2022 IEEE International Conference on Robotics and Automation (ICRA)
May 23-27, 2022. Philadelphia, PA, USA

Prototype-Voxel Contrastive Learning for LiDAR Point Cloud Panoptic Segmentation

Minzhe Liu[1], Qiang Zhou[2], Hengshuang Zhao[3], Jianing Li[1],
Yuan Du[1], Kurt Keutzer[4], Li Du[1*], Shanghang Zhang[5*]

***Abstract*— LiDAR point cloud panoptic segmentation, including both semantic and instance segmentation, plays a critical role in meticulous scene understanding for autonomous driving. Existing 3D voxelized approaches either utilize 3D sparse convolution that only focuses on local scene understanding, or add extra and time-consuming PointNet branch to capture global feature structures. To address these limitations, we propose an end-to-end Prototype-Voxel Contrastive Learning (PVCL) framework for learning stable and discriminative semantic representations, which includes voxel-level and prototype-level contrastive learning (CL). The voxel-level CL decreases intra-class distance and increases inter-class distance among sample representations, while the prototype-level CL further reduces the dependence of CL on negative sampling and avoids the influence of outliers from the same class, enabling PVCL to be more effective for outdoor point cloud panoptic segmentation. Extensive experiments are conducted on the public point cloud panoptic segmentation datasets, Semantic-KITTI and nuScenes, where evaluations and ablation studies demonstrate PVCL achieves superior performance compared with the state-of-the-art. Our approach ranks the top on the public leaderboard of Semantic-KITTI at the time of submission, and surpasses the published 2nd rank, EfficientLPS, by 1.7% in PQ.**

I. INTRODUCTION

LiDAR is one of the most important sensors for autonomous driving and other types of robots. Various scene perception tasks, such as 3D object detection [1], [2], [3], [4], [5], [6], [7], [8], [9], [10] and semantic segmentation [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], have been actively researched on large-scale outdoor data sets. However, these tasks are unable to meet the more holistic perception required by autonomous driving. Recently, LiDAR point cloud panoptic segmentation [21] is proposed in holistic perception as it can combine both semantic segmentation and instance segmentation. Especially, it can predict both semantic labels for the points in the scene and instance IDs for the points that belong to countable objects such as vehicles and people. However, such task is extremely challenging as the sparse and irregular space structure makes it very difficult for highly fine-grained scene understanding. Existing methods [22], [23], [24], [25], [26] can be mainly divided into two groups: one is to project a point cloud scan into 2D representation [22], [25]; the other is to voxelize the points into 3D voxels [23], [24], [26]. The projection-based method destroys the 3D spatial structure of the point cloud, and its post-process is quite complicated and time-consuming. The voxel-based method has demonstrated improved representation learning performance of point clouds, however, these methods merely focus on extracting local features and neglect the global context of the entire scene. Several methods [12], [27] add extra PointNet [28] branches to obtain a supplementary global feature structure, however, they also tremendously increase the network parameters and lower the computation efficiency. Due to the sparse and non-uniform distribution of outdoor point clouds, clustering-based methods for indoor 3D instance segmentation [29], [30], [25], [23], [24] also become ineffective. Therefore, a question is naturally raised: Is there a more effective and efficient way of point cloud representation learning to improve panoptic segmentation?

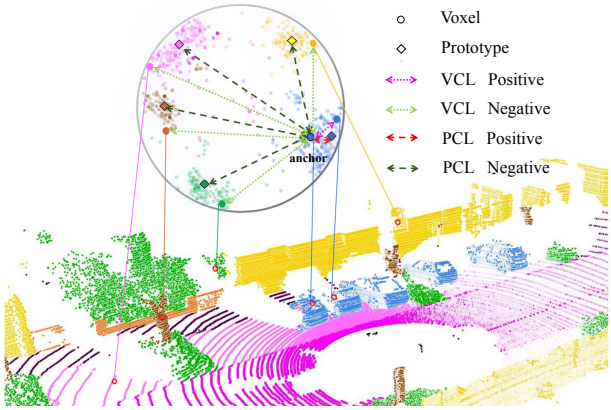

Fig. 1. Illustration of Prototype-Voxel Contrastive Learning module. It includes both voxel-level and prototype-level supervised contrastive learning. The colored circle represents the voxel and the colored diamond represents the prototype, which is the feature centroid of each semantic class group.

To address the limitations of existing works and achieve more effective and efficient panoptic segmentation, we propose an end-to-end Prototype-Voxel Contrastive Learning (PVCL) framework to learn more stable and discriminative semantic representations, which includes both voxel-level and prototype-level contrastive learning (CL). It is observed that conventional unsupervised contrastive learning may introduce sampling bias and suffer from performance decay because it may have negative samples from the same class [31]. The recent work Debiased CL [31] demonstrates both empirically and theoretically that such bias leads to significant performance drop compared with sampling nega-

This work was supported in part by the National Natural Science Foundation of China under Grant 62004097 and Grant 62004096.

[1]School of Electronic Science and Engineering, Nanjing University. [2]Institute for AI Industry Research (AIR), Tsinghua University. [3]The University of Hong Kong. [4]UC Berkeley. [5]Peking University.
*The corresponding authors of this paper. (E-mail: ldu@nju.edu.cn; shanghang@pku.edu.cn)

2022 IEEE International Conference on Robotics and Automation (ICRA) | 978-1-7281-9681-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRA46639.2022.9811638

tive samples only from truly different classes. To avoid such bias introduced by conventional CL [32], [33], we propose to explore the supervised voxel-level contrastive learning (VCL) [34] to learn better representations and improve the performance on the downstream tasks. As illustrated in Figure. 1, VCL draws the negative samples only from the classes that are different from the anchor and draws the positive samples from the same class of the anchor in a supervised learning manner. Yet compared with traditional supervised learning, our method can learn more robust and discriminative representations that have decreased intra-class distance and increased inter-class distance, verified by our empirical study. Our approach is especially beneficial to the outdoor LiDAR point cloud scenario since it more effectively encodes the feature correlations among multiple global points that cannot be covered by the conventional local convolutional operators.

According to previous study [32], [33], [35], potent CL depends on the amount and hardness of negative samples, and a large number of hard negative samples lead to better performance of CL. However, it is quite difficult and memory-consuming to select these hard negative samples. To address this issue, we further develop the prototype-level CL (PCL). PCL encourages each anchor to be closer to its assigned prototype (centroid of each class in the feature space), and farther away from different semantic prototypes, as illustrated in Figure. 1. Since the positive/negative samples are based on the prototype of each class in the scene, PCL alleviates the requirement for the hard sampling of the positive/negative pairs. Another advantage of PCL is that it avoids the influence of outliers and noisy samples of each semantic class and neutralizes the instability of the sampling process, which makes our method especially more robust to the complex outdoor LiDAR point cloud scenes. The design of PCL benefits to learning discriminative features of different semantic classes and avoids an excessive overlap of different semantic classes with similar characteristics. As a strong verification, the ablation study shows that combining both VCL and PCL outperforms VCL or PCL alone .

To this end, we develop the Prototype-Voxel CL with bi-level VCL and PCL. We build the PVCL module on top of a backbone design to form a novel end-to-end panoptic segmentation framework, as shown in Figure. 2. Such framework is innovative because PVCL is embedded in the supervised panoptic segmentation pipeline instead of working as a pretrain model. It efficiently constructs the feature connections among different points globally and addresses the challenge of point cloud sparsity, which essentially strengthens the representation learning for point cloud. To verify PVCL, we conduct extensive experiments on challenging datasets, Semantic-KITTI [36] and nuScenes [37]. Experiments and ablation study show PVCL significantly improves the performance of panoptic segmentation and outperforms the state-of-the-arts. Our method wins the 1st rank on the public leaderboard of Semantic-KITTI at the time of submission and surpasses EfficientLPS [22] by 1.7% in PQ, and DS-Net [24] by 3.2% in PQ (Panoptic Quality). PVCL achieves an even greater advantage in instance segmentation and outperforms EfficientLPS by 6.7% in PQ^{th}, and DS-Net by 4.7% in PQ^{th}. Our contributions are summarized as follows:

- We propose an end-to-end PVCL framework for point cloud panoptic segmentation, including both voxel-level and prototype-level CL. Such bi-level CL not only decreases intra-class distance and increases inter-class distance among sample representations, but also encourages representations to be closer to their assigned prototypes. To the best of our knowledge, it is the first CL framework for point cloud panoptic segmentation.
- PVCL contains a novel positive/negative pair construction strategy for outdoor point cloud scenes, allowing most points in the scene to participate in the representation learning and leading to a more stable model than normal random sampling.
- The proposed VCL avoids the bias introduced by conventional CL and considers many positives per anchor in addition to many negatives, leading to more balanced training and superior performance.
- The proposed PCL reduces the dependence of CL on negative sampling, and avoids the influence of outliers from the same class, enabling our method more effective for outdoor point cloud panoptic segmentation.
- Extensive experiments verify PVCL significantly outperforms the state-of-the-arts, and wins the 1st rank on the public leaderboard of Semantic-KITTI at the time of submission.

II. RELATED WORK

Point Cloud Panoptic Segmentation. Existing 3D panoptic segmentation works [21], [22], [23], [24], [25], [26] were generally similar in the overall framework, where the point-wise or voxel-wise features are learned through a feature extraction network and then mapped to the semantic segmentation branch and instance segmentation branch to obtain semantic class probability and coordinate offset respectively. Panoptic RangeNet [25] utilized RangeNet_{++} [38] as its backbone and proposed a range-image-based trilinear upsampling module. Panoster [23] proposed a simplified framework to incorporate a learning-based clustering solution. DS-Net [24] adopted cylindrical convolution for strong backbone design. EfficientLPS [22] proposed PCM convolution module that reshapes the convolution grid to capture local contextual information from neighboring pixels. Our PVCL outperforms these methods as shown in the empirical study.

Contrastive Learning. Contrastive learning has triggered its research on various tasks, achieving state-of-the-art performance on representation learning [39], [40]. In 2D representation learning, the representative works include SimCLR [7], MoCo [33], and ProtoNCE [41] While there are active researches on 2D contrastive learning, much fewer works have been done on 3D. PointContrast [35] is the first to propose unsupervised contrastive learning for 3D representation learning. Nevertheless, without access to labels, dissimilar (negative) points may be from the same class of the anchor, inducing bias to contrastive learning [31]. Apart

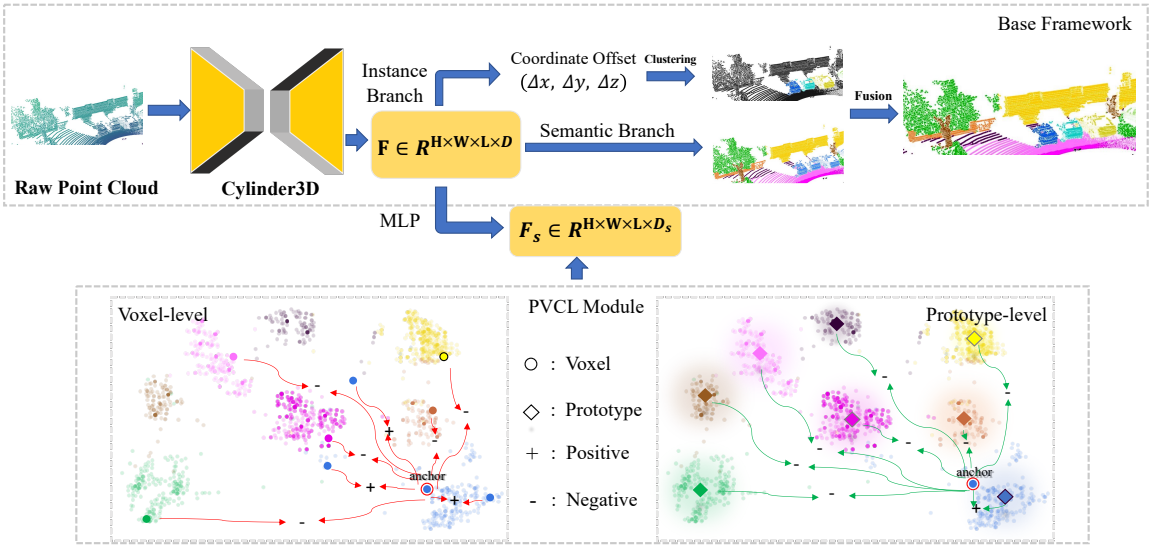

Fig. 2. **Detailed illustration** of our PVCL panoptic segmentation framework. The upper part is the base network and the lower part is the PVCL module, which consists of VCL and PCL. VCL constructs positive/negative pairs among anchor and samples from the same/different classes, while PCL constructs positive/negative pairs among anchor and prototypes from the same/different classes.

from unsupervised paradigm, supervised contrastive loss [34] was proposed to leverage label information.

III. METHODS

In this section, we first describe our base framework, and then introduce our proposed Prototype-Voxel Contrastive Learning approach. The PVCL module consists of two levels, voxel-level contrastive learning (VCL) and prototype-level contrastive learning (PCL). Additionally, a novel sampling pair construction strategy is introduced, where the sampling methods of anchors and positive/negative pairs are designed for voxel-wise features.

A. Framework Overview

Similar to those 3D panoptic segmentation network frameworks [22], [23], [24], [25], [26], our base framework is composed of three parts: feature extraction, semantic segmentation, and instance segmentation. For the feature extraction network, we refer to Cylinder3D [13], which distributes points more evenly than the normal cubic voxel partition and achieves higher performance. We assume that the raw point cloud $P \in \mathbb{R}^{N\times 3} = (x_i,\ y_i,\ z_i), i \in (0,\ N-1)$ is voxelized as $V \in \mathbb{R}^{H\times W\times L\times I}$, where $H \times W \times L$ represents the size of cylindrical partition, I represents the number of input features. After the feature extraction network, V is encoded into $F \in \mathbb{R}^{H\times W\times L\times D}$, where D represents the number of features per voxel. The semantic branch $\mathrm{b}_{\mathrm{sem}}$ maps F into categorical probability $S \in \mathbb{R}^{H\times W\times L\times C}$, where C represents the number of classes. The loss function L_{sem} of $\mathrm{b}_{\mathrm{sem}}$ is cross-entropy between S and its ground truth.

The instance branch $\mathrm{b}_{\mathrm{inst}}$ reduces voxel-wise to point-wise feature representation, and then maps them into the point-wise center offset $o_i = (\Delta x_i,\ \Delta y_i,\ \Delta z_i), i \in (0,\ N_{things})$, where N_{things} represents points belonging to things class according to the valid mask obtained by S. A L_1 regression loss function L_{inst} from [29] is utilized for $\mathrm{b}_{\mathrm{sem}}$, such loss is invariant to the offset vector norm and ensures that the points move towards their instance centroids.

$$L_{o\_reg} = \frac{1}{\sum_i m_i} \sum_i \|o_i - (\hat{c}_i - p_i)\| \cdot m_i \tag{1}$$

$$L_{o\_dir} = \frac{1}{\sum_i m_i} \sum_i \frac{o_i}{\|o_i\|_2} \cdot \frac{\hat{c}_i - p_i}{\|(\hat{c}_i - p_i)\|_2} \cdot m_i \tag{2}$$

$$L_{inst} = L_{o\_reg} + L_{o\_dir} \tag{3}$$

where m_i represents the valid mask where point belongs to things or stuff, m_i = 1 if point i belongs to things and m_i = 0 otherwise. $\widehat{c}_i$ is the centroid of object that point i belongs to. During inference, offset-shifted coordinates $c_i = (\Delta x_i + x_i,\ \Delta y_i + y_i,\ \Delta z_i + z_i), i \in (0,\ N_{things})$ are clustered to output instance labels for things points. For the semantic and instance labels obtained by the two branches respectively, we merge them together according to [24].

B. Prototype-Voxel Contrastive Learning

We introduce the two components of our novel Prototype-Voxel Contrastive Learning (PVCL) module, voxel-level contrastive learning (VCL) and prototype-level contrastive learning (PCL). In VCL, we also propose a novel sampling pair construction strategy for outdoor LiDAR voxel-wise feature representation. We project the point-wise feature vector F into the feature space $F_s \in \mathbb{R}^{H\times W\times L\times D_s}$ through MLP (e.g., 1×1 sparse convolution), where D_s represents the number of features in feature embedding space. In the feature embedding space, our proposed PVCL module regularizes the distances of features in voxel-to-voxel or voxel-to-prototype in a fully supervised setting and explores intra-class and inter-class feature structures.

1) Voxel-level Contrastive Learning (VCL): For each anchor, sampled from all voxels F_s, we search from the other voxels as its positive/negative pairs from the entire scene. The searched voxels that have the same semantic label with the anchor are used to build positive pairs, and the voxels with different semantic labels are used to build negative pairs. In some self-supervised contrastive learning methods, all samples in a batch are generally used as anchors. For each anchor, its positive pair is copied from the data augmentation, and the rest in the batch are its negative pairs. In supervised contrastive learning, samples with the same label as the anchor are also included as positive pairs. This sampling pair construction strategy requires a large batch size to meet the strict requirements of negative pairs (many and hard) in contrastive learning. To this end, memory banks [42], [43], which contribute to expanding the positive/negative pairs sampling pool, and harder sampling algorithms [43], [35] have emerged. However, in the outdoor LiDAR point cloud scene, the number of points/voxels in each scene is large and multi-category, which makes memory bank subordinate. Consequently, we propose a novel sampling pair construction strategy to sample the anchors and positive/negative pairs, which is explained in the next paragraph.

Compared with the normal random sampling method, we consider the integrity of the point cloud scene. The Farthest Point Sampling (FPS) [44] with wider coverage and less randomness is utilized to sample anchors **A** for the stability of CL. Due to the high granularity of the LiDAR point cloud, the number of points/voxels with the same semantic label is also sufficient, which eliminates the request to data augment a copy for the huge point cloud scene. We randomly sample as many positive pairs from the scene as possible. The sampling of negative pairs is considered to be an important role in CL. Due to the characteristics of LiDAR scanning, the object exists overlap or occlusion with other objects. In other words, for each anchor, the heterogeneous points around it should be more worthy of attention. Therefore, we search the neighbors nearest (NN) to the anchor as its negative pairs. For all anchors, the total number of samples for positive and negative pairs is fixed. We will demonstrate the effectiveness of our sampling pair construction strategy in the ablation study. A popular loss function for supervised CL takes the following form [34]:

$$L_{VCL} = \frac{1}{|A|} \sum_{i \in A} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(f_i \cdot f_p / \tau_v)}{\sum_{a \in M(i)} \exp(f_i \cdot f_a / \tau_v)} \tag{4}$$

where $P(i)$ is positive pairs of A_i, $A_i \in$A,$|P(i)|$ is its cardinality, $M(i)$ is sum of positive and negative pairs of A_i, and τ_v is a hyper-parameter for VCL.

2) Prototype-level Contrastive Learning (PCL): The prototype $P = p_i, i \in C$ is defined as the feature center of each semantic class in feature embedding space F_s. The sampled anchor is encouraged to be more similar to its own prototype p'. In the feature embedding space, the features of different semantic classes should be distributed in groups, and PCL strengthens and regularizes this inter-class distribution. Compared with VCL, the positive pair of PCL belongs to its own prototype and its negative pairs are the prototypes of other classes, hence PCL avoids sampling positive/negative pairs as before. In other words, PCL avoids the influence of outliers and noisy samples of each semantic class, which stabilizes the sampling process in supervised CL. The PCL loss we proposed is as follows:

$$L_{PCL} = \frac{1}{|A|} \sum_{i \in A} \log \frac{\exp(f_i \cdot p'_i / \tau_p)}{\sum_{a \in C} \exp(f_i \cdot p_a / \tau_p)} \tag{5}$$

where τ_p is a hyper-parameter for PCL.

We derive the gradient descent formula to show PCL is more stable than VCL when the gradient drops. According to [34]:

$$\frac{\partial L_{VCL}}{\partial f_i} = \frac{1}{\tau_v} \left\{ \sum_{p \in P(i)} f_p \left(F_{ip} - \frac{1}{|P(i)|} \right) + \sum_{n \in N(i)} f_n F_{in} \right\} \tag{6}$$

where F_{ip} is:

$$F_{ip} \equiv \frac{\exp(f_i \cdot f_p / \tau_v)}{\sum_{a \in A(i)} \exp(f_i \cdot f_a / \tau_v)} \tag{7}$$

After the same derivation:

$$\frac{\partial L_{PCL}}{\partial f_i} = \frac{1}{\tau_p} \left\{ p' \left(P_{ip'} - 1 \right) + \sum_{n \in N(i)} p_n P_{in} \right\} \tag{8}$$

where $P_{ip'}$ is:

$$P_{ip'} \equiv \frac{\exp(f_i \cdot p' / \tau_p)}{\sum_{a \in C} \exp(f_i \cdot p_a / \tau_p)} \tag{9}$$

As shown in Eq. 6 and 8, $p' \left(P_{ip'} - 1 \right)$ tends to be a more stable value than $\sum_{p \in P(i)} f_p \left(F_{ip} - \frac{1}{|P(i)|} \right)$, which clarifies why PVCL is more stable than VCL.

3) PVCL Joint Loss Function: After adding the PVCL module, the entire panoptic segmentation framework of the fully supervised setting is composed of four parts of loss, and the overall loss function is

$$Loss = L_{sem} + L_{inst} + \lambda_v \cdot L_{VCL} + \lambda_p \cdot L_{PCL} \tag{10}$$

where λ_v represents the weight for L_{VCL}, and λ_p represents the weight for L_{PCL}.

IV. EXPERIMENTS

To evaluate the effectiveness of our proposed methods, we conduct extensive experiments on the large-scale outdoor LiDAR point cloud data sets, Semantic-KITTI and nuScenes.

A. Dataset and Evaluation Metrics

The Semantic-KITTI contains 19130 scans for the training set, 4071 scans for the validation set, 20351 scans for the test set, and 19 classes including 8 things classes and 11 stuff classes. The nuScenes contains 28130 scans for training set, 6019 scans for validation set, and 16 classes including 10 things classes and 6 stuff classes. We report evaluation experiment results on test set of Semantic-KITTI and report

Method	PQ	PQ†	RQ	SQ	PQ^{th}	RQ^{th}	SQ^{th}	PQ^{st}	RQ^{st}	SQ^{st}	mIoU
RangeNet*++ [21]	37.1	45.9	47.0	75.9	20.2	25.2	75.2	49.3	62.8	76.5	52.4
KPConv* [21]	44.5	52.5	54.4	80.0	32.7	38.7	81.5	53.1	65.9	79.0	58.8
LPASD [25]	38.0	47.0	48.2	76.5	25.6	31.8	76.8	47.1	60.1	76.2	50.9
MOPT [26]	43.1	50.7	53.9	78.8	28.6	35.5	80.4	53.6	67.3	77.7	52.6
Panoster [23]	52.7	59.9	64.1	80.7	49.4	58.5	83.3	55.1	68.2	78.8	59.9
DS-Net [24]	55.9	62.5	66.7	82.3	55.1	62.8	87.2	56.5	69.5	78.7	61.6
EfficientLPS [22]	57.4	63.2	68.7	83.0	53.1	60.5	87.8	**60.5**	**74.6**	**79.5**	61.4
PCL (Ours)	58.0	64.3	68.5	83.7	57.9	64.6	**89.5**	58.1	71.4	79.5	62.5
VCL (Ours)	58.2	65.0	68.9	83.6	58.4	65.4	88.9	58.1	71.4	79.7	63.6
PVCL (Ours)	**59.1**	**65.7**	**69.6**	**84.0**	**59.8**	**66.7**	89.2	58.6	71.6	80.3	**64.0**

TABLE I. LiDAR-based panoptic segmentation results on the test set of Semantic-KITTI. All results in [%].

Method	PQ	PQ†	RQ	SQ	PQ^{th}	RQ^{th}	SQ^{th}	PQ^{st}	RQ^{st}	SQ^{st}	mIoU
DS-Net [24]	42.5	51.0	50.3	**83.6**	32.5	38.3	**83.1**	59.2	70.3	**84.4**	70.7
PCL (Ours)	63.9	67.2	77.2	81.2	59.0	72.3	79.6	68.7	81.3	79.8	71.7
VCL (Ours)	64.1	67.3	77.0	81.6	59.3	72.0	80.1	68.8	**81.3**	79.9	70.5
PVCL (Ours)	**64.9**	**67.8**	**77.9**	81.6	**59.2**	**72.5**	79.7	**67.6**	79.1	77.3	73.9

TABLE II. LiDAR-based panoptic segmentation results on the validation set of nuScenes. All results in [%].

Method	PQ	car	truck	bicycle	motorcycle	other-vehicle	person	bicyclist	motorcyclist	road	sidewalk	parking	other-ground	building	vegetation	trunk	terrain	fence	pole	traffic-sign
RangeNet*++ [21]	37.1	66.9	6.7	3.1	16.2	8.8	14.6	31.8	13.5	90.6	63.2	41.3	6.7	79.2	71.2	34.6	37.4	38.2	32.8	47.4
KPConv* [21]	44.5	72.5	17.2	9.2	30.8	19.6	29.9	59.4	22.8	84.6	60.1	34.1	8.8	80.7	77.6	53.9	42.2	49.0	46.2	46.8
Panoster [23]	52.7	84.0	18.5	36.4	44.7	30.1	61.1	69.2	51.1	90.2	62.5	34.5	6.1	82.0	77.7	55.7	41.2	48.0	48.9	59.8
DS-Net [24]	55.9	91.2	28.8	45.4	47.2	34.6	63.6	71.1	58.5	89.1	61.2	32.3	4.0	83.2	79.6	58.3	43.4	50.0	**55.2**	65.3
EfficientLPS [22]	57.4	85.7	30.3	37.2	47.7	43.2	**70.1**	66.0	44.7	**91.1**	**71.1**	**55.3**	**16.3**	**87.9**	80.6	52.4	**47.1**	53.0	48.8	61.6
PCL (Ours)	58.0	94.1	36.5	47.7	**49.5**	44.0	63.6	68.9	59.0	89.9	63.3	39.7	7.2	86.0	80.1	57.9	43.9	52.5	53.5	64.5
VCL (Ours)	58.2	93.8	**42.5**	45.5	45.4	**47.9**	62.9	72.4	56.9	89.5	64.4	34.7	9.0	84.8	80.4	**58.8**	43.4	53.3	54.9	66.1
PVCL (Ours)	**59.1**	**94.1**	39.2	**47.2**	47.1	43.9	62.9	**76.5**	**67.8**	88.8	64.3	38.4	10.5	85.5	**80.7**	58.6	42.3	**53.4**	55.1	**66.8**

TABLE III. Detailed per-class PQ results on the test set of Semantic-KITTI. All results in [%].

ablation studies on validation set of both Semantic-KITTI and nuScenes. *Panoptic Quality (PQ)*, *Segmentation Quality (SQ)*, *Recognition Quality (RQ)* and *Mean Intersection over Union (mIoU)* [21] are utilized to evaluate performance of panoptic segmentation. The above three metrics are also calculated separately on things and stuff classes which give PQ^{th}, SQ^{th}, RQ^{th}, PQ^{st}, SQ^{st}, RQ^{st}. PQ† uses IoU for a stuff class and PQ_c for a things class.

B. Implementation Details

Each scene is voxelized to $480 \times 360 \times 32$ voxels under cylindrical partition. The number of features in F is 128 and in F_s is 128. The projection layer is implemented as two 1×1 3D sparse convolutional layers with Batchnormalization and ReLU layers. This projection layer is applied during training and removed at inference time. In the loss function, the τ_v and τ_p are set as 0.07, the λ_v is set as 0.5, and λ_p is set as 5. We use a mini-batch size of 16 on 4 NVIDIA V-100 32GB GPUs, and use Adam as our optimizer, with an initial learning rate of 0.002. Note that we do not use any extra training data (e.g., EfficientLPS [22]). In PVCL, we use our pair construction strategy to sample 512 and 128/1024 points for the anchor and positive/negative pairs respectively. During testing, we average the semantic results with flipping without change or extra inference step.

C. Results and Analysis

Semantic-KITTI. We compare with several strong baseline results in order to validate the effectiveness of our PVCL framework. These baselines are introduced in related work, where RangeNet*++ and KPConv* are two sets of baselines proposed by [21], namely RangeNet++ [38] + PointPillars [6] and KPConv [14] + PointPillars[6], and LPASD represents Panoptic RangeNet proposed by [25]. PCL/VCL represents our method only with PCL/VCL module and PVCL presents utilizing both. Table I shows that our method wins first place in 7 out of 11 metrics and surpasses the best-published baseline method, EfficientLPS [22], by 1.7% in PQ and 2.6% in mIoU. Especially in thing classes, our method shows a greater improvement, 4.7% increase in PQ^{th} relative to DS-Net.

nuScenes. Since nuScenes recently released the official panoptic challenge, the evaluation method, combining semantic segmentation and 3D bounding boxes for panoptic segmentation, used by DS-Net is eliminated. As shown in Table II, our proposed PVCL is validated on the official panoptic challenge.

ID	Anchors	Positive	Negative	PQ	RQ	SQ	mIoU
1	Random	Random	Random	56.3	66.4	74.0	61.6
2	FPS	Random	Random	57.1	66.9	74.9	61.7
3	Random	Random	NN	57.5	67.9	77.8	63.3
4	FPS	NN	Random	58.7	68.7	74.8	63.6
Ours	FPS	Random	NN	**59.3**	**69.4**	**78.5**	**65.8**

TABLE IV. Detailed results of different pair construction strategies on the validation set of Semantic-KITTI, where NN represents sampling nearest points of anchors. All results in [%].

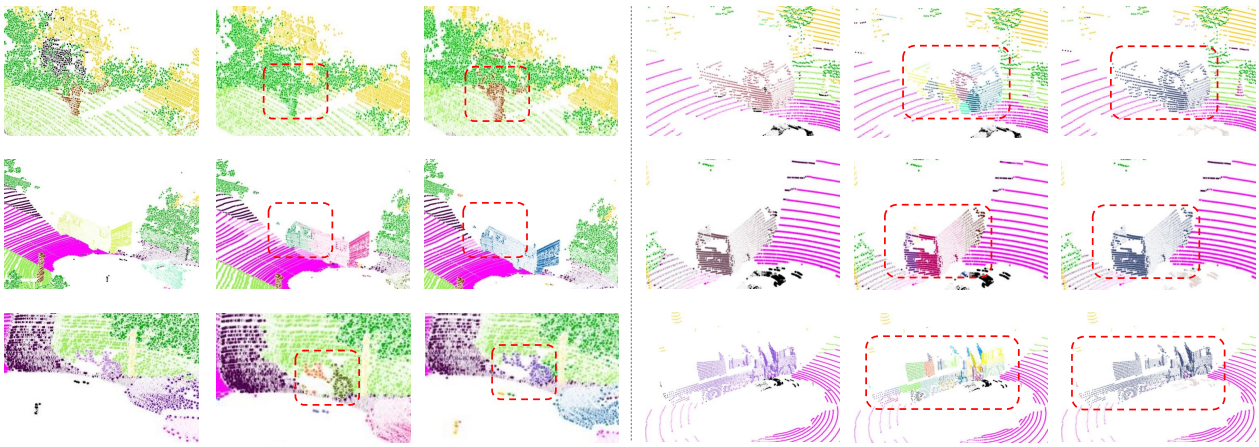

Fig. 3. Visual comparisons among Ground Truth, DS-Net, and PVCL (Ours) (from left to right) on Semantic-KITTI (left) and nuScenes (right).

Method	Inference time	PQ	RQ	SQ	mIoU
RangeNet$^*_{++}$	409ms	37.1	47.0	75.9	52.4
KPConv*	514ms	44.5	54.4	80.0	58.8
DS-Net	721ms	55.9	66.7	82.3	61.6
Ours	**405ms**	**59.1**	**69.6**	**84.0**	**64.0**

TABLE V. Single scan inference time of different methods, measured with NVIDIA V-100.

D. Ablation Study

Here we provide ablation studies on different modules and strategies to verify the efficacy of the proposed methods.

PCL/VCL and PVCL. As shown in Table I and III, our method with only PCL/VCL outperforms DS-Net by 2.1%/2.3% in PQ, EfficientLPS by 4.8%/5.3% in PQ$^{\text{th}}$. Compared with VCL, PCL lacks detailed voxel-to-voxel contrast, thence that PCL declines 0.2% on PQ. Combining with both, PVCL significantly outperforms PCL and VCL, and it reaches the highest for 4 out of 8 things classes.

Pair construction strategy. In order to validate our effective pair construction strategy, we randomly sample anchors and negative pairs respectively. We reorganize the three sampling methods of FPS, NN, and Random to verify the superiority of our FPS and NN sampling over Random sampling. Four comparative experiments are set to train 36k iterations on the training set, and the results are evaluated on the validation set, as shown in Table IV. On the Semantic-KITTI validation set, our pair construction strategy shows superiority over other random sampling strategies.

Computational efficiency. Nevertheless, the main motivation for our PVCL module is the maintained computational speed, because PVCL is abandoned during inference. In comparison to DS-Net and two-stage approaches, such as KPConv*, PVCL achieves both better performance and faster speed. As shown in Table V, compared with DS-Net, the inference time of PVCL is reduced by 43.8%.

E. Results Visualization

We visualize our results on the validation set of Semantic-KITTI (left) and nuScenes (right). Figure. 3 depicts qualitative comparisons of our proposed PVCL against DS-Net, which is the only work released code. We cannot compare the visualization results of EfficientLPS, because EfficientLPS has no public released code and detailed test set results. The focus of EfficientLPS is the feature maps fusion of semantic branches, while our focus is on the overall training framework. The two are not contradictory and can be combined. The visualization shows that our method improves point cloud perception. At the top of the left, DS-Net mixes vegetable and trunk together, and our method distinguishes well. Comparison of the two groups in the middle and at the bottom of the left show the advantages of our method in instance segmentation on Semantic-KITTI. As shown in Table III, we demonstrate that our method achieves the best performance in things class, and 7 out of 8 things classes reach the first place in PQ. The truck category far exceeds the second place by 12.2% in PQ ,the bicyclist category exceeds by 5.5% in PQ and the motorcyclist category exceeds by 9.4% in PQ. For the truck shown in the middle of the left, DS-Net erroneously divides it into two objects, and our method can segment it correctly. The bicycle is shown in the bottom of the left has segmentation errors in DS-Net, but PVCL is able to segment correctly. As for nuScenes, our proposed approach PVCL achieves satisfactory performance on large objects, like truck, bus and construction vehicle.

V. CONCLUSIONS

In this paper, we tackle the problem of LiDAR point cloud panoptic segmentation. We propose an end-to-end differentiable framework based on Prototype-Voxel CL, consisting of both VCL and PCL. To the best of our knowledge, it is the first supervised CL framework for point cloud panoptic segmentation. Extensive experiments show our method achieves SOTA performance for point cloud panoptic segmentation. For the future work, the recent progress of point cloud registration method [45] has shown superior performance, especially when the overlap between two point clouds is small, thus inspiring us to use multiple point cloud data for panoptic segmentation in the future.

REFERENCES

[1] Q. He, Z. Wang, H. Zeng, Y. Zeng, S. Liu, and B. Zeng, "SVGA-Net: Sparse Voxel-Graph Attention Network for 3D Object Detection from Point Clouds," *arXiv*, 2020.

[2] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 10 526–10 535, 2020.

[3] Y. Zhang, D. Huang, and Y. Wang, "PC-RGNN: Point cloud completion and graph neural network for 3D object detection," *arXiv*, 2020.

[4] D. Zhou, J. Fang, X. Song, L. Liu, J. Yin, Y. Dai, H. Li, and R. Yang, "Joint 3D instance segmentation and object detection for autonomous driving," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1836–1846, 2020.

[5] Z. Wang and K. Jia, "Frustum ConvNet: Sliding Frustums to Aggregate Local Point-Wise Features for Amodal," *IEEE International Conference on Intelligent Robots and Systems*, pp. 1742–1749, 2019.

[6] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 12 689–12 697, 2019.

[7] Q. Chen, L. Sun, Z. Wang, K. Jia, and A. Yuille, "Object as Hotspots: An Anchor-Free 3D Object Detection Approach via Firing of Hotspots," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12366 LNCS, pp. 68–84, 2020.

[8] Z. Luo, Z. Cai, C. Zhou, G. Zhang, H. Zhao, S. Yi, S. Lu, H. Li, S. Zhang, and Z. Liu, "Unsupervised domain adaptive 3d detection with multi-level consistency," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8866–8875.

[9] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "IPOD: Intensive point-based object detector for point cloud," *arXiv*, 2018.

[10] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3D Proposal Generation and Object Detection from View Aggregation," *IEEE International Conference on Intelligent Robots and Systems*, pp. 5750–5757, 2018.

[11] M. Gerdzhev, R. Razani, E. Taghavi, and B. Liu, "Tornado-Net: Multiview total variation semantic segmentation with diamond inception module," *arXiv*, pp. 1–10, 2020.

[12] H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, and S. Han, "Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution," 2020. [Online]. Available: http://arxiv.org/abs/2007.16100

[13] 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," 2020. [Online]. Available: http://arxiv.org/abs/2011.10033

[14] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "KPConv: Flexible and Deformable Convolution for Point Clouds," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-Octob, pp. 6410–6419, apr 2019. [Online]. Available: http://arxiv.org/abs/1904.08889

[15] 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," mar 2020. [Online]. Available: http://arxiv.org/abs/2003.14032

[16] C. Xu, B. Wu, Z. Wang, W. Zhan, P. Vajda, K. Keutzer, and M. Tomizuka, "SqueezeSegV3: Spatially-Adaptive Convolution for Efficient Point-Cloud Segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12373 LNCS, pp. 1–19, 2020.

[17] Y. Liu, Q. Fan, S. Zhang, H. Dong, T. Funkhouser, and L. Yi, "Contrastive multimodal fusion with tupleinfonce," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 754–763.

[18] 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.

[19] S. J. Li, X. Chen, Y. Liu, D. Dai, C. Stachniss, and J. Gall, "Multi-scale interaction for real-time lidar data segmentation on an embedded platform," *arXiv*, vol. 14, no. 8, pp. 1–11, 2020.

[20] 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," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2019-May, pp. 4376–4382, 2019.

[21] J. Behley, A. Milioto, C. Stachniss, J. Behley, A. Milioto, and C. Stachniss, "A Benchmark for LiDAR-based Panoptic Segmentation based on KITTI," mar 2020. [Online]. Available: http://arxiv.org/abs/2003.02371

[22] K. Sirohi, R. Mohan, D. Büscher, W. Burgard, and A. Valada, "EfficientLPS: Efficient LiDAR Panoptic Segmentation," pp. 1–20, 2021. [Online]. Available: http://arxiv.org/abs/2102.08009

[23] S. Gasperini, M.-A. N. Mahani, A. Marcos-Ramiro, N. Navab, and F. Tombari, "Panoster: End-to-end Panoptic Segmentation of LiDAR Point Clouds," 2020. [Online]. Available: http://arxiv.org/abs/2010.15157

[24] F. Hong, H. Zhou, X. Zhu, H. Li, and Z. Liu, "LiDAR-based Panoptic Segmentation via Dynamic Shifting Network," pp. 1–17, 2020. [Online]. Available: http://arxiv.org/abs/2011.11964

[25] A. Milioto, J. Behley, C. McCool, and C. Stachniss, "LiDAR Panoptic Segmentation for Autonomous Driving," *Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, no. ii, 2020.

[26] J. V. Hurtado, R. Mohan, W. Burgard, and A. Valada, "MOPT: Multi-Object Panoptic Tracking," 2020. [Online]. Available: http://arxiv.org/abs/2004.08189

[27] Z. Liu, H. Tang, Y. Lin, and S. Han, "Point-voxel CNN for efficient 3d deep learning," *arXiv*, no. NeurIPS, 2019.

[28] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 77–85, 2017.

[29] L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia, "PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation," apr 2020. [Online]. Available: http://arxiv.org/abs/2004.01658

[30] L. Zhao and W. Tao, "JSNet: Joint Instance and Semantic Segmentation of 3D Point Clouds," dec 2019. [Online]. Available: http://arxiv.org/abs/1912.09654

[31] C.-Y. Chuang, J. Robinson, L. Yen-Chen, A. Torralba, and S. Jegelka, "Debiased contrastive learning," *arXiv preprint arXiv:2007.00224*, 2020.

[32] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised Feature Learning via Non-parametric Instance Discrimination," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

[33] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9726–9735, 2020.

[34] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised Contrastive Learning," no. NeurIPS, pp. 1–23, 2020. [Online]. Available: http://arxiv.org/abs/2004.11362

[35] S. Xie, J. Gu, D. Guo, C. R. Qi, L. J. Guibas, and O. Litany, "PointContrast: Unsupervised Pre-training for 3D Point Cloud Understanding," 2020. [Online]. Available: http://arxiv.org/abs/2007.10985

[36] 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," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-Octob, pp. 9296–9306, apr 2019. [Online]. Available: http://arxiv.org/abs/1904.01416

[37] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," 2020.

[38] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss, "RangeNet ++: Fast and Accurate LiDAR Semantic Segmentation," in *IEEE International Conference on Intelligent Robots and Systems*. Institute of Electrical and Electronics Engineers Inc., nov 2019, pp. 4213–4220.

[39] T. Li, X. Chen, S. Zhang, Z. Dong, and K. Keutzer, "Cross-domain sentiment classification with contrastive learning and mutual information maximization," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8203–8207.

[40] X. Yue, Z. Zheng, S. Zhang, Y. Gao, T. Darrell, K. Keutzer, and A. S. Vincentelli, "Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 834–13 844.

[41] J. Li, P. Zhou, C. Xiong, R. Socher, and S. C. Hoi, "Prototypical contrastive learning of unsupervised representations," *arXiv preprint arXiv:2005.04966*, 2020.

[42] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised Feature Learning via Non-parametric Instance Discrimination," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

[43] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring Cross-Image Pixel Contrast for Semantic Segmentation," 2021. [Online]. Available: http://arxiv.org/abs/2101.11939

[44] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 5100–5109, 2017.

[45] Y. Wang, C. Yan, Y. Feng, S. Du, Q. Dai, and Y. Gao, "Storm: Structure-based overlap matching for partial point cloud registration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.