Waterfall-Net: Waterfall Feature Aggregation for Point Cloud Semantic Segmentation

Hui Shuai(✉), Xiang Xu, and Qingshan Liu

Engineering Research Center of Digital Forensics, Ministry of Education, School of Computer and Software, Nanjing University of Information Science, Nanjing 210044, China
huishuai13@nuist.eud.cn, qsliu@nuist.edu.com

Abstract. In this paper, we observe that the point cloud density affects the performance of different categories in 3D point cloud semantic segmentation. Most existing point-based methods implicitly deal with this density issue via extracting multi-scale features in a single forward path. Instead, we propose a Waterfall-Net that explicitly utilizes the density property via cross-connected cascaded sub-networks. In Waterfall-Net, three sub-networks successively process the input point cloud. Each sub-network handles the point features sampled at different densities, obtaining the information at various densities. The output features of one sub-network are up-sampled via a learnable up-sample method and fed into the next sub-network. This Sub-Network Fusing aligns the density of two sub-networks and maintains the contextual information. Meanwhile, Sub-Stage Fusing fuses the sub-stage features between successive sub-networks according to the density. Such waterfall-like feature aggregation ensembles all the features from different densities and enhances the model learning ability. We empirically demonstrate the effectiveness of the Waterfall-Net on two benchmarks. Specifically, it achieves 72.2% mIoU on S3DIS and 55.7% mIoU on SemanticKitti.

Keywords: Point cloud semantic segmentation · Density property · Feature aggregation

1 Introduction

Point cloud semantic segmentation is a fundamental task in 3D scene analysis. It plays a vital role in many applications, such as autonomous driving and robotics. Recently, many methods have obtained promising performance on several benchmarks [2,3]. In this paper, we focus on the point-based methods [7,13,19,31] directly processing the 3D points as no information conversion occurs.

For point-based methods, we observe that the density of the input points can significantly affect the performance of different categories. As shown in Fig. 1, we feed the point cloud of S3DIS randomly sampled at different densities into the

S. Yu et al. (Eds.): PRCV 2022, LNCS 13536, pp. 28–40, 2022.
https://doi.org/10.1007/978-3-031-18913-5_3

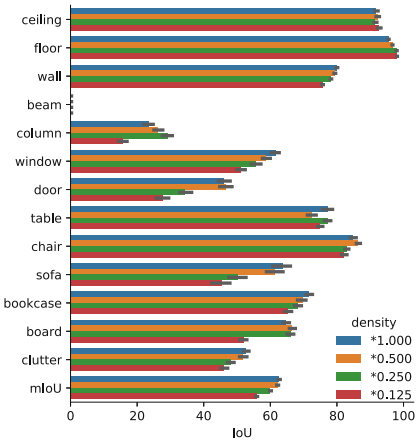

Fig. 1. The performance of RandLA-Net on S3DIS at different input densities. Different categories achieve the best performance at different densities.

RandLA-Net and evaluate the IoU of all the categories. The results demonstrate that a particular category is better resolved at a certain density. Some categories are better predicted at high resolution while some other categories achieve the best performance at lower densities. The reasons for this phenomenon are twofold. First, points sampled at different densities can reflect the geometric property of the objects. For example, when points are sampled at low density, the performance of ceiling and floor increases because low density can smooth the noise on these large planes. On the contrary, the performance of the window and clutter decreases as these objects need more detailed geometric information for accurate prediction. Second, the density of the sampling procedure affects the receptive field of the sampled points. At different densities, the K nearest neighbor points of a point will provide contextual information of different scopes. So, how can we take advantage of all the superiority of different densities?

The naive idea to utilize multiple densities is to combine the results predicted at a range of densities, but we do not know the best density for each category. An alternative is to aggregate the feature arising from various densities. Some previous works utilize points sampled via various rules or features with different receptive fields intuitively. PointNet++ [17] employs the density adaptive layer to aggregate multi-scale features from the neighborhood of different scopes, dealing with the non-uniform sampling density in the point cloud. JSNet [33] fuses the features from multi-layers with concatenating and adding operations at the end of the backbone. RandLA-Net [7] and KPConv [19] et al. both use a U-shape encoder-decoder structure, in which features with different receptive fields are fused via skip connections between the encoder and decoder. These methods extract multi-scale features in a single "funnel" forward path, and the density of the point decreases as the depth of the layer increases. The features from early layers are detailed features at high density and the features

from layers in the back are abstract features at low density, but some features of various semantics-and-density combinations (such as abstract features with high density) are missing. Obviously, we can obtain more abundant features in different states via further exploring the density property.

This paper establishes a network that can sufficiently utilize the density property with cascaded sub-networks. As visualized in Fig. 2, the proposed method employs a U-shape encoder-decoder architecture. The encoder consists of three sub-networks. Each sub-network handles the point features sampled at different densities. In this way, the encoder can extract clues at different densities. The output of one sub-network acts as the input of the next one via a learnable up-sampling, to fully utilize the contextual information and provide features of multiple granularity. From another perspective, the cascaded sub-networks is a polishing process for the features. This mechanism is termed as sub-networks fusing. Meanwhile, each sub-network consists of several sub-stages. The corresponding sub-stage features from adjacent sub-networks are fused according to the same density, and we term this mechanism as sub-stage fusing. Overall, the features from different layers are cross-connected like a waterfall. Such waterfall feature aggregation assembles features at different states and enhances the model learning ability. Thus, we name the proposed method as Waterfall-Net. We evaluate the Waterfall-Net on two standard benchmarks, S3DIS and SemanticKitti, and the results demonstrate that the proposed method can significantly improve the baseline's performance. Our main contributions are summarized as follows:

1. We observe that the density of input point cloud can significantly affect the performance of different categories.
2. We propose a Waterfall-Net to take advantage of the density property. It extracts abundant features at different densities and aggregate them in an waterfall-like manner for better prediction.
3. We propose a learnable up-sampling method for point cloud feature interpolation. It can adaptively incorporate the contextual clues for interpolation and outperforms the rule-based methods.

2 Related Work

Point-Based Point Cloud Semantic Segmentation: Point-based methods process the raw point cloud directly. PointNet [16] is the first method to employ point-wise MLP and symmetry function for point cloud analysis. Based on PointNet, PointNet++ [17] and PointSIFT [8] use shared point-wise MLPs for point-wise manipulation and adopt aggregate modules to capture the context information. Subsequently, Francis et al. [6] further utilizes K-means and KNN in both the world space and the latent feature space to regularize feature learning. To overcome the drawback that MLPs only process points individually, PointWeb [30] and RSCNN [13] design some measurements to explore the relationship between the point pairs in a local region. Besides, PCCN [23] and KP-FCNN [19] explore effective convolution operations for point clouds. Along with the rise

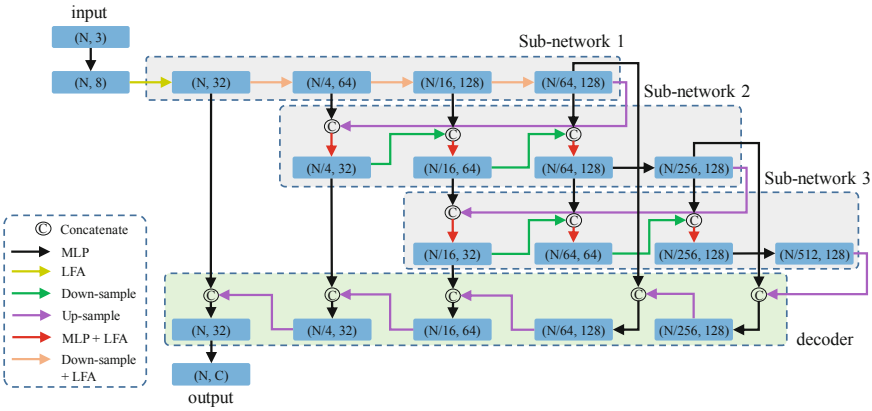

Fig. 2. The overview architecture of Waterfall-Net. The Waterfall-Net employs an encoder-decoder architecture. The encoder consists of three sub-networks, each sub-network handles point features sampled at a certain density. These sub-networks are cross-connected via Sub-Network Fusing and Sub-Stage Fusing mechanisms.

of Graph Neural Networks, PyramNet [34] and GAC [22] use graph-based networks to model point clouds and capture the underlying shapes and geometric structures. DGCNN [24] proposes an Edge-Conv based on a graph to recover the topological information of the point cloud. Beyond that, RandLA-Net [7] employs a local feature aggregation module on randomly sampled points to progressively increase the receptive field. These methods process the point cloud in a single forward path, and the density of the points decreases as the depth increase. However, the density property we observed is not explicitly considered by previous methods. In this paper, we take full use of the density property through several cascaded sub-networks.

Multi-scale Feature Fusing: In image analysis methods, the multi-scale feature is very important to deal with the scale space. The straightforward method is using image pyramid [1] to feed multi-resolution images into multiple networks and aggregate the output [20]. To improve efficiency, PSPNet [32] and Deeplab series [4] aggregate features from different scales. Hourglass [15] and its extension [9] combine the low-level and high-level features with short-cut connections. HRNet [21] and DFANet [11] construct several subnetworks for different resolution and conduct multi-scale fusion repeatedly. For the 3D point cloud, the multi-scale feature fusing methods are relatively few. PointNet++ [17] employs the density adaptive layer to aggregate multi-scale features from the neighborhood of different scopes like PSPNet. JSNet [33] fuses the features from multi-layers with concatenating and adding operations at the end of the backbone. RandLA-Net [7], KPConv [19], and GAC [22] et al. fuse features with different receptive fields via skip connections between the encoder and decoder. These methods extract multi-scale features in a single forward path, but the density property of the point cloud is not fully utilized. Inspired by the idea of HRNet

and DFANet, we establish a network that processes point features at multiple densities in cascaded sub-networks. In this way, the information at various densities is extracted and the network architecture is not bloated.

3 Waterfall-Net

Waterfall-Net employs an encoder-decoder architecture, following a typical semantic segmentation fashion. To sufficiently utilize the density property, we design a Cascaded Sub-networks Encoder to extract informative features at different densities. These features are fused via Sub-Network Fusing and Sub-Stage Fusing. Additionally, to align the density of different sub-networks, we propose a learnable upsample method that increases the density of the point cloud.

3.1 Cascaded Sub-Networks Encoder

The Cascaded Sub-networks Encoder transforms the input point cloud $\mathbf{P}$ into latent features via stacked sub-networks. Each sub-network consists of 4 sub-stages, it decreases the number of points via Random Sample (RS) and expands the dimension of per-point feature via the Local Feature Aggregate (LFA) module inherited from RandLA-Net [7]. We represent the feature of j-th stage in i-th sub-network as $\mathbf{F}_i^j$, $i \in \{1, 2, 3\}$ and $j \in \{1, 2, 3, 4\}$. The multiple sub-networks are cross-connected via Sub-Network Fusing and Sub-Stage Fusing mechanisms.

Sub-Network Fusing joins successive sub-networks via transmitting the features in a cascaded manner. The output of the i-th sub-network acts as the input of the $(i+1)$-th sub-network. Each sub-network's input is formulated as:

$$\mathbf{F}_i^1 = \begin{cases} \mathrm{LFA}(\mathcal{M}(\mathbf{P})) & \text{if } i = 1 \\ \mathrm{LFA}(\mathcal{M}(\mathbf{F}_{i-1}^2 \oplus \mathrm{UP}(\mathbf{F}_{i-1}^4))) & \text{otherwise} \end{cases} \tag{1}$$

where $\mathcal{M}$ represents the MLP, $\oplus$ is the concatenation operation, and UP is the upsample operation will be introduced in the next subsection. Note that $\mathbf{F}_{i-1}^4$ has a sparser density than $\mathbf{F}_{i-1}^2$, so the upsample operation is necessary to align their density. The preceding sub-network extracts features with high semantic awareness but at low density. With the interpolation of upsample operation, the Sub-Network Fusing inherits the semantic awareness from the previous sub-network and retains structure details at high density. Thus, this mechanism provides semantics-and-density combined features of more variety.

Sub-Stage Fusing establishes connections between sub-stages in adjacent sub-networks. The feature F_{i-1}^j and F_i^{j-1} contribute to the feature F_i^j together. The intermediate feature of each sub-stage in different sub-network is:

$$\mathbf{F}_i^j = \begin{cases} \mathrm{LFA}(\mathrm{RS}(\mathbf{F}_i^{j-1})) & \text{if } i = 1 \\ \mathrm{LFA}(\mathcal{M}(\mathrm{RS}(\mathbf{F}_i^{j-1}) \oplus \mathbf{F}_{i-1}^{j+1})) & \text{otherwise} \end{cases} \tag{2}$$

where $j > 1$. We first decrease the density of $\mathbf{F}_i^{j-1}$. Then, $\mathbf{F}_{i-1}^{j+1}$ and $\mathbf{F}_i^{j-1}$ are concatenated and fed to LAF. This mechanism constructs more informative features via aggregating features with different information granularity.

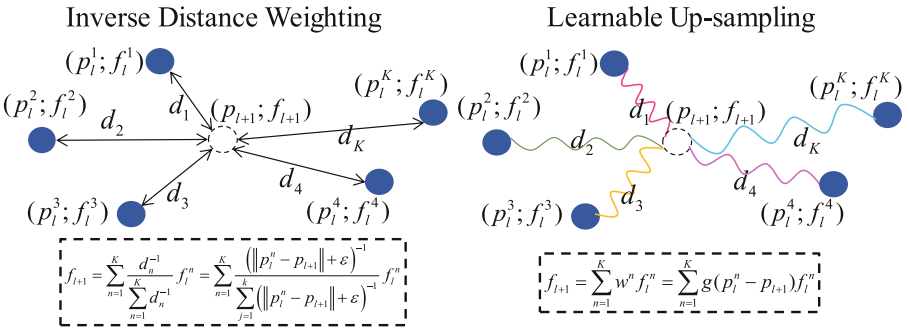

Fig. 3. IDW interpolates the features of the target point via weighting the surrounding points with inverse distance as weighing coefficients. The learnable up-sampling employs a network to estimate the weighting coefficients.

With these fusing methods, named waterfall feature aggregation, the multiple sub-networks work in a complementary manner. The feature of various semantic-and-density combinations is abstracted and fused. Through the waterfall-like cross connections in sub-networks, the information flow can be transferred through an arbitrary network pipeline.

3.2 Learn to Upsample

In the Cascaded Sub-network Encoder, the point cloud is randomly down-sampled. However, point cloud semantic segmentation aims to assign a semantic label for each point, which needs to propagate features from down-sampled points to the original points. RandLA-Net [7] employs the nearest interpolation that simply duplicates the features, making the interpolated features undistinguishable. More works, such as PointNet++ [17], uses Inverse Distance Weighted (IDW) average on k nearest neighbours. However, IDW only considers the distance between points while missing the needed semantic clues of the point cloud.

We propose a learnable up-sampling that can better capture the semantic clues of point cloud with a data-driven network. As shown in Fig. 3, given the point cloud $\mathbf{P}_l$ and its feature $\mathbf{F}_l$, the learnable up-sampling propagates $\mathbf{F}_l$ to $\mathbf{F}_{i+1}$, i.e. the feature of $\mathbf{P}_{l+1}$ ($\mathbf{P}_i \subset \mathbf{P}_{l+1}$). For one point p_{l+1} in $\mathbf{P}_{l+1}$, the nearest K points of it, i.e. $\{p_l^n\}_{n=1}^K$, are firstly selected from P_l. Then, a shared MLP is designed to encode the relative position between p_{l+1} and $\{p_l^n\}_{n=1}^K$ into weights, followed by a softmax function for normalization:

$$w^n = g(p_l^n - p_{l+1}; \mathbf{W}) \tag{3}$$

where p_{l+1} and p_l^n are the three-dimensional coordinates of points, $\mathbf{W}$ is the parameters in MLP, and $g()$ represents the MLP followed by softmax function. $w^n \in \mathbb{R}^1$ represents the weight between p_{l+1} and p_l^n. Compared with the distance-based weight in [7,17], the learned weight derives from the relative position and holistically optimized parameters. Thus, it can better reflect the distribution of the point cloud, making the semantic information more distinguishable. Finally, the features of nearest points are weighted summed as follows:

$$f_{l+1} = \sum_{n=1}^{K} (w^n \cdot f_l^n) \tag{4}$$

where f_l^n represents the features of p_l^n, and f_{l+1} is the feature p_{l+1}.

4 Experiments

In this section, we first conduct ablation studies to verify the effectiveness of our design on area 5 in S3DIS. Then, the results of Waterfall-Net in three popular datasets are reported, including S3DIS and SemanticKITTI.

4.1 Analysis of Waterfall-Net Architecture

We verify the effect of each design, including the Sub-network Fusing, Sub-stage Fusing, and Learnable upsampling. The quantitative results are reported on area5 in S3DIS.

Waterfall Feature Aggregation: In this part, we set RandLA-Net as the baseline. Then, a multi-scale block that extracts features from three parallel branches with various densities is embedded between the encoder and decoder. For Sub-Network Fusion verification, we remove the horizontal connection in the Waterfall-Net (marked as purple arrows between the sub-networks). For Sub-Stage Fusion verification, we remove the vertical connection between the stages (marked as black arrows between the stages), and the encoder acts as a forward network that consists of three cascaded subnetworks. The results of all the above-mentioned modules and Waterfall-Net are shown in Table 1. The RandLA-Net obtains mIoU of 62.8% in area5. However, embedding a multi-scale

Table 1. Performance of different modules on S3DIS area5

RandLA-Net	Multi-scale block	Sub-network fusing	Sub-stage fusing	mIoU(%)
✓				62.8
✓	✓			62.3
✓		✓		64.3
✓			✓	64.4
✓		✓	✓	66.1

block into RandLA-Net decreases the mIoU by 0.5%. The features at the end of the encoder are all derived from the same forward path. Thus, the multi-scale block is hard to introduce more abundant features but causes extra parameters, leading to over-fitting. When we only employ Sub-Network Fusing, our method obtains the mIoU of 64.3%, improving the baseline by 1.5%. When we only employ Sub-Stage Fusing, the performance of the baseline is improved by 1.6%, to 64.4%. It means that both the Sub-Network Fusing and Sub-Stage Fusing are beneficial for the discriminative feature extracting. Finally, combining the Sub-Network Fusing and Sub-stage Fusing into the waterfall feature aggregation, Waterfall-Net obtains the mIoU of 66.1%, by 3.3%. The improvement is larger than the sum of the gain arising from Sub-Network Fusing and Sub-Stage Fusing individually. That is to say, the Sub-Network Fusing and the Sub-Stage Fusing are complementary to each other.

Up-Sampling Method: The up-sampling method bridges the density gap between successive sub-networks in Waterfall-Net. It should increase the density of the point cloud and keep the contextual information simultaneously. For comparison, we implement the nearest-neighbor interpolation, inverse distance interpolation, and the learnable up-sampling as the up-sampling method. Their results are shown in Table 2. The nearest neighbor interpolation replicates the feature of the nearest point but the contextual information of other surroundings is not considered. Eventually, it achieves a performance of 64.3%. The inverse distance interpolation exploits all the surrounding points and uses the inverse of distance as the weighting coefficients. It outperforms the nearest neighbor interpolation by 0.6% and achieves the mIoU of 64.9%. However, its weighting coefficients are based on geometrical prior but what we need is the semantic information. Our proposed learnable up-sampling also uses all the surrounding points, and the weighting coefficients are inferred by a neural network. It explores the semantic relation in a data-driven manner and achieves the mIoU of 66.1%, outperforming the inverse distance interpolate by 1.2%.

Table 2. Performance of different up-sample methods on S3DIS area5

Method	mIoU(%)
Nearest neighbor interpolate	64.3
Inverse distance interpolate	64.9
Learnable up-sampling	**66.1**

Other Basic Block: Waterfall-Net is implemented based on the Local Feature Aggregation module proposed in RandLA-Net. To verify the general applicability of waterfall feature aggregation, we also implemented the Waterfall-Net based on the hierarchical point set feature learning layer proposed in PointNet++. The result of the PointNet++ based Waterfall-Net in S3DIS area5 is presented in Table 3. The waterfall-Net improves the performance of PointNet++ to 54.1%, by 3.2%. It demonstrates that the waterfall feature aggregation can generally

improve the performance of point-based methods for point cloud semantic segmentation.

Table 3. Quantitative results of different basic block on S3DIS (area 5)

Method	mIoU (%)	Ceil.	Floor	Wall	Beam	Col.	Wind.	Door	Table	Chair	Sofa	Book.	Board	Clut.
PointNet [16]	41.1	88.8	97.3	69.8	0.1	3.9	46.3	10.8	52.6	58.9	5.9	40.3	26.4	33.2
PointNet++ [17]	50.9	90.7	98.1	75.5	0.0	2.7	35.8	31.9	70.8	73.9	25.7	54.1	42.5	49.8
PointNet++ & Waterfall-Net	54.1	90.9	98.3	79.8	0.0	10.3	38.4	30.8	74.5	77.2	47.3	59.3	40.2	55.7
RandLA-Net [7]	62.8	91.5	96.0	80.6	0.0	26.1	62.5	47.6	76.4	84.1	60.7	71.3	65.5	54.1
RandLA-Net & Waterfall-Net	66.1	92.9	97.8	83.3	0.0	30.8	61.5	54.5	77.6	89.4	79.9	72.1	63.9	55.7

Table 4. Quantitative results of different approaches on S3DIS (6-fold cross validation)

Method	OA (%)	mACC (%)	mIoU (%)	Ceil.	Floor	Wall	Beam	Col.	Wind.	Door	Table	Chair	Sofa	Book.	Board.	Clut.
PointNet [16]	78.6	66.2	47.6	88.0	88.7	69.3	42.4	23.1	47.5	51.6	54.1	42.0	9.6	38.2	29.4	35.2
SPG [10]	85.5	73.0	62.1	89.9	95.1	76.4	62.8	47.1	55.3	68.4	73.5	69.2	63.2	45.9	8.7	52.9
PointCNN [12]	88.1	75.6	65.4	94.8	97.3	75.8	63.3	51.7	58.4	57.2	71.6	69.1	39.1	61.2	52.2	58.6
PointWeb [30]	87.3	76.2	66.7	93.5	94.2	80.8	52.4	41.3	64.9	68.1	71.4	67.1	50.3	62.7	62.2	58.5
ShellNet [29]	87.1	-	66.8	90.2	93.6	79.9	60.4	44.1	64.9	52.9	71.6	**84.7**	53.8	64.6	48.6	59.4
PointASNL [27]	**88.8**	79.0	68.7	**95.3**	**97.9**	81.9	47.0	48.0	**67.3**	70.5	71.3	77.8	50.7	60.4	63.0	**62.8**
KPConv [19]	-	79.1	70.6	93.6	92.4	**83.1**	63.9	**54.3**	66.1	**76.6**	57.8	64.0	**69.3**	**74.9**	61.3	60.3
RandLA-Net [7]	88.0	82.0	70.0	93.1	96.1	80.6	62.4	48.0	64.4	69.4	69.4	76.4	60.0	64.2	65.9	60.1
Waterfall-Net	88.5	**82.4**	**72.2**	94.7	97.8	82.8	**64.2**	53.9	64.8	70.5	**74.2**	78.3	66.0	65.3	**66.7**	60.0

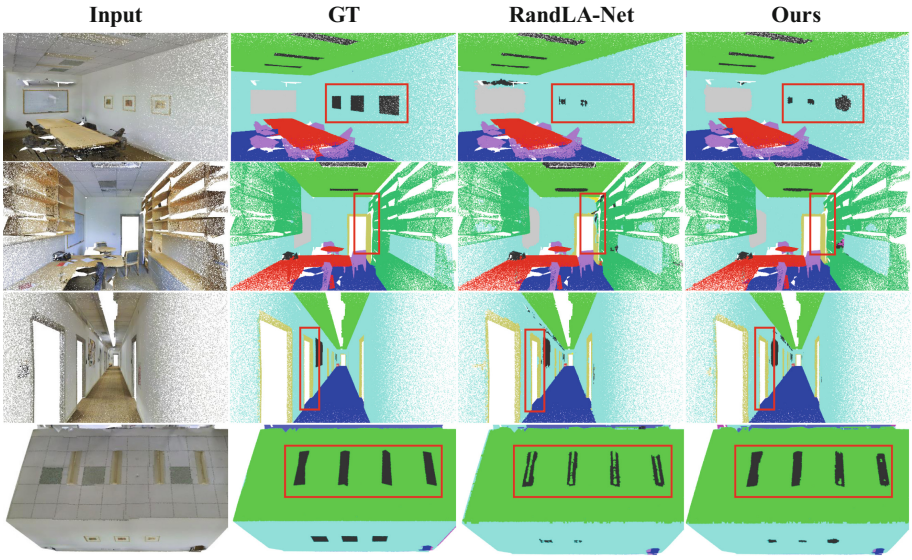

Fig. 4. Qualitative results of Waterfall-Net on S3DIS.

4.2 Results and Visualization

To verify the effectiveness of Waterfall-Net, we conduct it in two benchmarks: S3DIS and SemanticKitti, both indoor and outdoor scenarios.

Table 5. Quantitative results of different approaches on SemanticKITTI

Method	Size	mIoU(%)	road	sidewalk	parking	other-ground	building	car	truck	bicycle	motorcycle	other-vehicle	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traffic-sign
PointNet [16]	50K pts	14.6	61.6	35.7	15.8	1.4	41.4	46.3	0.1	1.3	0.3	0.8	31.0	4.6	17.6	0.2	0.2	0.0	12.9	2.4	3.7
SPG [10]		17.4	45.0	28.5	0.6	0.6	64.3	49.3	0.1	0.2	0.2	0.8	48.9	27.2	24.6	0.3	2.7	0.1	20.8	15.9	0.8
PointNet++ [17]		20.1	72.0	41.8	18.7	5.6	62.3	53.7	0.9	1.9	0.2	0.2	46.5	13.8	30.0	0.9	1.0	0.0	16.9	6.0	8.9
TangentConv [18]		40.9	83.9	63.9	33.4	15.4	83.4	90.8	15.2	2.7	16.5	12.1	79.5	49.3	58.1	23.0	28.4	8.1	49.0	35.8	28.5
PointASNL [27]		46.8	87.4	74.3	24.3	1.8	83.1	87.9	39.0	0.0	25.1	29.2	84.1	52.2	70.6	34.2	57.6	0.0	43.9	**57.8**	36.9
PointNL [5]		52.2	90.5	72.5	48.3	19.0	81.6	92.1	9.8	**42.6**	37.4	20.0	78.5	54.5	62.7	49.2	**57.8**	**28.3**	50.2	41.7	55.8
RandLA-Net [7]		53.9	90.7	73.7	60.3	20.4	86.9	**94.2**	**40.1**	26.0	25.8	**38.9**	81.4	61.3	66.8	49.2	48.2	7.2	56.3	49.2	47.7
SqueezeSeg [25]	64×2048 pixels	29.5	85.4	54.3	26.9	4.5	57.4	68.8	3.3	16.0	4.1	3.6	60.0	24.3	53.7	12.9	13.1	0.9	29.0	17.5	24.5
SqueezeSegV2 [26]		39.7	88.6	67.6	45.8	17.7	73.7	81.8	13.4	18.5	17.9	14.0	71.8	35.8	60.2	20.1	25.1	3.9	41.1	20.2	36.3
DarkNet53Seg [3]		49.9	**91.8**	74.6	64.8	**27.9**	84.1	86.4	25.5	24.5	32.7	22.6	78.3	50.1	64.0	36.2	33.6	4.7	55.0	38.9	52.2
RangeNet53++ [14]		52.2	**91.8**	**75.2**	**65.0**	27.8	87.4	91.4	25.7	25.7	34.4	23.0	80.5	55.1	64.6	38.3	38.8	4.8	58.6	47.9	55.9
PolarNet [28]		54.3	90.8	74.4	61.7	21.7	90.0	93.8	22.9	40.3	30.1	28.5	**84.0**	**65.5**	67.8	43.2	40.2	5.6	61.3	51.8	**57.5**
Waterfall-Net	50K pts	**55.7**	91.0	75.0	63.0	23.4	**90.4**	93.0	30.8	40.5	**38.6**	30.0	82.3	58.2	**68.0**	**49.9**	47.6	5.6	**63.9**	49.6	57.1

S3DIS: We use the 6-fold cross-validation for fair comparison, following previous methods [7,16,17]. Results are reported in Table 4. Waterfall-Net outperforms RandLA-Net in all criteria and achieves superior results over previous point-based methods on mACC and mIoU. Compared with RandLA-Net, Waterfall-Net obtains obvious improvements in a large plane (ceil, floor, wall) and complex geometry (table, chair). It means waterfall feature aggregation can improve the categories that need clues of different granularity. In other words, the multiple density property is more sufficiently utilized in Waterfall-Net. Figure 4 displays some results in S3DIS. Some elaborate objects surrounded by large objects are misclassified by RandLA-Net, while the Waterfall-Net can handle these issues properly.

SemanticKitti: We follow the official split of training and testing set and evaluate the results in the competition server. Results are presented in Table 5. The Waterfall-Net outperforms all the point-based methods and improves the performance of RandLA-Net in most categories. It is inferior to RandLA-Net in the categories that have few samples as the number of samples in SemanicKitti is unbalanced. The topic of unbalance sample is another tough issue in point cloud analysis but out of the scope of our research.

5 Conclusion

In this paper, we present a Waterfall-Net to take advantage of the density property of different categories. It extracts more informative features with cascaded sub-networks. The sub-networks are connected via Sub-Network Fusing and the sub-stages in sub-networks are connected via Sub-Stage Fusing. Such a waterfall feature aggregation strategy provides more abundant semantics-and-density feature combinations. Quantitative experimental results and analysis on S3DIS and SemanticKitti demonstrate the effectiveness of our method.

References

1. Adelson, E.H., Anderson, C.H., Bergen, J.R., Burt, P.J., Ogden, J.M.: Pyramid methods in image processing. RCA Eng. **29**(6), 33–41 (1984)
2. Armeni, I., et al.: 3d semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1534–1543 (2016)
3. Behley, J., et al.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9297–9307 (2019)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
5. Cheng, M., Hui, L., Xie, J., Yang, J., Kong, H.: Cascaded non-local neural network for point cloud semantic segmentation. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8447–8452. IEEE (2020)

6. Engelmann, F., Kontogianni, T., Schult, J., Leibe, B.: Know what your neighbors do: 3d semantic segmentation of point clouds. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11131, pp. 395–409. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11015-4_29
7. Hu, Q., et al.: RandLA-Net: efficient semantic segmentation of large-scale point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11108–11117 (2020)
8. Jiang, M., Wu, Y., Zhao, T., Zhao, Z., Lu, C.: PointSIFT: a sift-like network module for 3d point cloud semantic segmentation. arXiv preprint arXiv:1807.00652 (2018)
9. Ke, L., Chang, M.-C., Qi, H., Lyu, S.: Multi-scale structure-aware network for human pose estimation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11206, pp. 731–746. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01216-8_44
10. 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, pp. 4558–4567 (2018)
11. Li, H., Xiong, P., Fan, H., Sun, J.: DFANet: deep feature aggregation for real-time semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9522–9531 (2019)
12. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B.: PointCNN: convolution on χ-transformed points. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 828–838 (2018)
13. Liu, Y., Fan, B., Xiang, S., Pan, C.: Relation-shape convolutional neural network for point cloud analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8895–8904 (2019)
14. Milioto, A., Vizzo, I., Behley, J., Stachniss, C.: Rangenet++: fast and accurate lidar semantic segmentation. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4213–4220. IEEE (2019)
15. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
16. 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, pp. 652–660 (2017)
17. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: deep hierarchical feature learning on point sets in a metric space. Adv. Neural Inf. Process. Syst. **30**, 1–10 (2017)
18. 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, pp. 3887–3896 (2018)
19. 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/CVF International Conference on Computer Vision, pp. 6411–6420 (2019)
20. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 648–656 (2015)
21. Wang, J., et al.: Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **43**, 3349–3364 (2020)

22. Wang, L., Huang, Y., Hou, Y., Zhang, S., Shan, J.: Graph attention convolution for point cloud semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10296–10305 (2019)
23. 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, pp. 2589–2597 (2018)
24. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph CNN for learning on point clouds. ACM Trans. Graph. (TOG) **38**(5), 1–12 (2019)
25. Wu, B., Wan, A., Yue, X., Keutzer, K.: SqueezeSeg: convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3d lidar point cloud. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 1887–1893. IEEE (2018)
26. Wu, B., Zhou, X., Zhao, S., Yue, X., Keutzer, K.: SqueezeSegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 4376–4382. IEEE (2019)
27. Yan, X., Zheng, C., Li, Z., Wang, S., Cui, S.: PointaSNL: robust point clouds processing using nonlocal neural networks with adaptive sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5589–5598 (2020)
28. Zhang, Y., et al.: PolarNet: an improved grid representation for online lidar point clouds semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9601–9610 (2020)
29. Zhang, Z., Hua, B.S., Yeung, S.K.: ShellNet: efficient point cloud convolutional neural networks using concentric shells statistics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1607–1616 (2019)
30. Zhao, H., Jiang, L., Fu, C.W., Jia, J.: PointWeb: enhancing local neighborhood features for point cloud processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5565–5573 (2019)
31. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16259–16268 (2021)
32. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
33. Zhao, L., Tao, W.: JSNet: joint instance and semantic segmentation of 3d point clouds. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12951–12958 (2020)
34. Zhiheng, K., Ning, L.: PyramNet: point cloud pyramid attention network and graph embedding module for classification and segmentation. arXiv preprint arXiv:1906.03299 (2019)