clouds directly. It learns single point features by sharing MLPs, but they do not consider contextual information. Subsequently, Pointnet++ [5] was proposed and solved the problem of PointNet. Meanwhile, more and more methods were proposed, which can be roughly categorized into: point-based methods [4–9], voxel-based methods [10–12], and Transformer-based methods [13, 14]. Point-based methods usually use the K-nearest-neighbor algorithm or spherical algorithm to aggregate the point cloud into a region, and then use the convolution operation on this local region. Voxel-based methods first divide the point cloud into fixed-size voxels, which are then processed using standard convolutional neural networks. Whereas Transformer-based methods typically have the best performance, this type of method first converts the point cloud into a Patch and then processes it using a designed transformer structure. Taken together, while these three types of methods have achieved increasingly better results, these methods are susceptible to uncertainties. It is undeniable that there has been a small amount of work on point cloud uncertainty [15]. But its not straightforward to use uncertainty to guide segmentation.

Therefore, we incorporate uncertainty factors into the point cloud segmentation process, which allows for a more comprehensive understanding of the point cloud data and provides more accurate and robust segmentation results, which is valuable for real-time applications such as autonomous driving. Therefore, we propose a neighborhood uncertainty point cloud segmentation method. The proposed method replaces the uncertainty of each point with its neighborhood uncertainty, and then uses a loss that dynamically adjusts the model according to the uncertainty of the point, which can make the model more focused on deterministic data. This can serve to filter noise, smooth anomalies, and improve model stability and model accuracy. To summarize, our contribution is as follows:

- To attenuate the effect of noise in data, we propose a neighborhood uncertainty method. The proposed method first aggregates domain knowledge and then uses this aggregated knowledge to calculate the uncertainty. The uncertainty reflects to some extent how confident the model is about a certain region;
- To make the model more focused on the deterministic region, we propose an uncertainty regularization method. By fusing the uncertainty degree into the loss function, the model can be made to focus more on the deterministic data and attenuate the effect of the uncertain data.
- To evaluate the effectiveness of the proposed method, we conducted experiments on the large-scale point cloud datasets S3DIS and Toronto3D are, both of which achieved leading results.

2 Related work

2.1 2D image segmentation

The research of deep learning on 2D images started earlier, and a large amount of work has been accumulated, we list a few recent representative works here [16–20]. SG-Net

[16] is a single-stage spatial granularity network characterized by feature sharing, high mask quality, high tracking robustness, and efficient inference time. In order to enhance the instance discrimination ability of query segmenters, [17] designed a new training framework to enhance query-based models through discriminative query embedding learning. It explores two fundamental properties of the relationship between queries and instances, i.e., dataset-level uniqueness and transformation isomorphism. The approach achieves significant performance improvements. In [18], on the other hand, a coarse-to-fine instance segmentation method is proposed, which uses optical flow techniques to propagate instance masks across video frames, conditioning the appearance flow on the input video frames,so that it takes into account the scene context changes, and the overall model is trained end-to-end to jointly optimize the loss function on multiple tasks of instance segmentation, optical flow estimation, and flow consistency. This coarse and fine prediction combined with the conditionalized decomposition of appearance streams makes the video instance segmentation results more stable and robust, surpassing the performance of previous work on benchmark datasets. Overall, these methods achieve increasingly better results and give good theoretical support for 3D point cloud segmentation tasks, but cannot be directly used to handle point cloud segmentation tasks.

2.2 3D point cloud segmentation

In recent years, with the development of artificial intelligence and the generalization of hardware acquisition devices, research on 3D point cloud segmentation tasks has begun to emerge, which can provide better geometric perception for intelligences, which is not available for 2D images. Current point cloud segmentation methods can be categorized into point-based methods [4–9, 21–23], voxel-based methods [10–12, 24], and transformer-based methods [13, 14].

Point-based methods perform point cloud segmentation by learning the feature representation and semantic information of point cloud data. PointNet [4] is one of the first point-based segmentation methods proposed to directly deal with disordered point cloud data. The main idea of PointNet is to deal with each point in the point cloud one by one, instead of treating the whole point cloud as a single entity. This is achieved by extracting features from each point using a shared multilayer perceptron (MLP) and then performing a maximum pooling operation to aggregate these features into a global representation of the point cloud. One obvious drawback of this method is the inability to learn local features. Subsequently, many improved network structures have been proposed, PointNet++ [5] is an extended and improved version of PointNet with more expressive power and better performance in dealing with point cloud data.The core idea of PointNet++ is to extract and aggregate features from point cloud data through a hierarchical structure. It introduces a module called “PointNet Set Abstraction”, which is used to extract features from local regions and gradually aggregate global features. Subsequently, DGCNN (Dynamic Graph CNN) [7] was proposed as a deep learning model for point cloud analysis, which is based on the idea of Graph Convolutional Neural Networks (GCNs) and Dynamic Graph Construction, and is able to effectively learn and analyze features from point cloud data. Unlike traditional grid-based convolution, DGCNN is able to handle irregular and inhomogeneous point cloud data for objects of various shapes and sizes. Then KPConv [9] (Kernel Point Convolution) was proposed as a deep learning model for point cloud analysis, which is based on the idea of Convolutional Neural Networks (CNNs) and introduces adaptive convolutional kernel points to process the point cloud data. The core idea of KPConv is to represent the point cloud data as a set of reference points called

"Kernel Points", which are used to define the point cloud data. The core idea of KPConv is to represent the point cloud data as a set of reference points called "Kernel Points", which are used to define the shape and weight of the convolution kernel. Each Kernel Point has an adaptive convolutional kernel that automatically adjusts to the position and characteristics of the points around it. Cui et al. [21] proposed a point cloud analysis method based on dynamic graph convolutional network, called GAG-CNN. it uses dynamic graph to represent the point cloud, and constructs a changing graph representation based on the input point cloud, which is more consistent with the point cloud structure. It applies the geometric attention mechanism to learn the importance of different parts of the point cloud, and uses graph convolution operation to capture the local features and global structure information of the point cloud, and builds a dynamic lexicon to encode the semantic information of the point cloud, which can be trained end-to-end and used for the tasks of point cloud classification and segmentation. Experimental results show that this method achieves high accuracy on ShapeNet and S3DIS datasets. In order to differentiate between different points, [22] proposed an efficient point cloud semantic segmentation method using spatially adaptive convolution. The method learns how to aggregate information from neighbors based on the location of the points, and employs different sensory fields for different points. It achieves adaptive convolution by encoding spatial coordinates into features and designing a location-aware weight generation module, which can better capture local geometric structures and is more efficient than ordinary convolution.Cloud-RAIN [23], on the other hand, solves the problem that the existing point cloud networks do not take into account the reflective symmetry of the 3D shapes, which makes it difficult to summarize the 3D shapes well, and has a poor generalization ability. They realize the isovariance of features to reflections by symmetrization and achieve good segmentation performance.

Voxel-based methods are PointNetVLAD [24], which is a combination of PointNet and VLAD (Vector of Locally Aggregated Descriptors). It converts the point cloud data into a voxel grid and locally aggregates the features within each voxel. The overall point cloud representation is then obtained by encoding the features of each voxel using VLAD coding. By clustering or classifying the encoded features, segmentation of the point cloud can be performed.

Transformer-based methods have been developing very rapidly in recent years, broadly speaking, Pointformer [13], Pointformer++ [14], and so on. First of all, we look at Pointformer, which is a Transformer-based point cloud segmentation method, which represents the point cloud data as a series of point feature vectors, and aggregates and updates the point features through the self-attention mechanism and the multilayer perceptron, and utilizes Transformer's parallel computing capability and global attention mechanism to achieve efficient and accurate point cloud segmentation. To address the shortcomings of Pointformer, Pointformer++ introduces a multi-layer hybrid attention mechanism to effectively fuse local and global information. It also introduces contextual relationship modeling of point cloud data and feature compression techniques to improve the accuracy and efficiency of point cloud segmentation. These methods have achieved better and better results, but they are susceptible to noise, which greatly affects model performance and is not suitable for scenarios with high security requirements.

2.3 Local knowledge aggregation and uncertainty

In computer vision, the merit of the local features of the network and the robustness of the model are crucial for the vision task. Therefore, there are some excellent works [15, 25–

27] have been proposed. In [25], a fully convolutional feature aggregation network structure, DenserNet, is proposed. This method extracts multi-scale features from low to high resolution through multiple parallel branches and the feature outputs from each branch are fused by convolution to finally produce a dense feature map. In [26], the global attention module is used to fuse the features from different regions to extract the key region features, and the local attention module is used to further decompose the features in the key region at a finer granularity, which achieves a very good performance. Although the above works have achieved good results, these works are susceptible to human-designed noise and have poor security and robustness. For example, [27] proposed a new unimodal adversarial attack method, where only a single unimodal input needs to be fine-tuned to influence the model's multimodal fusion decision. Therefore, it is crucial to improve the model robustness and attenuate the effect of noise on the model. The first uncertainty-based LiDAR panorama segmentation method was proposed in [15]. It can simultaneously predict the categories and instances of all scanned objects,and give the uncertainty measure of each prediction, which can effectively improve the robustness and interpretability of LiDAR panorama segmentation. This work is relative to ours in that they do not directly utilize the uncertainty to guide the segmentation, which may be less effective.

3 Method

3.1 Problem definition

Suppose we have a point cloud scene, denoted as$P \in R^{N\times D}$, where N denotes the number of points and D denotes the dimension of each point. The goal of point cloud segmentation is to learn a classifier F that predicts the category Y of each point, i.e., $Y = F(P)$, given a training set. The total number of categories is denoted by C.

3.2 General overview

There is no doubt that there are unavoidable noise and unpredictable regions in point cloud data, but most current point cloud segmentation methods do not consider these issues. Therefore their methods tend to be overconfident about the prediction results, which is not feasible in practical applications. Therefore, we propose a point cloud segmentation method for domain uncertainty, which enables the model to focus on more certain regions while weakening the effect of uncertain data. As shown in Fig. 1, the proposed method consists of input, network structure, uncertainty estimation, and output. For the input data, more to improve the model generalization ability and robustness, we use the data enhancement operation. As for the network structure, we use Backbone with two shared parameters, which can improve the model's ability to learn knowledge. Finally, to further enhance the robustness of the model, we use an uncertainty estimation method by which the uncertainty of each point can be obtained. Finally, by fusing the uncertainty to the cross-entropy loss, the final output can be obtained.

3.3 Network architecture

We use RandLA-Net as Backbone because it has fast processing speed for large-scale point clouds and achieves very good segmentation results. RandLa-Net includes Local Feature Aggregation (LFA), Random Downsampling (RS), Upsampling (US) and MLP

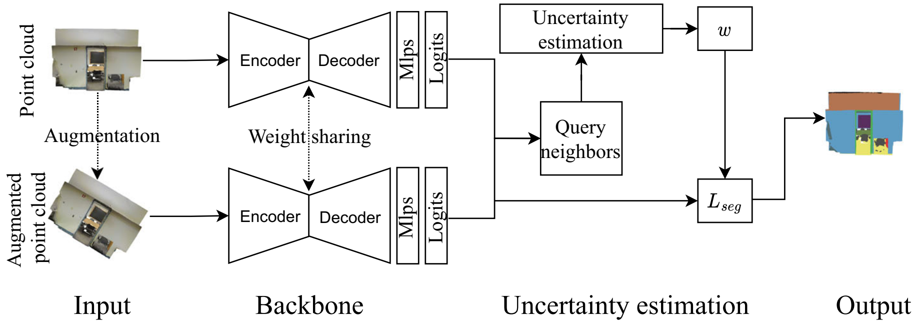

Fig. 1 Our method. First, the original and augmented data are entered into Backbone to get the predictive distribution. Then, the neighborhood predictive distribution is obtained based on this predictive distribution, and then the uncertainty is obtained. Finally, we can obtain the final prediction based on the uncertainty and cross-entropy loss. w denotes the uncertainty.L_{seg} is the loss function

operations. Specifically, when given a point cloud P with N input points, RandLa-Net performs feature extraction hierarchically, producing feature representations with different dimensions.RandLa-Net focuses on extracting high-level semantic features, capturing the global structure, semantic information, and contextual relationships in the point cloud. However, we found that it is not very robust to noise in the dataset. Therefore, we propose an improved method in this paper based on RandLA-Net. We improve the data input level of RandLA-Net in the first place by using 3 in data enhancement. Then for the model structure, we use the RandLA-Net structure with shared parameters, and finally for the model output and loss function, we use the uncertainty to further improve this RandLA-Net.

3.4 Data enhancement

For the input, more to improve the generalization ability and robustness of the model, we use three kinds of data augmentation, which are scene-level transformation, point perturbation, and random flip. For a point cloud P, we can split it into coordinates $P \in R^{N\times 3}$ and others (e.g., color, normal vector). Our scene-level transformations and random flips are both operations on point cloud coordinates. For scene-level transformation, we first define a random selection matrix $T^r \in R^{3\times 3}$, and then multiply P with T^r to obtain the transformed point cloud P^r, i.e., $P^r = P \bullet T^r$, where

$$T^r = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

and the transformation angle θ obeys the uniform distribution $U \in (2, 2\pi)$, and “$\bullet$” denotes matrix multiplication. For the flip, we only use the flip centered on the Y-axis, denoted as $P^m = P \bullet T^m$, where $T^m = diag(1, -1, 1)$. For the point perturbation operation, we define a perturbation matrix $T^j \in R^{N\times 3}$, which is a Gaussian noise distributed between $[-0.05, 0.05]$, then the perturbed point cloud $P^j = P + T^j$. We randomly select one of the enhancements from P^r, P^m, P^j at training time and send it to the network training along with the original point cloud P. We believe that by doing this, we can serve to expand the data, improve the model learning ability, and have to take advantage of the final uncertainty modeling.

3.5 Neighborhood uncertainty estimation

In order to further improve the robustness of the boosting model, so that the model focuses more on deterministic regions and less on regions of relative uncertainty, we propose a method for domain uncertainty estimation. The basic idea of our approach is that similar samples may have the same labels, and we believe that features from semantically similar points should lie in the same feature space. Thus, the uncertainty of a point is accomplished by aggregating the knowledge from its nearest neighboring samples. Then the uncertainty estimation of a single point is achieved by aggregating features from similar samples, and depending on the size of the uncertainty, we can then make the model focus on the certainty region with high confidence data, which will serve to filter noise and smooth out anomalous data. The method used to aggregate neighbor knowledge and generate uncertainty is shown in Fig. 2.

Suppose there is a current point, such as the red point labeled “?” in Fig. 2. The domain point and domain prediction result of this point can be obtained by KNN firstly, and then the domain knowledge can be aggregated by average weighting operation as shown in the following equation.

$$I(p) = KNN(p) \tag{2}$$

$$\hat{p}_t^{(c)} = \frac{1}{K} \sum_{i \in I} p_i^{(c)} \tag{3}$$

The prediction of a localized region consisting of a set of points is more plausible than the likelihood estimation of a single point, therefore, we use the domain knowledge aggregation above. Next, we calculate the entropy value after this aggregation, as shown in the following equation. If the entropy value is larger, it means that the uncertainty of the corresponding region is larger. Then, by the same token, if the entropy value is smaller, it means that the corresponding region has less uncertainty and the model is more confident, and it also means that the region has less noise.

$$H(\hat{p}_t) = E[I(\hat{p})] = -\sum_{c=1}^{C} \hat{p}_t \log_2 \hat{p}_t \tag{4}$$

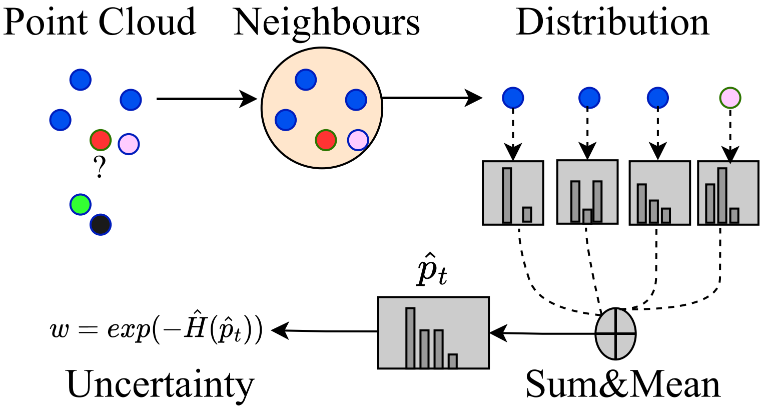

Fig. 2 Uncertainty estimation. “?” indicates that the neighborhood uncertainty is to be calculated for the red point.w denotes uncertainty

Next, we normalized this entropy value and applied the negative exponential function, and we got the final uncertainty. Here, we used the negative exponential function for the following reason, i.e., the function allows the model to put more weight on the low entropy data and relatively less weight on the high entropy data.

$$\hat{H}(\hat{p}_t) = \frac{H(\hat{p}_t)}{\log_2 C} \tag{5}$$

$$w = \exp(-\hat{H}(\hat{p}_t)) \tag{6}$$

3.6 Total loss

In order for the model to put more weight into the deterministic region, we characteristically blend the uncertainty into the traditional cross-entropy loss, i.e., as shown in the following equation:

$$L_{seg} = -\frac{1}{CL} \sum_{i=1}^{L} w_i \sum_{c=1}^{C} y_{ic}^{l} \log \frac{\exp(\hat{y}_{ic}^{l})}{\sum_{c=1}^{C} \exp(\hat{y}_{ic}^{l})} \tag{7}$$

With the fusion described above, the model can then target learning based on the size of uncertainty. Specifically, if some data has less uncertainty, the model puts more weight on it. If the uncertainty of some data is large, the model will pay less attention to this part of the data.

4 Experiments

In order to verify the validity of the proposed method, we conduct quantitative and qualitative analysis experiments in this section. We begin this section by describing the dataset used in the experiments, the specific experimental details, followed by the quantitative analysis, qualitative analysis, and finally the ablation experiments.

4.1 Dataset

The S3DIS [28] dataset is a comprehensive indoor scene dataset that includes six large-scale indoor scenes. Each scene contains 271 rooms with approximately 10^6 points and 13 semantic categories in each room. We use six attributes as inputs, including XYZ coordinates and RGB colors.

The Toronto3D [29] dataset is a point cloud of urban roads acquired using a vehicle-mounted laser system with 78.3 million points and 13 semantic categories.

4.2 Implementation details

In our experiments, we use RandLa-Net as the backbone network. The initial learning rate is set to 0.01 and decreases by 5% after each epoch. We use a neighborhood size of 15. The number of points input to the network is 40,960 for S3DIS and Toronto3D. We trained our method on an Nvidia 2080Ti GPU for 100 epochs. However, during the re-implementation of the comparison method, we had to reduce the batch size due to GPU limitations, which may lead to slightly different results compared to the original paper.

Table 1 Quantitative results on Area-5 of S3DIS

Methods	mIoU(%)	Ceil	Floor	Wall	Beam	Col	Win	Door	Table	Chair	Sofa	Book	Board	Clutter
PNet [4]	41.1	88.8	97.3	69.8	0.1	3.9	46.3	10.8	58.9	52.6	5.9	40.3	26.4	33.2
PCNN [30]	57.3	92.3	98.2	79.4	0.0	17.16	22.8	62.1	74.4	80.6	31.7	66.7	62.1	56.7
SPG [31]	58.0	89.4	96.9	78.1	0.0	42.8	48.9	61.6	84.7	75.4	69.8	52.6	2.1	52.2
SPH3D [32]	59.5	93.3	97.1	81.1	0.0	33.2	45.8	43.8	79.7	86.9	33.2	71.5	54.1	53.7
PWeb [33]	60.3	92	98.5	79.4	0.0	21.1	59.7	34.8	76.3	88.3	46.9	69.3	64.9	52.5
RandLA [6]	62.5	91.8	97.1	80.2	0.0	19.2	60.9	35.2	77.6	87.4	67.1	71.8	70.3	54.5
Ours	65.1	92.5	98	81.9	0.0	35.8	58.7	67.7	79.7	86.9	66.9	70.3	57.3	50.8

‘PNet’, ‘PCNN’, ‘SPG’, ‘PWeb’, ‘RandLA’ respectively denote PointNet, PointCNN, SPGraph, PointWeb, RandLA-Net

Table 2 Quantitative results on Toronto3D

Methods	oAcc	mIoU	Road	Road mrk.	Natural	Building	Util line	Pole	Car	Fence
PNet2 [5]	84.88	41.81	89.27	0.00	69.06	54.16	43.78	23.30	52.00	2.95
PNet2(MSG) [5]	92.56	59.47	92.90	0.00	86.13	82.15	60.96	62.81	76.41	14.43
DGCNN [7]	94.24	61.79	93.88	0.00	91.25	80.39	62.40	62.32	88.26	15.81
KPFCNN [9]	95.39	69.11	94.62	0.06	96.07	91.51	87.68	81.56	85.66	15.72
RandLA [6]	97.44	75.11	96.50	62.07	94.05	89.83	84.57	73.04	82.46	18.38
Ours	97.83	78.87	97.47	70.04	94.98	93.10	82.67	68.84	92.42	31.41

‘PNet2’ denotes Pointnet++. ‘RandLA’ denotes RandLA-Net

For S3DIS, we chose a batch size of 2 for the training set and 6 for the validation set. Regarding Toronto, we used a batch size of 1 for the training set and 4 for the validation set. We included position and color information in the model. We used the mean intersection on the concurrent set (mIoU,%) as an evaluation metric to assess the performance of the model.

4.3 Quantitative analysis

4.3.1 Results on Area-5 of S3DIS

In Table 1, we compare the results of the proposed method with several current popular methods on Area-5 of S3DIS. PointNet, as an early segmentation network, achieves 41.1% mIoU and relatively low results on each category. Subsequent methods PointCNN, SPGraph, SPH3D, and PointWeb achieved increasingly better results, reaching 57.3%, 58.0%, 59.5%, and 60.3% mIoU, respectively. The most recent method, RandLa-Net, achieved 62.5% mIoU on Area-5, greatly surpassing the above methods by, respectively 21.1%, 5.2%, 4.5%, 3%, and 2.2% of mIoU. our proposed method is exactly based on the work done by RandLa-Net because of its fast processing speed and segmentation performance for large-scale point clouds, but we found that it is susceptible to noise and has relatively poor robustness. Therefore, we do the improvement based on RandLa-Net to propose the method in this paper, and the results are shown in the last row of Table 1, and it can be seen that our method achieves 65.1% mIoU.Specifically, the proposed method in 'Wall', 'Col', and 'Door' categories achieve the best results.

4.3.2 Results on Toronto3D

In Table 2, we compare the results of our proposed method with several current classical point cloud semantic segmentation networks on Toronto3D. Our method achieves an accuracy of 97.83, all outperforming PointNet++, PointNet++(MSG), DGCNN, KPFCNN, and RandLA-Net networks by 12.95%, 5.27%, 3.59%, 2.44%, and 0.39%, respectively. And for the metric of average intersection, our method achieves 78.87% mIoU, which is 37.06%,

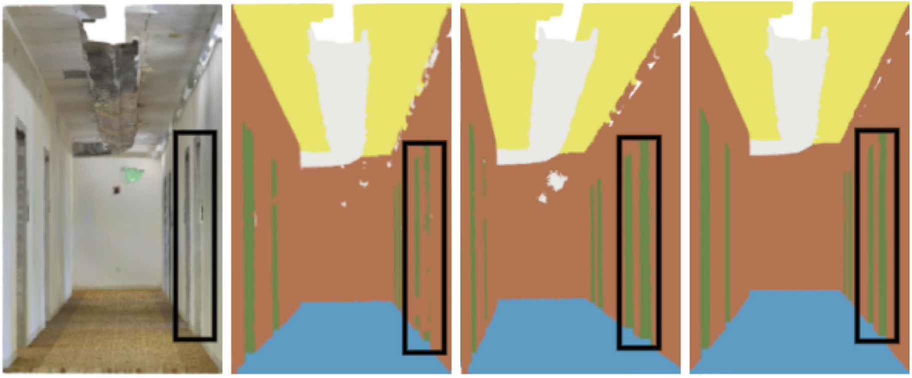
(a) Raw Data (b) RandLA-Net (c) Our Method (d) Ground Truth

Fig. 3 Qualitative results of different methods on the S3DIS. The black box region demonstrates the superiority of the proposed method

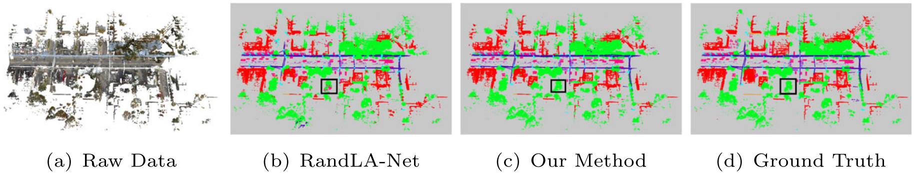

Fig. 4 Qualitative results of different methods on the Toronto3D. The black box region demonstrates the superiority of the proposed method

19.4%, 17.08%, 9.76%, and 3.76% mIoU better than the above methods, respectively. Specifically, the proposed method is in ‘Road’, ‘Road marking’, ‘Building’, ‘Car’, ‘Fence’ achieve excellent results of 97.47%, 70.04%, 93.10%, 92.42%, and 31.41% mIoU, respectively, which are better than the compared methods, proving the effectiveness of the proposed method.

4.4 Qualitative analysis

The role and significance of qualitative analysis in computer vision is to provide an in-depth understanding of visual features, patterns and semantic information by observing, understanding and interpreting image data. It plays a vital role in feature extraction and selection, information understanding and interpretation, data preprocessing and cleaning, result interpretation and evaluation, human-computer interaction and user experience. It provides support and guidance for the research and application of computer vision. Therefore, in order to prove the effectiveness of the proposed method more intuitively, we conducted qualitative analysis experiments. As shown in Figs. 3 and 4, we list the visualization results of the original point cloud data, RandLA-Net, Ground Truth, and the proposed method, respectively. It can be seen that the proposed method has more accurate segmentation results and is closer to Ground Truth, while the classical RandLA-Net has slightly worse segmentation results. Our analysis suggests that this is due to the fact that the classical RandLA-Net is not robust enough to noise and the model is slightly less generalized.

4.5 Abalation study

This section focuses on proving the effectiveness of the various components of the proposed method, as shown in Table 3. First, (#1) if only the base network is used, i.e., trained only with the officially provided RandLA-Net, a mIoU of 62.53% is achieved on Area-5 of S3DIS.(#2) When we incorporate data augmentation into the network, a mIoU of 64.16% is achieved, which is a relative improvement of 1.63% of the mIoU.(#3) And when we add the incorporate the proposed domain uncertainty, a final mIoU of 65.1% is achieved on Area-5, (#3) which is a 2.57% improvement in mIoU relative to the classical RandLa-Net (#1), and (#3) which is a 0.94% improvement in mIoU relative to #2. These results demonstrate the effectiveness of

Table 3 Ablations of different components on Area-5 of S3DIS

	Data augmentation	Uncertainty	mIoU(%)
#1			62.53
#2	✓		64.16
#3	✓	✓	65.10

Table 4 Generalizability analysis of uncertainty on different backbone

	Original	Uncertainty	PointNet++ [5]	KPCONV [9]	PCT [13]
#1	✓		52.8	63.8	69.5
#2	✓	✓	54.1	65.6	70.9

the proposed data augmentation and the domain uncertainty. We analyze that the significant improvement in the segmentation results is due to the fact that the proposed method can increase the amount of data and attenuate the effect of noise to a certain extent, which in turn can lead to a more robust knowledge of model learning.

4.6 Further analysis

4.6.1 Generalizability analysis of uncertainty on different backbone

In order to demonstrate the generalizability of the proposed uncertainty point cloud segmentation network on different Backbone, we conducted this set of experiments, we chose three excellent networks in the field of point cloud segmentation, Pointnet++ [5], KPCONV [9], and PointTransformer [13] as Backbone for the experiments respectively. Table 4 shows that the proposed method is significantly improved on different backbone networks, proving its effectiveness.

4.6.2 Experiments on more datasets

In order to demonstrate the generalizability of the proposed method on the dataset, we conducted the set of experiments. We performed experiments on Semantic3D [34], Scannet-v2 [35] in addition to the above experiments on S3DIS [28] and Toronto3D [29]. Table 5 shows that the proposed method significantly improved over Backbone (Randla-Net) on Semantic3D and Scannet-v2, proving the proposed method's good generalization.

4.6.3 Analysis of different neighborhood points

Intuitively, the larger the domain value, the richer the local information that can be learned, but also the more complex it is, the more difficult it is to learn, and the more computation is required. In order to further explore the relationship between the number of domain points and uncertainty, we conducted this set of experiments. The results in the Fig. 5 show that 1) when we increase the domain points from 4 to 15, the segmentation performance, improves from 62.3% to 65.1%. This indicates that a small domain value brings a smaller feeling field and does not have the generalization of the domain; 2) while when the domain points are increased

Table 5 Experiments on Semantic3D and Scannet-v2

	Original	Uncertainty	Semantic3D [34]	Scannet-v2 [35]
#1	✓		69.7	58.9
#2	✓	✓	73.6	62.3

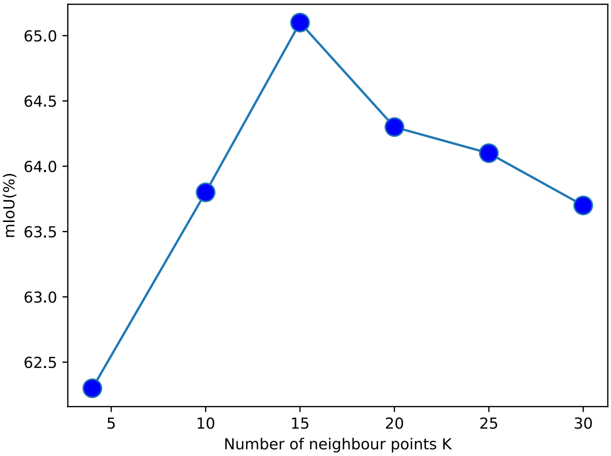

Fig. 5 The mIoU scores of our method for different choices of K in KNN

from 15 to 30, the segmentation performance instead decreases from 65.1% to 63.8%, which indicates that a larger domain value brings a more complex geometric information, and then by calculating the uncertainty the unique information of each point may be lost.

5 Conclusion

In this paper, a domain uncertainty point cloud segmentation method is proposed. First, local knowledge is aggregated through the domain, and then, the point-by-point uncertainty is generated and fused to the objective function based on the domain knowledge, which can act as a different degree of penalty and regularization for different points. Finally, we have experimented the proposed method on the large indoor dataset S3DIS give you large outdoor dataset Toronto3D to demonstrate the effectiveness of the proposed method. For future work, we will explore more efficient uncertainty methods.

Acknowledgements This research was funded by the Natural Science Project of China for Young and middle-aged(No.2022KY1786).

Author Contributions **Yong Bao:** Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing original draft. **Wen Haibiao:** Conceptualisation, Methodology. **Zhang Bao Qing:** Validation, Software.

Data Availability The **S3DIS** dataset used in the article can be obtained at http://buildingparser.stanford.edu/dataset.html. The **Toronto3D** dataset used in the article can be obtained at https://opendatalab.org.cn/Toronto-3D/download. The **Semantic3D** dataset used in the article can be obtained at http://www.semantic3d.net/view_dbase.php?chl=1. The **Scannet-v2** dataset used in the article can be obtained at https://github.com/ScanNet/ScanNet.

Declarations

Conflicts of interest The authors declare that they have no conflicts of interest.

Ethical standard Data used in the present study are publicly available, and ethical approval and informed consent were obtained in each original study.

Competing interests The authors have no relevant financial or nonfinancial interests to disclose.

References

1. Chen X, Ma H, Wan J, Li B, Xia T (2016) Multi-view 3d object detection network for autonomous driving. 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 6526–6534
2. Blanc T, Beheiry ME, Caporal C, Masson J-B, Hajj B (2020) Genuage: visualize and analyze multidimensional single-molecule point cloud data in virtual reality. Nature Methods 1–3
3. Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R (2020) Nerf: representing scenes as neural radiance fields for view synthesis. arXiv:2003.08934
4. Qi C, Su H, Mo K, Guibas LJ (2016) Pointnet: deep learning on point sets for 3d classification and segmentation. 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 77–85
5. Qi C, Yi L, Su H, Guibas LJ (2017) Pointnet++: deep hierarchical feature learning on point sets in a metric space. In: NIPS
6. Hu Q, Yang B, Xie L, Rosa S, Guo Y, Wang Z, Trigoni A, Markham A (2019) Randla-net: efficient semantic segmentation of large-scale point clouds. 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 11105–11114
7. Wang Y, Sun Y, Liu Z, Sarma SE, Bronstein MM, Solomon JM (2018) Dynamic graph cnn for learning on point clouds. ACM Trans Graph (TOG) 38:1–12
8. Lin Y, Yan Z, Huang H, Du D, Liu L, Cui S, Han X (2020) Fpconv: learning local flattening for point convolution. 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 4292–4301
9. Thomas H, Qi C, Deschaud J-E, Marcotegui B, Goulette F, Guibas LJ (2019) Kpconv: flexible and deformable convolution for point clouds. 2019 IEEE/CVF international conference on computer vision (ICCV), pp 6410–6419
10. Choy CB, Gwak J, Savarese S (2019) 4d spatio-temporal convnets: Minkowski convolutional neural networks. 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3070–3079
11. Graham B, Engelcke M, Maaten L (2017) 3d semantic segmentation with submanifold sparse convolutional networks. 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 9224–9232
12. Huang S-S, Ma Z-Y, Mu T-J, Fu H, Hu S (2021) Supervoxel convolution for online 3d semantic segmentation. ACM Trans Graph (TOG) 40:1–15
13. Zhao H, Jiang L, Jia J, Torr PHS, Koltun V (2020) Point transformer. 2021 IEEE/CVF international conference on computer vision (ICCV), pp 16239–16248
14. Wu X, Lao Y, Jiang L, Liu X, Zhao H (2022) Point transformer v2: grouped vector attention and partition-based pooling. arXiv:2210.05666
15. Sirohi K, Marvi S, Büscher D, Burgard W (2022) Uncertainty-aware lidar panoptic segmentation. 2023 IEEE international conference on robotics and automation (ICRA), pp 8277–8283
16. Liu D, Cui Y, Tan W, Chen Y (2021) Sg-net: spatial granularity network for one-stage video instance segmentation. 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 9811–9820
17. Wang W, Liang J, Liu D (2022) Learning equivariant segmentation with instance-unique querying. arXiv:2210.00911
18. Qin Z, Lu X, Nie X, Liu D, Yin Y, Wang W (2023) Coarse-to-fine video instance segmentation with factorized conditional appearance flows. IEEE/CAA J Automatica Sinica 10:1192–1208
19. Cui Y, Yan L, Cao Z, Liu D (2021) Tf-blender: temporal feature blender for video object detection. 2021 IEEE/CVF international conference on computer vision (ICCV), pp 8118–8127
20. Liang J, Zhou T, Liu D, Wang, W (2023) Clustseg: clustering for universal segmentation. arXiv:2305.02187
21. Cui Y, Liu X, Liu H, Zhang J, Zare A, Fan B (2021) Geometric attentional dynamic graph convolutional neural networks for point cloud analysis. Neurocomputing 432:300–310
22. Xu C, Wu B, Wang Z, Zhan W, Vajda P, Keutzer K, Tomizuka M (2020) Squeezesegv3: spatially-adaptive convolution for efficient point-cloud segmentation. In: European conference on computer vision. https://api.semanticscholar.org/CorpusID:214802232
23. Cui Y, Ruan L, Dong H, Li Q, Wu Z, Zeng T, Fan F (2023) Cloud-rain: point cloud analysis with reflectional invariance. arXiv:2305.07814

24. Uy MA, Lee GH (2018) Pointnetvlad: deep point cloud based retrieval for large-scale place recognition. 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 4470–4479
25. Liu D, Cui Y, Yan L, Mousas C, Yang B, Chen Y (2020) Densernet: weakly supervised visual localization using multi-scale feature aggregation. In: AAAI conference on artificial intelligence. https://api.semanticscholar.org/CorpusID:227305257
26. Yan L, Cui Y, Chen Y, Liu D (2021) Hierarchical attention fusion for geo-localization. ICASSP 2021 - 2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 2220–2224
27. Cheng Z, Choi H, Liang J, Feng S, Tao G, Liu D, Zuzak M, Zhang X (2023) Fusion is not enough: single-modal attacks to compromise fusion models in autonomous driving. arXiv:2304.14614
28. Armeni I, Sax S, Zamir AR, Savarese S (2017) Joint 2d-3d-semantic data for indoor scene understanding
29. Tan W, Qin N, Ma L, Li Y, Du J, Cai G, Yang K, Li J (2020) Toronto-3d: a large-scale mobile lidar dataset for semantic segmentation of urban roadways. In: 2020 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 797–806. https://doi.org/10.1109/CVPRW50498.2020.00109
30. Li Y, Bu R, Sun M, Chen B (2018) Pointcnn. arXiv:1801.07791
31. Landrieu L, Simonovsky M (2017) Large-scale point cloud semantic segmentation with superpoint graphs. 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 4558–4567
32. Lei H, Akhtar N, Mian AS (2019) Spherical kernel for efficient graph convolution on 3d point clouds. IEEE Trans Patt Anal Mach Intell 43:3664–3680
33. Zhao H, Jiang L, Fu C-W, Jia J (2019) Pointweb: enhancing local neighborhood features for point cloud processing. 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 5560–5568
34. Hackel T, Savinov N, Ladicky L, Wegner JD, Schindler K, Pollefeys M (2017) Semantic3d.net: a new large-scale point cloud classification benchmark. ISPRS annals of photogrammetry, remote sensing and spatial information sciences
35. Dai A, Chang AX, Savva M, Halber M, Funkhouser T, Nießner M (2017) Scannet: richly-annotated 3d reconstructions of indoor scenes. In: Proc. computer vision and pattern recognition (CVPR). IEEE

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.