2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Multi-to-Single Knowledge Distillation for Point Cloud Semantic Segmentation

Shoumeng Qiu[1], Feng Jiang[1], Haiqiang Zhang[2], Xiangyang Xue[1], Jian Pu[1*]

***Abstract*— 3D point cloud semantic segmentation is one of the fundamental tasks for environmental understanding. Although significant progress has been made in recent years, the performance of classes with few examples or few points is still far from satisfactory. In this paper, we propose a novel multi-to-single knowledge distillation framework for the 3D point cloud semantic segmentation task to boost the performance of those hard classes. Instead of fusing all the points of multi-scans directly, only the instances that belong to the previously defined hard classes are fused. To effectively and sufficiently distill valuable knowledge from multi-scans, we leverage a multilevel distillation framework, i.e., feature representation distillation, logit distillation, and affinity distillation. We further develop a novel instance-aware affinity distillation algorithm for capturing high-level structural knowledge to enhance the distillation efficacy for hard classes. Finally, we conduct experiments on the SemanticKITTI dataset, and the results on both the validation and test sets demonstrate that our method yields substantial improvements compared with the baseline method. The code is available at `https://github.com/skyshoumeng/M2SKD`.**

I. INTRODUCTION

3D point cloud semantic segmentation aims to classify every point of a given scan into a certain class, which is crucial for various applications such as the navigation of autonomous vehicles [1], [2]. Although significant progress has been made in recent years [3]–[6], with increased distance to the sensor, almost all approaches perform worse with sparse point clouds [7]. This is especially true for classes with few examples or only a small number of points in a single scan, such as motorcyclists, trucks and poles. The task becomes very hard even for human eyes. However, in realistic situations, such as the autonomous driving scenario, accurate segmentation of sparse objects can be of great importance. For example, an object with few points, such as pedestrians and bicyclists, could be seriously affected by incorrect segmentation and can eventually lead to crashing into curbs and other road traffic accidents.

In recent years, several studies have been proposed to address the distance-dependent sparsity problem. For example, SqueezeSeg [8] was proposed to generate a denser range image, which meant that all point cloud information was retained in the range image. However, since it adopted the commonly used 2D convolution operations for feature extraction, it could not explore the 3D geometric pattern very well. The Cylinder3D framework [9] proposed a cylindrical partition and a symmetrical 3D convolution network to better explore the 3D geometric pattern and tackle the difficulties caused by sparsity and varying density. However, as the method lacked an explicit investigation of feature representation learning for sparse objects, the segmentation performance on these objects was still less than satisfactory. PVKD [10] proposed a point-to-voxel knowledge distillation approach for model compression. To enhance the distillation efficacy of the minority classes and distant objects, a difficulty-aware sampling strategy was employed to more frequently sample these hard classes. However, there were also shortcomings in this method. One of them was that there may be many background points in the supervoxels, which could cause disturbance in distillation learning for the hard classes. Another problem was that knowledge distillation was only performed in a single scan, lacking the exploration of sequential information.

Therefore, to boost the segmentation performance of instances with only a small number of points in a single scan, we propose a novel multi-to-single knowledge distillation framework for the 3D point cloud semantic segmentation task. Specifically, we try to generate more dense points for these hard instances by combining multi-scans into a single large point cloud. However, since the sequential information cannot be obtained in inference, the enhanced point cloud is only used as a supervision for better feature representation learning in model training. To this aim, we propose a simple yet effective priori-based sparse fusion strategy to make the model focus on the hard classes naturally and reduce the computational cost at the same time: instead of fusing all the points of multiple past scans directly, only the instances belonging to the previously defined classes (hard classes) are fused. We adopt point cloud registration technology to obtain more precise position alignment and accurate fusion results. For knowledge distillation, we leverage a multilevel distillation framework for knowledge translation, i.e., feature representation distillation, logit distillation, and affinity distillation. For affinity distillation, we develop a novel instance-aware affinity distillation method to capture high-level structural data more effectively for each instance, and the distillation efficiency for the hard classes can also be enhanced in the training procedure.

To evaluate the effectiveness of our proposed method, we conduct experiments on the SemanticKITTI dataset [7], and the results on both the validation and test sets demonstrate

This work was supported by Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01), ZJ Lab, and Shanghai Center for Brain Science and Brain-Inspired Technology, NSFC Project (62176061) and STCSM Project (No.22511105000).

[1]Fudan University, [2]Mogo Auto.

that our algorithm can outperform the baseline method by a large margin. We also conduct several ablation studies to examine the efficacy of each component. Qualitative and quantitative results are reported with a detailed description, and an analysis is also given for future works.

In summary, our contributions include the following:

1. We propose a novel multi-to-single knowledge distillation framework for 3D point cloud semantic segmentation.

2. We propose a simple yet effective priori-based sparse fusion strategy to make the model focus on hard classes naturally and enhance the distillation efficacy in training.

3. Multilevel knowledge distillation is adopted to fully exploit the information in the sequences, and a novel instance-aware affinity distillation is further proposed to better capture the high-level structural knowledge in each instance.

4. We conduct experiments on the SemanticKITTI dataset, and superior performance is achieved over the baseline method by a significant margin.

II. RELATED WORK

A. LiDAR Semantic Segmentation

3D point cloud semantic segmentation is a fundamental task for the navigation of autonomous vehicles. The distance-dependent sparsity of point clouds presents a great challenge to the semantic segmentation task. SqueezeSeg [8] proposed generating a denser range image by exploiting the way the rotating scanner captures the point cloud data, so the distance-dependent sparsity problem could be partially solved. 3DMiniNet [11] proposed learning a 2D representation from the raw points through a projection operation, which could effectively extract local and global information from the 3D data, showing promising and effective results. Cylinder3D [9] proposed a new framework to better explore the 3D geometric pattern and tackle these difficulties caused by sparsity and varying density through cylindrical partition and asymmetrical 3D convolution networks. (AF)2-S3Net [12] demonstrated a multi-branch attentive feature fusion module in the encoder and an adaptive feature selection module in the decoder that could simultaneously capture and emphasize fine details for smaller instances while focusing on global contexts embodied in larger instances. In [13], based on the observation that the distribution of objects is severely biased, it proposed a location-guided feature module to extract features with input-dependent convolutions in different regions. AMVNet [14] proposed a modular-and-hierarchical late fusion approach with an assertion-guided sampling strategy, where features of the uncertain points are sampled to a point head for more accurate predictions. RPVNet [15] proposed a deep fusion framework with a gated fusion module that aimed to utilize different view advantages and alleviate their own shortcomings in fine-grained segmentation tasks. Although the above models have shown impressive performance on various benchmarks, the performance of classes with few points is still far from satisfactory. By combining multi-scans into a single large point cloud, the point representation of objects can be enhanced. However, how to make use of sequential information to boost the segmentation performance on these objects with few points has not been sufficiently exploited.

B. Knowledge Distillation

Knowledge distillation was introduced in the seminal work [16], which aimed to minimize the KL divergence of soft class probabilities between the teacher and the student model. Knowledge distillation has been considered an effective approach for both model compression and model accuracy boosting.

SKD [17] presented two structured knowledge distillation strategies, pair-wise distillation and holistic distillation, and the pair-wise and high-order consistency was enforced between the outputs of the compact and cumbersome networks. GID [18] introduced relation-based knowledge for distillation on object detection tasks and further integrated it with response-based and feature-based knowledge, making the student model even surpass the teacher model. PVKD [10] proposed a point-to-voxel knowledge distillation approach and a difficulty-aware sampling strategy to enhance the distillation efficacy, specifically in hard cases. In [19], they proposed a new knowledge distillation method by reinterpreting the output from the teacher network to a re-represented latent domain, which made student model learning much easier, and it also proposed an affinity distillation module to help the student network capture long-term dependencies from the teacher network. WKD [20] proposed a novel knowledge distillation method that decomposed the images into different frequency bands with discrete wavelet transformation and then only distilled the high-frequency bands. Although previous distillation approaches have shown excellent performance in broad applications in machine learning, distillation between a single scan and sequential information has not been fully explored. To the best of our knowledge, we are the first to propose the multi-to-single scan point cloud knowledge distillation.

III. PROPOSED METHOD

In this section, we first present the details of our proposed priori-based sparse fusion strategy, and then the multilevel knowledge distillation approach is introduced. The overview of our proposed framework is shown in Figure 1.

A. Multi-scan Fusion

1) Priori-based Sparse Fusion: By exploiting information from a sequence of multiple past scans, the segmentation performance of the current scan can be improved. However, in the standard multi-scan fusion method, the adjacent scans are simply combined by the global pose translation. Such a simple combination approach has one obvious problem, i.e., the fusion result of the current scan contains a large number of point clouds, and large quantities of computational resources are needed. In fact, in the semantic segmentation task, the performance for some ground or structure categories is already fairly good, so the sequential information is of little help to the segmentation results of these categories.

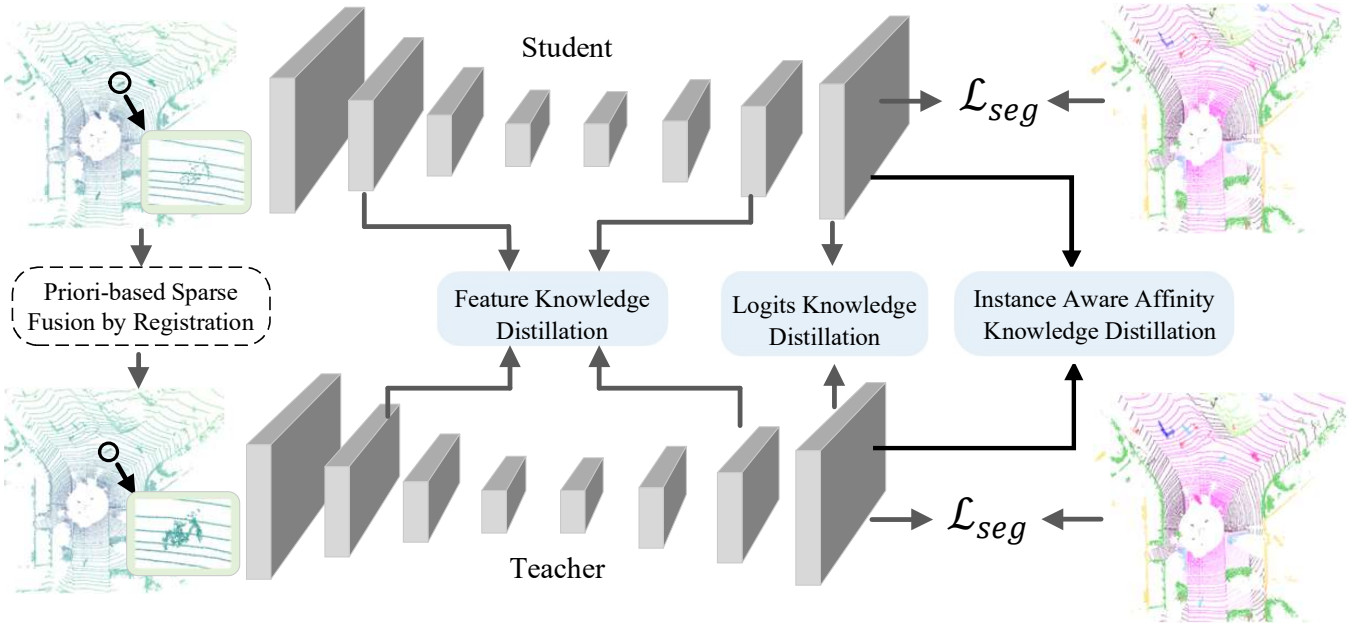

Fig. 1. The pipeline of our proposed multi-to-single knowledge distillation framework. There are two branches in our framework. One is the single-scan segmentation pipeline, and the other is the multi-scan fused segmentation pipeline. It should be noted that only the instances belonging to the previously defined hard classes are fused, and the points of multi-scans are precisely aligned by the point cloud registration algorithm. We leverage a multilevel knowledge distillation framework to make knowledge distillation more sufficient. Furthermore, an instance-aware affinity distillation approach is proposed to better capture high-level structural knowledge and enhance distillation efficacy for hard classes.

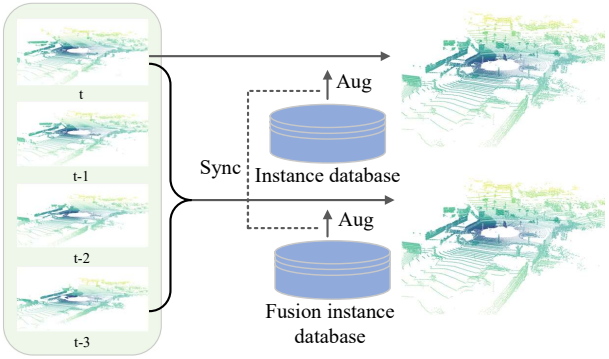

Fig. 2. Our copy-paste instance augmentation pipeline. For considerations of computational efficiency, first, data preprocessing is conducted for the whole dataset. Then the instances in the single scan and the corresponding fused instances are stored separately. During training, the instances in both the instance database and the fusion instance database are sampled synchronously and then pasted to the single-scan and multi-fused scans, respectively.

To address the abovementioned problem, we improve the conventional multi-scan fusion into a sparse fusion based on a priori. Specifically, instead of fusing all the points of past multi-scans directly, only the points belonging to the previously defined classes are fused. The classes with with a lower IoU are selected as the candidate classes to be enhanced by exploiting the sequential information. Our approach has two obvious advantages compared to conventional multi-scan fusion methods. One is the reduced computational cost of the fused point cloud, and the other is to force the model to focus on difficult samples in the knowledge distillation process.

For objects without instance IDs in the SemanticKITTI dataset [21], such as traffic signs, we proposed a simple method, for instance, ID generation. In this method, we only try to generate an instance ID for each object using semantic labels. However, as there may be many objects belonging to the same class in each scan, the algorithm must be carefully designed. Next, we provide a detailed description of the instance ID generation method. Specifically, we first filter the point cloud based on semantic labels, and only points that belong to the specific class are preserved. Then, the farthest point sampling is performed on the preserved points. If the distance between the points sampled currently and previously is less than a certain threshold, the sampling operation is stopped. Next, we take the sampled points as key points, and then a distance-based clustering algorithm is performed with these sampled points. Finally, the unique instance IDs were assigned to each cluster. The instance ID generation pipeline is shown in Figure 3.

To further boost the performance of the hard class, we also adopt the copy-paste instance augmentation approach [22], [23] in our knowledge distillation framework. The augmentation pipeline is shown in Figure 2.

Another challenging problem is to tackle the moving and non-moving classes in each scan, i.e., cars and buildings. The straightforward fusion method based on the global pose translation cannot handle moving objects. To solve this problem, we separate the fusion process into two different parts: moving objects and non-moving objects, respectively. Non-

moving object fusion can be simply achieved by the global pose translation and multi-scan point cloud combination. Because there is no pose translation information for each object, moving object fusion becomes more difficult. To cope with this problem, we adopt point cloud registration technologies to obtain more precise fusion results, and the details of the fusion by registration approach are presented in the following section.

2) Fusion by Registration: In the priori-based sparse fusion, one of the most straightforward ways is that for each moving instance, we simply move the point cloud of the corresponding instance in the adjacent scans to the current position. However, in this case, determining the precise location of the instance becomes crucial. We found experimentally that the commonly used centroid-based methods work well in most cases, and the centroid-based approach is also very efficient in practice. However, in regard to the sparser instances, such as the bicyclist and motorcyclist in the distance, or the car in a turn, the simple centroid-based alignment method will become fail.

To achieve a more precise fusion result, we adopt point registration technology to solve the misalignment problem in the fusion operation. Specifically, for each candidate instance in the current scan, we first gather all the corresponding instances in the adjacent scans, and then we use the centroid-based method as a fast way to provide a feasible initialization for the following registration algorithm. Finally, a more precise registration-based point cloud fusion operation is performed.

Noted that in the SemanticKITTI dataset, the instance ID is consistent in the adjacent scans, so we can use the ground truth data in the above registration alignment operation to find the same instance in the adjacent scans. In other datasets where a consistent instance ID cannot be obtained from the annotation data, one feasible way is to generate them using object-tracking methods.

B. Multilevel Knowledge Distillation

Different from the standard knowledge distillation task, which aims to facilitate the training of a lightweight student model under the supervision of a sophisticated teacher model, we try to distill the enhanced feature representation from the multi-scan combination input to the single scan input.

1) Feature Representation Distillation: To make the knowledge distillation from multi-scans more sufficient, we leverage a multilevel distillation framework for knowledge translation, i.e., feature representation distillation, logit distillation, and affinity distillation. For the feature-level knowledge distillation, we select the feature from two different middle layers of the segmentation network for knowledge distillation learning, i.e., the feature after the second down-sampling block and the feature after the third up-sampling block. As mentioned in subsection III-A.1, we aim to encourage the network to focus on the features of the predefined hard classes. Since only the classes whose segmentation performance is below a predefined threshold are fused by the priori-based sparse fusion operation, we simply constrain the knowledge distillation regions at the points of fusion areas. Finally, referring to the methods described in [24], we adopt the smooth-L1 distance between the student and teacher model outputs, which is directly optimized for feature-level knowledge distillation.

$$\mathcal{L}_{FD} = \begin{cases} \frac{1}{Nf_c}\sum_{i=1}^{N}\sum_{k=1}^{f_c}\frac{1}{2T}\|F_{(i,k)}-f_{(i,k)}\|^2 & \|F_{(i,k)}-f_{(i,k)}\|<T \\ \frac{1}{Nf_c}\sum_{i=1}^{N}\sum_{k=1}^{f_c}\frac{1}{2T}\|F_{(i,k)}-f_{(i,k)}\|-0.5T^2 & \text{otherwise} \end{cases}, \quad (1)$$

where N is the number of sampled hard points, f_c is the channels of the feature, $F_{(i,k)}$ and the $f_{(i,k)}$ are the features of the teacher model and the student model respectively, and T is a predefined threshold parameter.

2) Logits Distillation: In the logits output layer, knowledge distillation is also conducted. Similar to feature representation distillation, only the specific regions are considered in the logit distillation. Following much prior work, we use the Kullback-Leibler divergence loss to minimize the KL divergence of class probabilities between the multi-scan and single-scan outputs. Soft labels are always considered to contain much more information than one-hot labels, which can play the roles of supervisory signals and regularizations at the same time. Therefore, in our experiment, we adopt the softened version of the teacher and student model logits by multiplying the logits with a temperature parameter P.

$$\mathcal{L}_{SLD} = \frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C}\varphi(F_l(i,c))\frac{\varphi(F_l(i,c))}{\varphi(f_l(i,c))}, \quad (2)$$

where C is the channels of the logits, $\varphi()$ is the softmax function.

3) Affinity Distillation: Distilling the knowledge of the output features is insufficient since it considers only individual elements while the structural information of the surrounding environment is ignored. However, structural knowledge is crucial to the 3D point cloud semantic segmentation model as the input points are unordered, especially regarding objects with only a small number of points. One possible solution is the relational knowledge distillation technology proposed in [25]. Unlike pixel-wise feature representation knowledge distillation, relational knowledge distillation attempts to distill the pair-wise relationship of all point features. However, as mentioned in [10], the affinity distillation process became computationally expensive and extremely difficult to learn since the similarity matrix had too many elements. In [10], they improved the previous relational knowledge distillation approach by dividing the whole point cloud into several super voxels and only sampling part of them for the relation distillation learning. However, direct applying this method is also problematic, as we mainly focus on obtaining better feature representations of classes with only a small number of points, which is most likely to be objects such as bicyclists or motorcyclists. The supervoxel division mechanism is not quite suitable for these small objects. Hence, we develop a simple yet effective instance-aware affinity distillation

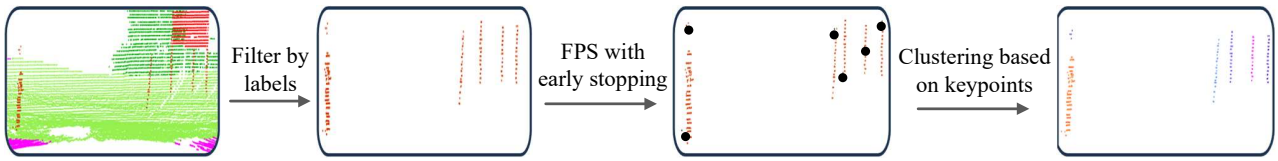

Fig. 3. The instance ID generation pipeline for specific classes.

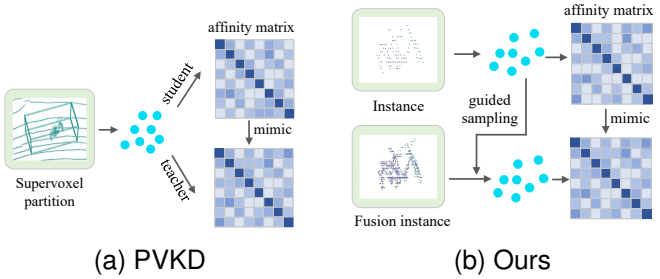

Fig. 4. Comparison between our proposed instance-aware affinity knowledge distillation approach and the affinity knowledge distillation method proposed in PVKD [10]. Note that the interference of background points is totally removed in our approach, and the high-level structural knowledge of each instance is captured more precisely and effectively.

strategy to overcome the above issues. We only calculate the pair-wise similarity of point clouds belonging to the same instances. As these objects always have only a small number of points, the simpler similarity matrix is computationally efficient and much easier to learn. In addition, the model can focus on each instance in the distillation process automatically. The comparison between our instance-aware affinity distillation approach and the affinity distillation proposed in PVKD is shown in Figure 4.

The instance-aware affinity matrix is calculated according to the following equation:

$$A_{matrix}(i,j) = \frac{Fea(i)^T Fea(j)}{\|Fea(i)\|^2 \|Fea(j)\|^2} \quad i,j \in \mathcal{S}_k, \tag{3}$$

where Fea denotes the features of the teacher model or the student model and $\mathcal{S}_k$ is the point set of the kth instance.

Then, the affinity distillation loss is formulated as:

$$\mathcal{L}_{IAAD} = \sum_{k=1}^{|\mathcal{S}|} \frac{1}{|\mathcal{S}_k|^2} \sum_{i=1}^{|\mathcal{S}_k|} \sum_{j=1}^{|\mathcal{S}_k|} \|A_m(i,j) - A_s(i,j)\|^2. \tag{4}$$

C. The Final Objective Function

Our final loss function is composed of four terms, i.e., the main segmentation task loss for the point-wise prediction of both the student and teacher models, the smooth-L1 loss for the feature representation distillation, the Kullback-Leibler divergence loss for soft logits distillation, and the 12 loss for the instance-aware affinity distillation:

$$\begin{aligned} \mathcal{L}_{final} =& \mathcal{L}_{seg}^S + \beta_1 \mathcal{L}_{seg}^T \\ &+ \beta_2 \mathcal{L}_{FD} + \beta_3 \mathcal{L}_{SLD} + \beta_4 \mathcal{L}_{IAAD}, \end{aligned} \tag{5}$$

where β_1, β_2, β_3, and β_4 are the loss coefficients used to balance the contribution of the knowledge distillation loss to the main segmentation task loss. In our experiments, if not otherwise stated, β_1, β_2 and β_3 are set as 0.5, 0.01, 0.1, and 0.1.

IV. EXPERIMENTAL RESULTS

A. Datasets

SemanticKITTI [7] is a large-scale dataset for semantic scene understanding using LiDAR sequences, which is based on the KITTI Vision Benchmark and has a dense semantic annotation for the entire KITTI Odometry Benchmark, making it a standard dataset for evaluating LiDAR semantic segmentation methods. This dataset presents challenges on rare classes such as motorcyclists and other-ground due to the limited training examples. The most frequent class, "vegetation", has 4.82×10^7 times more points than the least frequent class, "motorcyclist", which is heavily imbalanced. In our experiments, we defined the bicycle, motorcycle, truck, other-vehicle, person, bicyclist, motorcyclist, and traffic-sign as the hard classes, and we adopt the method in [26] for point cloud registration.

B. Main Results

As official guidance suggests, we use mean intersection-over-union (mIoU) over all classes as the evaluation metric. The metric can be formalized as follows:

$$mIoU = \frac{1}{n} \sum_{c=1}^{n} \frac{TP_c}{TP_c + FP_c + FN_c}, \tag{6}$$

where TP_c denotes the number of true positive points for class c, FP_c denotes the number of false positives, and FN_c is the number of false negatives. As the name suggests, the final IoUs are calculated for each class separately and then the mean is taken.

To verify the effectiveness of our method, we apply our method to both the PolarNet [37] and Cylinder3D [9] baselines. The experimental results are shown in Table I. Note that the Cylinder3D* represents the result reproduced by the model released by the author without test augmentation. For a fair comparison, our model is only trained on the training set. It can be seen that we surpass the PolarNet baseline by 3.7% mIoU, and surpass the Cylinder3D baseline by 1.7% mIoU. Specifically, for hard classes such as bicyclist and motorcyclist, the performance gain of our proposed multi-to-single knowledge distillation framework becomes more significant, we surpass the PolarNet baseline by 11.2% and

TABLE I

QUANTITATIVE COMPARISON OF CYLINDER3D TRAINED WITH OUR FRAMEWORK. THE RESULTS ARE REPORTED IN TERMS OF THE MIOU ON THE SEMANTICKITTI TEST SET. ∗ REPRESENTS THE RESULT REPRODUCED BY THE OFFICIALLY RELEASED CODE AND MODELS. FROM TOP TO BOTTOM, THE METHODS ARE GROUPED INTO POINT-BASED, PROJECTION-BASED, AND MULTI-VIEW FUSION MODELS.

Methods	*car*	*bicycle*	*motorcycle*	*truck*	*other-vehicle*	*person*	*bicyclist*	*motorcyclist*	*road*	*parking*	*sidewalk*	*other-ground*	*building*	*fence*	*vegetation*	*trunk*	*terrain*	*pole*	*traffic-sign*	mIoU
LatticeNet [27]	88.6	12.0	20.8	43.3	24.8	34.2	39.9	60.9	88.8	64.6	73.8	25.5	86.9	55.2	76.4	67.9	54.7	41.5	42.7	51.3
PointNL [28]	92.1	42.6	37.4	9.8	20.0	49.2	57.8	28.3	90.5	48.3	72.5	19.0	81.6	50.2	78.5	54.5	62.7	41.7	55.8	52.2
RandLa-Net [3]	94.2	26.0	25.8	40.1	38.9	49.2	48.2	7.2	90.7	60.3	73.7	20.4	86.9	56.3	81.4	61.3	66.8	49.2	47.7	53.9
KPConv [5]	96.0	30.2	42.5	33.4	44.3	61.5	61.6	11.8	88.8	61.3	72.7	31.6	90.5	64.2	84.8	69.2	69.1	56.4	47.4	58.8
SqueezeSegV2 [8]	82.7	21.0	22.6	14.5	15.9	20.2	24.3	2.9	88.5	42.4	65.5	18.7	73.8	41.0	68.5	36.9	58.9	12.9	41.0	39.6
RangeNet53 [29]	91.4	25.7	34.4	25.7	23.0	38.3	38.8	4.8	91.8	65.0	75.2	27.8	87.4	58.6	80.5	55.1	64.6	47.9	55.9	52.2
3D-MiniNet-KNN [11]	90.5	42.3	42.1	28.5	29.4	47.8	44.1	14.5	91.6	64.2	74.5	25.4	89.4	60.8	82.8	60.8	66.7	48.0	56.6	55.8
SqueezeSegV3 [30]	92.5	38.7	36.5	29.6	33.0	45.6	46.2	20.1	91.7	63.4	74.8	26.4	89.0	59.4	82.0	58.7	65.4	49.6	58.9	55.9
SalsaNext [31]	91.9	48.3	38.6	38.9	31.9	60.2	59.0	19.4	91.7	63.7	75.8	29.1	90.2	64.2	81.8	63.6	66.5	54.3	62.1	59.5
KPRNet [32]	95.5	54.1	47.9	23.6	42.6	65.9	65.0	16.5	93.2	73.9	80.6	30.2	91.7	68.4	85.7	69.8	71.2	58.7	64.1	63.1
MVLidarNet [33]	87.1	34.9	32.9	23.7	24.9	44.5	44.3	23.1	90.3	56.7	73.0	19.1	85.6	53.0	80.9	59.4	63.9	49.9	51.1	52.5
MPF [34]	93.4	30.2	38.3	26.1	28.5	48.1	46.1	18.1	90.6	62.3	74.5	30.6	88.5	59.7	83.5	59.7	69.2	49.7	58.1	55.5
TORNADONet [35]	94.2	55.7	48.1	40.0	38.2	63.6	60.1	34.9	89.7	66.3	74.5	28.7	91.3	65.6	85.6	67.0	71.5	58.0	65.9	63.1
AMVNet [14]	96.2	59.9	54.2	48.8	45.7	71.0	65.7	11.0	90.1	71.0	75.8	32.4	92.4	69.1	85.6	71.7	69.6	62.7	67.2	65.3
GFNet [36]	96.0	53.2	48.3	31.7	47.3	62.8	57.3	44.7	93.6	72.5	80.8	31.2	94.0	73.9	85.2	71.1	69.3	61.8	68.0	65.4
PolarNet [37]	93.8	40.3	30.1	22.9	28.5	43.2	40.2	5.6	90.8	61.7	74.4	21.7	90.0	61.3	84.0	65.5	67.8	51.8	57.5	54.3
PolarNet+Ours	93.5	45.9	36.3	27.6	34.9	55.0	51.4	15.8	91.1	64.7	73.8	26.1	92.5	67.0	84.6	63.4	67.4	50.7	59.5	58.0
Cylinder3D* [9]	96.7	60.1	57.4	43.2	49.6	70.0	65.1	12.0	91.6	64.6	76.0	24.3	90.0	63.4	84.8	70.7	67.6	62.0	64.0	63.9
Cylinder3D+Ours	96.4	60.8	54.8	42.8	51.2	69.1	67.8	34.8	92.2	66.5	76.7	30.4	91.1	65.7	85.5	69.8	68.6	60.7	61.0	65.6

TABLE II

ABLATION STUDY OF EACH COMPONENT ON THE FINAL PERFORMANCE ON THE SEMANTICKITTI VALIDATION SET.

Cylinder3D	PWKD	IAAD	mIoU
√			64.1
√	√		65.6
√	√	√	66.2

10.2%, and surpass the Cylinder3D baseline by 2.7% and 22.8% respectively.

C. Ablation Study

We performed ablation studies to examine the efficacy of the major components of the proposed method. The ablation results are shown in Table II. We reproduce the results of Cylinder3D by the officially released code. PWKD denotes the point-wise knowledge distillation, including the feature distillation and the logit distillation, IAAD denotes the instance-aware affinity distillation.

D. Visualization

Finally, we also visualize the performance gain of our approach in different scans and mainly focus on two of the hard classes, including truck and bicyclist. The result is shown in Figure 5. Note that in scene A, our model gives better results on the prediction of the truck segmentation. In scene B, the baseline model misclassified a bicyclist as a person, as these two classes are very similar; however, our method gives the correct predictions.

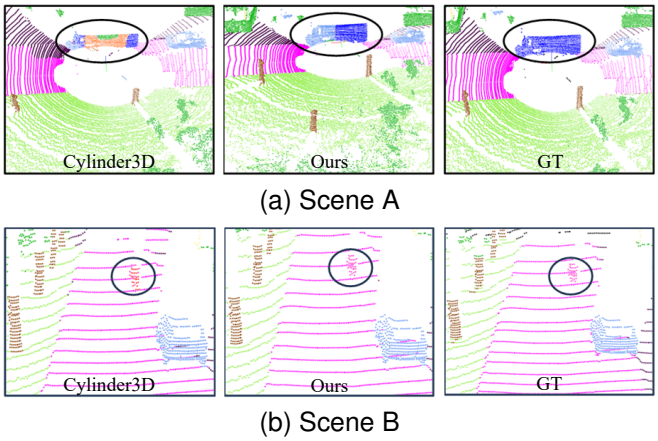

Fig. 5. Visualization comparison of the baseline (Cylinder3D) and our proposed method on the SemanticKITTI dataset.

V. CONCLUSIONS

In this work, we have developed a simple but effective multi-to-single knowledge distillation framework for the 3D point cloud semantic segmentation task. By training the model with supervision from multiple adjacent scans combined, the segmentation performance of the single scan can be improved. We adopt the priori-based sparse fusion strategy to fuse only the predefined hard classes, which can make the model focus on the hard classes naturally and enhance the distillation efficacy in the training procedure. We also propose an instance-aware affinity distillation to better capture the high-level structural knowledge in each object in knowledge distillation. The multi-to-single knowledge distillation framework can also be considered as a plug-in component and can be integrated into the recently developed point cloud semantic segmentation approaches. We conduct experiments on the SemanticKITTI dataset and the results demonstrate that our algorithm can outperform the baseline method by a significant margin. One of our future works is to extend our proposed approach to other 3D point cloud understanding tasks, such as 3D object detection.

REFERENCES

[1] J. Zhang, X. Zhao, Z. Chen, and Z. Lu, "A review of deep learning-based semantic segmentation for point cloud," *IEEE Access*, vol. 7, pp. 179 118–179 133, 2019.

[2] H. Lu and H. Shi, "Deep learning for 3d point cloud understanding: a survey," *arXiv preprint arXiv:2009.08920*, 2020.

[3] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, and A. Markham, "Randla-net: Efficient semantic segmentation of large-scale point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 108–11 117.

[4] H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, and S. Han, "Searching efficient 3d architectures with sparse point-voxel convolution," in *European conference on computer vision*. Springer, 2020, pp. 685–702.

[5] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6411–6420.

[6] Y. Ren, S. Zhao, and L. Bingbing, "Object insertion based data augmentation for semantic segmentation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 359–365.

[7] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9297–9307.

[8] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer, "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)*. IEEE, 2019, pp. 4376–4382.

[9] X. Zhu, H. Zhou, T. Wang, F. Hong, Y. Ma, W. Li, H. Li, and D. Lin, "Cylindrical and asymmetrical 3d convolution networks for lidar segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9939–9948.

[10] Y. Hou, X. Zhu, Y. Ma, C. C. Loy, and Y. Li, "Point-to-voxel knowledge distillation for lidar semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8479–8488.

[11] I. Alonso, L. Riazuelo, L. Montesano, and A. C. Murillo, "3d-mininet: Learning a 2d representation from point clouds for fast and efficient 3d lidar semantic segmentation," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5432–5439, 2020.

[12] R. Cheng, R. Razani, E. Taghavi, E. Li, and B. Liu, "2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 547–12 556.

[13] G. Xian, C. Ji, L. Zhou, G. Chen, J. Zhang, B. Li, X. Xue, and J. Pu, "Location-guided lidar-based panoptic segmentation for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, 2022.

[14] V. E. Liong, T. N. T. Nguyen, S. Widjaja, D. Sharma, and Z. J. Chong, "Amvnet: Assertion-based multi-view fusion network for lidar semantic segmentation," *arXiv preprint arXiv:2012.04934*, 2020.

[15] J. Xu, R. Zhang, J. Dou, Y. Zhu, J. Sun, and S. Pu, "Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 024–16 033.

[16] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[17] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2604–2613.

[18] X. Dai, Z. Jiang, Z. Wu, Y. Bao, Z. Wang, S. Liu, and E. Zhou, "General instance distillation for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7842–7851.

[19] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 578–587.

[20] L. Zhang, X. Chen, X. Tu, P. Wan, N. Xu, and K. Ma, "Wavelet knowledge distillation: Towards efficient image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 464–12 474.

[21] M. Aygun, A. Osep, M. Weber, M. Maximov, C. Stachniss, J. Behley, and L. Leal-Taixé, "4d panoptic lidar segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5527–5537.

[22] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2918–2928.

[23] Z. Zhou, Y. Zhang, and H. Foroosh, "Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[24] L. Du, X. Ye, X. Tan, E. Johns, B. Chen, E. Ding, X. Xue, and J. Feng, "Ago-net: Association-guided 3d point cloud object detection network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[25] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976.

[26] H. Yang and L. Carlone, "A quaternion-based certifiably optimal solution to the wahba problem with outliers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1665–1674.

[27] R. A. Rosu, P. Schütt, J. Quenzel, and S. Behnke, "Latticenet: Fast point cloud segmentation using permutohedral lattices," *arXiv preprint arXiv:1912.05905*, 2019.

[28] M. Cheng, L. Hui, J. Xie, J. Yang, and H. Kong, "Cascaded non-local neural network for point cloud semantic segmentation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8447–8452.

[29] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss, "Rangenet++: Fast and accurate lidar semantic segmentation," in *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2019, pp. 4213–4220.

[30] C. Xu, B. Wu, Z. Wang, W. Zhan, P. Vajda, K. Keutzer, and M. Tomizuka, "Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 1–19.

[31] T. Cortinhal, G. Tzelepis, and E. Erdal Aksoy, "Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds," in *International Symposium on Visual Computing*. Springer, 2020, pp. 207–222.

[32] D. Kochanov, F. K. Nejadasl, and O. Booij, "Kprnet: Improving projection-based lidar semantic segmentation," *arXiv preprint arXiv:2007.12668*, 2020.

[33] K. Chen, R. Oldja, N. Smolyanskiy, S. Birchfield, A. Popov, D. Wehr, I. Eden, and J. Pehserl, "Mvlidarnet: Real-time multi-class scene understanding for autonomous driving using multiple views," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 2288–2294.

[34] Y. A. Alnaggar, M. Afifi, K. Amer, and M. ElHelw, "Multi projection fusion for real-time semantic segmentation of 3d lidar point clouds," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1800–1809.

[35] M. Gerdzhev, R. Razani, E. Taghavi, and L. Bingbing, "Tornado-net: multiview total variation semantic segmentation with diamond inception module," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 9543–9549.

[36] H. Qiu, B. Yu, and D. Tao, "Gfnet: Geometric flow network for 3d point cloud semantic segmentation," *arXiv preprint arXiv:2207.02605*, 2022.

[37] Y. Zhang, Z. Zhou, P. David, X. Yue, Z. Xi, B. Gong, and H. Foroosh, "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*, 2020, pp. 9601–9610.