

Clustering based Point Cloud Representation Learning for 3D Analysis

Tuo Feng¹, Wenguan Wang², Xiaohan Wang², Yi Yang^{2,*}, Qinghua Zheng³

¹ ReLER, AAIL, University of Technology Sydney ² ReLER, CCAI, Zhejiang University ³ Xi'an Jiaotong University

<https://github.com/FengZicai/Cluster3Dseg/>

Abstract

Point cloud analysis (such as 3D segmentation and detection) is a challenging task, because of not only the irregular geometries of many millions of unordered points, but also the great variations caused by depth, viewpoint, occlusion, etc. Current studies put much focus on the adaption of neural networks to the complex geometries of point clouds, but are blind to a fundamental question: how to learn an appropriate point embedding space that is aware of both discriminative semantics and challenging variations? As a response, we propose a clustering based supervised learning scheme for point cloud analysis. Unlike current de-facto, scene-wise training paradigm, our algorithm conducts within-class clustering on the point embedding space for automatically discovering subclass patterns which are latent yet representative across scenes. The mined patterns are, in turn, used to repaint the embedding space, so as to respect the underlying distribution of the entire training dataset and improve the robustness to the variations. Our algorithm is principled and readily pluggable to modern point cloud segmentation networks during training, without extra overhead during testing. With various 3D network architectures (i.e., voxel-based, point-based, Transformer-based, automatically searched), our algorithm shows notable improvements on famous point cloud segmentation datasets (i.e., 2.0-2.6% on single-scan and 2.0-2.2% multi-scan of SemanticKITTI, 1.8-1.9% on S3DIS, in terms of mIoU). Our algorithm also demonstrates utility in 3D detection, showing 2.0-3.4% mAP gains on KITTI.

1. Introduction

During the last few years, point cloud analysis, such as 3D segmentation, has attracted increasing research effort, due to the wide applications in autonomous driving, intelligent robotics, airborne laser scanning, and virtual reality. In particular, the advances in deep learning significantly pushed forward the state-of-the-art in this field. Applying standard neural networks which are specialized for grid-like

data, such as natural images, to point clouds is nontrivial, as point data are unorganized and irregular. To adapt neural networks to the geometries of point data, considerable effort has been made and representative achievements include: i) *projection-/voxel-based networks* [1–11] that project irregular point clouds to regular representations, so that mature 2D/3D convolution can be applied for segmentation; and ii) *point-based networks* [12–16] that ingest raw point clouds directly, by using permutation-invariant operator [17–21], graph convolution [22], customized convolution [23–25], or self-attention (Transformer) based architecture [26–28].

Nevertheless, the challenges in point cloud analysis stem not only from the intrinsic non-Euclidean nature of point data, but also from the large intra-class variations caused by depth, occlusion, viewpoint, shape, etc. Despite various fancy point structure-aware network designs and their encouraging results, a fundamental issue was long ignored: *how to learn a good point embedding space that is discriminative for semantic categorization yet robust for point data variations?*

Mitigating this issue demands a powerful learning regime that is aware of latent variation modes (or representative fine-grained patterns) – comprehensively describing the potential structure of point data. However, in practice, it is infeasible to precisely annotate, or even roughly identify, the underlying data patterns in point clouds. This may be the reason behind the common choice that point cloud segmentation is learned as point-wise classification; any fine-grained patterns that the point data may possess are left to be ‘mysteriously’ learned through the supervision from high-level semantic tags.

These novel insights motivate us to devise a clustering analysis based training scheme for point cloud segmentation. It complements the standard supervised learning of point-wise classification with unsupervised clustering and regularization of the feature space. Specifically, clustering is conducted inside each labeled semantic class to automatically discover informative yet hidden subclass patterns without explicit annotation. The discovered subclass patterns essentially capture the underlying fine-grained distribution of the whole training dataset. They are then used to reshape the point embedding space, achieved by explicitly inspiring inter-subclass-/cluster discriminativeness, and reducing intra-

¹Corresponding author: Yi Yang.

subclass/-cluster variation. Such regularized representation space in turn facilitates the discovery of typical within-class variation modes, and benefits point recognition eventually.

Our learning algorithm enjoys several appealing advantages: **First**, it raises a *dataset-level context-aware* training strategy. Unlike the current *de-facto*, scene-wise training paradigm, our algorithm groups point features across training scenes, and conducts clustering based representation learning. By probing the global data distribution, our algorithm encourages the highly flexible feature space to be discretized into a few distinct subcluster centers, easing the difficulty of the final semantic classification. **Second**, it is *efficient* for large-scale point cloud training. To avoid time-consuming clustering of massive point data, we opt the Sinkhorn-Knopp algorithm [29, 30] that solves cluster assignment using fast matrix-vector algebra [31]. Moreover, to follow closely the drifting representation during network training, a momentum update strategy is adopted for online approximation of the subcluster centers. **Third**, it is *principled* enough to be seamlessly incorporated into the training process of any modern point cloud segmentation networks, without bringing extra computation burden or model parameters during inference.

For thorough evaluation, we approach our training algorithm on four remarkable point cloud segmentation models, *i.e.*, Cylinder3D [16] (*voxel*-based), KPConv [25] (*point*-based), PTV1 [26] (*Transformer*-based), SPVNAS [32] (*neural architecture search* (NAS) based), and conduct experiments on 3D point cloud segmentation for urban scenes (*i.e.*, SemanticKITTI [33] single-scan) and indoor environments (*i.e.*, S3DIS [34]) as well as 4D segmentation of point cloud sequences (*i.e.*, SemanticKITTI [33] multi-scan). Results show that our algorithm owns **2.2-2.6%**, **1.9-2.2%**, **1.8%**, and **2.0%** mIoU gains over Cylinder3D, KPConv, PTV1, and SPVNAS, respectively. Our algorithm even promotes 3D detectors Second [35] and PointPillar [36] by **2.7-3.4%** and **2.0-2.2%** mAP on KITTI [37], verifying its high generality.

2. Related Work

Deep Learning for Static Point Cloud Segmentation. In general, existing algorithms for single-scan point cloud segmentation can be categorized into two schools, depending on the underlying data representation: **i)** *Projection-based* methods first transform unstructured point sets to regular 2D grid [2, 5, 7, 8, 10, 38–40], or 3D voxel [3, 4, 9, 16, 41–46], to enable the usage of vanilla 2D/3D convolution operation. However, 2D projection based methods are likely to discard critical geometric cues and require expensive 2D-3D back-projection after 2D segmentation, yet voxel-based architectures typically suffer from significant computation and memory usage. **ii)** *Point-based* methods, pioneered by PointNet [17, 18], directly learn point-wise features from raw point clouds, usually through 1) local feature pooling [14, 15, 21, 47–54], 2) graph convolution [22, 51, 55–

62], 3) kernel-based convolution [13, 19, 25, 63–67], and 4) attention-based aggregation [20, 26–28, 68]. Compared with projection-based approaches, point-based methods tend to be computationally efficient and are capable of preserving point-wise semantics as well as local geometries. Unfortunately, their performance in large-scale, urban scenarios is still not desirable [69].

Deep Learning for Dynamic Point Cloud Segmentation. 4D semantic segmentation is rather difficult as point cloud sequences are spatially irregular yet temporally ordered. Existing approaches for dynamic point cloud segmentation can be broadly classified into two groups, in terms of the spatial-temporal information fusion strategy: **i)** *Early fusion based* methods [6, 70] directly process point cloud sequences via adapting the standard convolution to the heterogeneous characteristics of point clouds in spatial and temporal domains. In this way, they allow spatial-temporal information fusion throughout the networks. **ii)** *Late fusion based* methods [28, 71–74] are typically built upon existing single-scan point cloud processing models for spatial information extraction, and devoted to leveraging temporal information to enrich static features and hence to boost segmentation.

Despite their dazzling network designs, existing static/dynamic point cloud segmentation models generally follow a *scene-wise* training protocol, which treats each point data as an individual training sample and accumulates all the point classification errors within each scene for network parameter optimization. As a result, they ignore the rich relations between points across different scenes, and fail to regularize the feature embedding space from a holistic view. In contrast, through automatic, *class-wise* data clustering, our training algorithm grasps the latent structure of the whole training dataset, which draws on a key insight that meaningful, latent data structure, like subclass semantics, fine-grained patterns, and intra-class variation modes, are common and stable across scenes. As we will show, representation learned in such a way is desirable for detailed analysis of point clouds.

Self-supervised Representation Learning and Clustering.

Our algorithm relies on automated discovery of unknown subclasses, achieved by clustering point data with only coarse-grained class labels. Thus it bears some resemblance to self-supervised learning techniques which learn meaningful representations from massive unlabeled data. A spectrum of recent unsupervised representation learning methods [75–79] build upon the *instance discrimination* task that considers each data instance of the dataset as its own class [80, 81]. They conduct noise contrastive estimation [82], a special form of contrastive learning [83, 84], to compare instances, and also show promise in dense 2D/3D representation learning [58, 85–93]. Another line of methods [31, 81, 94–103] discriminates between groups of images with similar features instead of individual images, by jointly performing unsupervised representation learning and clustering.

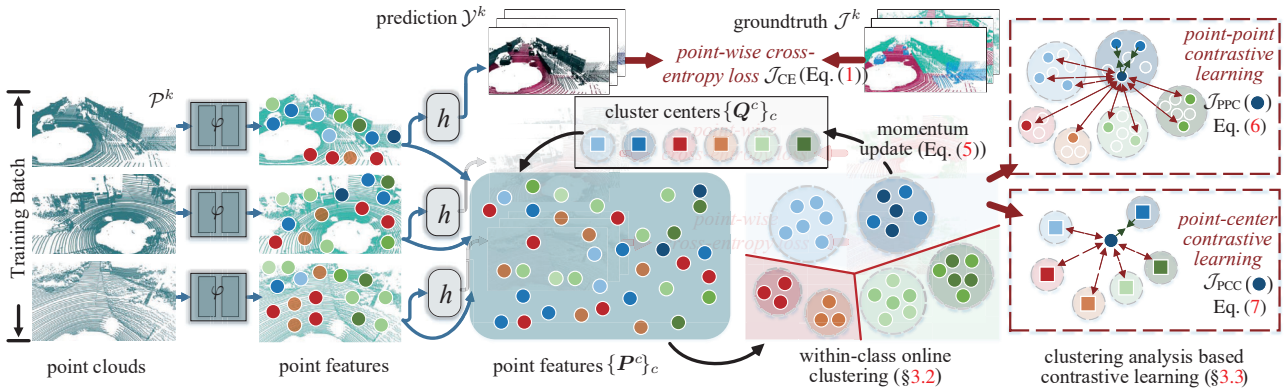


Figure 1: Overview of our clustering based supervised learning algorithm for point cloud segmentation.

In this work, we resort to clustering to probe the underlying structure of large-scale point sets and discover fine-grained patterns within manually-labeled, high-level semantic classes. We reinforce the standard supervised training paradigm of point recognition with clustering analysis based point representation learning, which regularizes the feature space by respecting the inherent structure of point data. This represents the first effort, as far as we know, that explores automatic, fine-grained pattern mining in the context of fully supervised learning of point cloud segmentation.

3. Proposed Algorithm

3.1. Problem Statement and Algorithm Overview

In the context of *fully supervised* learning of point cloud segmentation, current common practice is to learn a point recognition network from a training dataset $\{\mathcal{P}^k, \mathcal{L}^k\}_k$. Here $\mathcal{P}^k = \{p_n^k \in \mathbb{R}^{3+x}\}_{n=1}^N$ is the k -th point cloud containing N points with 3D position and other auxiliary information (e.g., color, intensity); $\mathcal{L}^k = \{l_n^k \in \mathcal{C}\}_{n=1}^N$ contains semantic labels for the points in \mathcal{P}^k , where \mathcal{C} is the label list, e.g., $\mathcal{C} = \{car, road, \dots\}$. The segmentation network is achieved as $h \circ \varphi: \mathcal{P} \mapsto \mathcal{L}$, where $\varphi: \mathbb{R}^{N \times (3+x)} \mapsto \mathbb{R}^{N \times d}$ is a *feature extractor* (■ in Fig. 1) that embeds points in \mathcal{P} into a d -dimensional feature space, and $h: \mathbb{R}^{N \times d} \mapsto \mathbb{R}^{N \times |\mathcal{C}|}$ is a *segmentation head* (□) usually consisting of a small MLP, mapping point features into the discriminative semantic space for point-wise, $|\mathcal{C}|$ -way classification. Thus the whole network is typically learned by minimizing the point-wise cross-entropy loss¹:

$$\mathcal{J}_{CE}(p_n) = -\log P(l_n|p_n) = -\log \frac{\exp(y_{n,l_n})}{\sum_{c \in \mathcal{C}} \exp(y_{n,c})}, \quad (1)$$

where $\mathbf{y}_n = [y_{n,c}]_c \in \mathbb{R}^{|\mathcal{C}|}$ is the vector of categorical scores (*logits*) for point p_n , i.e., $\mathbf{y}_n = h(\mathbf{p}_n)$, and $\mathbf{p}_n \in \mathbb{R}^d$ is the feature of p_n obtained from φ . For the feature extractor φ , there already have many candidates (e.g., voxel-/point-based 3D networks) elaborately designed to capture the specific geo-

¹In practice, some other losses (e.g., lovasz loss [104]) can be used as complementary, but this does not affect our conclusion.

metries of point data. However, point clouds yield rich and diverse patterns, e.g., fine-grained semantics, intra-class variations, etc. These patterns reflect underlying data structures; they are informative yet challenging for semantic understanding, and even hard to be identified. Thus it is usually the case that simply learning the segmentation network $h \circ \varphi$ from the supervision of easily-acquired high-level semantic tags (i.e., Eq. 1), without considering the underlying data structures.



We instead devise a *clustering analysis* based supervised learning framework (Fig. 1). Our algorithm not only learns point recognition with pre-given semantic tags, but more essentially, it automatically discovers and encodes latent structures of point data into the feature space φ . Features learned in such strategy are expected to be more discriminative for (fine-grained) semantics and robust for intra-class variations, hence facilitating final dense recognition of point clouds.

At each training iteration, our algorithm has two phases. In **phase 1**, we perform online clustering over massive points inside of each labeled classes. The purpose is to search for subclass patterns which are hard to be labeled yet significant across scenes. In **phase 2**, in addition to optimizing the whole segmentation network $h \circ \varphi$ with the point-wise classification loss \mathcal{L}_{CE} as usual, we leverage deterministic cluster assignments as an auxiliary constraint to shape the feature space φ . The improved features, in turn, enable more reliable within-class clustering, and eventually boost point recognition. Independent of a certain point segmentation network, our training scheme is powerful and general.

3.2. Online Clustering based Subclass Pattern Mining

Our algorithm is built upon an intuitive insight: capturing underlying data structures can facilitate point representation learning and semantic recognition. Thus the first major question arises: *how to automatically and efficiently discover underlying data structures, which cannot be explicitly labeled, from massive training points?* This motivates us to conduct unsupervised clustering inside each labeled class $c \in \mathcal{C}$ so as to automatically mine representative yet latent subclass patterns.

To scale our algorithm to millions of point data, we formulate such within-class clustering as optimal transport, which can be efficiently solved using Sinkhorn Iteration [30]. In addition, to overcome the computational expensive process of cluster center computation, which requires a full epoch over the entire dataset after every update of the representation, we adopt a momentum update strategy for proceeding online clustering simultaneously with network batch training.

For each class $c \in \mathcal{C}$, we assume it contains M latent, fine-grained patterns. Hence there are a total of $M \times |\mathcal{C}|$ unobservable patterns are desired to be discovered from the training dataset $\{\mathcal{P}^k, \mathcal{L}^k\}_k$. To do so, we perform within-class clustering on the point embedding space φ . As a result, the training points belonging to class c , *i.e.*, $\mathcal{P}^c = \{p_n | l_n = c\}$, are partitioned into M subclasses, and the M patterns of class c can be intuitively represented as the corresponding cluster centers. Let $\mathbf{Q}^c = [\mathbf{q}_1^c, \dots, \mathbf{q}_M^c] \in \mathbb{R}^{d \times M}$ denote the M cluster centers of class c (*e.g.*,  in Fig. 1), and $\mathbf{P}^c = [\mathbf{p}_1^c, \dots, \mathbf{p}_{N^c}^c] \in \mathbb{R}^{d \times N^c}$ all the features² of points belonging to class c (*e.g.*, , where $p^c \in \mathcal{P}^c$ and $N^c = |\mathcal{P}^c|$). The cluster assignment can be represented as a binary matrix, $\mathbf{A}^c \in \{0, 1\}^{M \times N^c}$, where the (m, i) -th element of \mathbf{A}^c indicates whether assigning the i -th point of \mathcal{P}^c to the m -th cluster center, *i.e.*, the m -th subclass, of c . The clustering inside class c can be achieved as the optimization of the assignment matrix \mathbf{A}^c , *i.e.*, maximizing the similarity between the point features and cluster centers:

$$\min_{\mathbf{A}^c \in \mathcal{A}^c} \langle \mathbf{A}^{c\top}, -\log \mathbf{S}^c \rangle, \quad (2)$$

$$\mathcal{A}^c = \{\mathbf{A}^c \in \{0, 1\}^{M \times N^c} | \mathbf{A}^{c\top} \mathbf{1}_M = \mathbf{1}_{N^c}, \mathbf{A}^c \mathbf{1}_{N^c} = \frac{N^c}{M} \mathbf{1}_M\}$$

where $\mathbf{S}^c = \text{softmax}(\mathbf{Q}^{c\top} \mathbf{P}^c)$ refers to the similarity matrix between cluster centers and points, $\langle \cdot \rangle$ is the Frobenius dot-product, \log is applied element-wise, and $\mathbf{1}_M$ denotes the vector of ones in dimension M . For the solution space \mathcal{A}^c , the former constraint enforces that each point is assigned to exactly one subclass, and the later imposes an equipartition constraint [31, 81] to inspire the N^c points to be grouped into M subclasses of equal size. The equipartition constraint helps avoid the degenerate solution where all the point samples are partitioned to a single cluster [88, 96]. By relaxing \mathbf{A}^c to be an element of *transportation polytope* [30], *i.e.*, $\mathcal{A}^c = \{\mathbf{A}^c \in \mathbb{R}_+^{M \times N^c} | \mathbf{A}^{c\top} \mathbf{1}_M = \frac{1}{N^c} \mathbf{1}_{N^c}, \mathbf{A}^c \mathbf{1}_{N^c} = \frac{1}{M} \mathbf{1}_M\}$, the label assignment task can be viewed as an instance of the *optimal transport* problem, which can be efficiently solved by a fast version of the Sinkhorn-Knopp algorithm [30]:

$$\min_{\mathbf{A}^c \in \mathcal{A}^c} \langle \mathbf{A}^{c\top}, -\log \mathbf{S}^c \rangle + \frac{1}{\lambda} \text{KL}(\mathbf{A}^c || \frac{1}{MN^c} \mathbf{1}_M \mathbf{1}_{N^c}^\top), \quad (3)$$

where KL is the Kullback-Leibler divergence, and λ is the strength of the regularization. The solution of Prob. (3) over

²Point feature has been projected to the unit sphere: $\mathbf{p} = \mathbf{p} / \|\mathbf{p}\|_2$; \mathbf{p} is reused without causing ambiguity.

the set \mathcal{A}^c can be written as:

$$\mathbf{A}^{c*} = \text{diag}(\mathbf{u})(\mathbf{S}^c)^\lambda \text{diag}(\mathbf{v}), \quad (4)$$

where exponentiation is meant element-wise. $\mathbf{u} \in \mathbb{R}^M$ and $\mathbf{v} \in \mathbb{R}^{N^c}$ are two vectors of scaling coefficients, obtained using a small number of matrix-vector multiplications via iterative Sinkhorn-Knopp algorithm [30]. Due to the drift of the point representation caused by iterative network training, after each training batch of point clouds, re-computing the cluster assignment would cost a pass over the full data. To circumvent such computationally expensive procedure of offline cluster assignment, we restrict the transportation polytope to the minibatch, through approximating the cluster centers \mathbf{Q}^c with momentum. As in [88], at each training iteration, each cluster center \mathbf{q}_m^c of class c is updated as:

$$\mathbf{q}_m^c \leftarrow \mu \mathbf{q}_m^c + (1 - \mu) \bar{\mathbf{p}}_m^c, \quad (5)$$

where $\mu \in [0, 1]$ is the momentum coefficient, and $\bar{\mathbf{p}}_m^c$ indicates mean feature vector of the points that are assigned to cluster center \mathbf{q}_m^c in the current batch. The cluster centers are initialized randomly and gradually updated every batch, following smoothly the changing of the representation φ . These designs lead to scalable and online clustering, allowing to automatically mine latent subclass patterns from massive point data. The clustering is very efficient on GPU; in practice, assigning 50K points into 40 clusters takes only 60 ms. We visualize clustering results ($M=2$) of five classes in Fig. 2, where subclasses under the same class are associated with similar colors. As seen, points with similar patterns are grouped together, thus the underlying data distribution of each class can be comprehensively captured.

3.3. Clustering Analysis based Point Cloud Representation Learning

Through within-class clustering, we search for latent structures in point clouds, and detect locally discriminative patterns, *i.e.*, the cluster centers $\{\mathbf{q}_m^c\}_{m,c}$. The next question is: *how to leverage these fine-grained patterns to aid point cloud representation learning?* To answer this, we complement the supervised point-wise classification loss \mathcal{J}_{CE} (Eq. (1)) with a clustering analysis based contrastive learning strategy, which poses structured and direct supervision for point representation. In particular, with the deterministic cluster assignments in §3.2, we conduct contrastive representation learning over both point-point and point-center pairs. This allows us to fully exploit relations between any two points and local data structures, and directly optimize the point feature space φ .

Point-Point Contrastive Learning. Our point-point contrastive learning is achieved by comparing pairs of points to push away point representations from different subclasses while pulling together those from the same subclass. The corres-

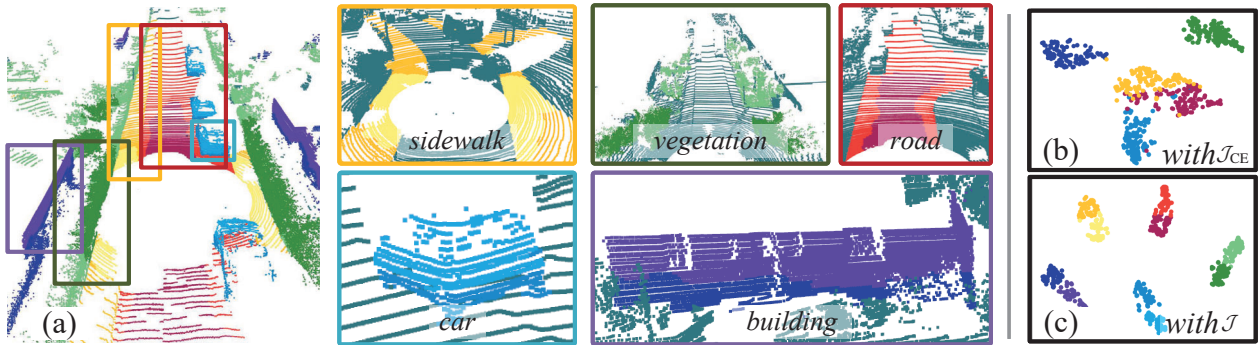


Figure 2: (a) Our clustering results for five classes, *i.e.*, *sidewalk*, *vegetation*, *road*, *car*, and *building*. (b-c) t-SNE visualization of point features $\{\mathbf{P}^c\}_c$ learned with \mathcal{J}_{CE} (Eq. (5)) and \mathcal{J} (Eq. (8)). We set $M = 2$ here, see supplementary for analysis.

ponding training objective for each point p_n is defined as:

$$\mathcal{J}_{PPC}(p_n) = \frac{1}{|\mathcal{O}_{p_n}|} \sum_{p^+ \in \mathcal{O}_{p_n}} -\log \frac{\exp(\mathbf{p}_n \cdot \mathbf{p}^+ / \tau)}{\exp(\mathbf{p}_n \cdot \mathbf{p}^+ / \tau) + \sum_{p^- \in \mathcal{N}_{p_n}} \exp(\mathbf{p}_n \cdot \mathbf{p}^- / \tau)}, \quad (6)$$

where $\tau > 0$ is a scalar temperature parameter, \mathcal{O}_{p_n} and \mathcal{N}_{p_n} denote collections of *positive* and *negative* samples, respectively, for p_n . Training points belonging to the same cluster of p_n are positive samples, while being assigned to other clusters are negative. Note that the positive (negative) samples are not limited to a same training point cloud. To further boost our point-point contrastive learning, we follow the common practice in unsupervised representation learning [75, 78, 105] to build a memory bank per cluster, leading to $M \times |\mathcal{C}|$ memory banks totally. The memory banks gather point features of corresponding clusters from previous training batches, hence increasing the quantity and diversity of positive and negative samples. These designs deliver a *cross-scene* training scheme, rather than the current *de facto* scene-wise training paradigm that ignores the rich correspondences among points across different scenes. Minimizing Eq. (6) leads to a well-structured embedding space φ , where points with similar patterns are grouped close to each other while points with dissimilar patterns are separated.

Point-Center Contrastive Learning. With a similar spirit of point-point contrastive learning, *i.e.*, inspiring intra-cluster compactness and inter-cluster separation, our point-center contrastive learning strategy contrasts the similarities between points and cluster centers on the embedding space φ :

$$\mathcal{J}_{PCC}(p_n) = -\log \frac{\exp(\mathbf{p}_n \cdot \mathbf{q}^+ / \tau)}{\sum_{c,m} \exp(\mathbf{p}_n \cdot \mathbf{q}_m^c / \tau)}, \quad (7)$$

where \mathbf{q}^+ refers to the cluster center of point p_n . Eq. (7) lets p_n find out the assigned cluster center \mathbf{q}^+ from all the centers $\{\mathbf{q}_m^c\}_{c,m}$, so as to decrease the distance between \mathbf{p}_n and \mathbf{q}^+ , while increasing the distance between \mathbf{p}_n and other cluster centers. Since cluster centers are representative of the dataset, Eq. (7) provides a cheaper and more direct way to impose dataset-level context, or underlying data structures, on feature space optimization, compared with the point-point

contrastive learning (Eq. (6)). In practice, we find combining the two cluster-analysis based contrastive learning strategies yields the best performance (see detailed experiments in §4.5). One may also view point-center contrastive learning from an *information bottleneck* perspective [97, 106], wherein the deterministic clustering imposes a natural bottleneck and discretizes the embedding space φ as a finite set of cluster centers, *i.e.*, $\{\mathbf{q}_m^c\}_{c,m}$, through minimizing Eq. (7), as opposed to learning φ as a continuous vector space.

Overall Training Objective. The standard cross-entropy loss \mathcal{J}_{CE} in Eq. (1) is essentially a unary training objective that is only aware of point-wise semantic discrimination, without accounting for any underlying data structure and pairwise relations between training points. The clustering analysis based contrastive losses, *i.e.*, \mathcal{J}_{PPC} in Eq. (6) and \mathcal{J}_{PCC} in Eq. (7), are pairwise training objectives that exploit locally representative patterns for structure-aware, distance based point representation learning. Thus we assemble these two complementary training targets as our overall learning objective:

$$\mathcal{J} = \mathcal{J}_{CE} + \alpha(\mathcal{J}_{PPC} + \mathcal{J}_{PCC}). \quad (8)$$

Our training algorithm alternately performs within-class clustering over the point embedding space φ , and optimizes the whole segmentation network $h \circ \varphi$ with the semantic labels $\{\mathcal{L}^k\}_k$ and cluster assignments $\{\mathcal{A}^c\}_c$. As such, meaningful clusters capture fine-grained data structures and become informative supervisory signals for point representation learning; in turn, discriminative representations help obtain meaningful clusters and eventually ease point recognition. In Fig. 2 (b-c), we provide visualization of point embeddings learned by \mathcal{J}_{CE} and \mathcal{J} . As seen, after additionally considering clustering analysis based training targets, the point embedding space becomes more structured.

3.4. Algorithm Details

Online Clustering (§3.2). We group point samples of each class into M subclasses for exploiting latent structures of the entire dataset. We empirically set $M = 40$ and the momentum coefficient in Eq. (5) $\mu = 0.9999$ (related experiments can

be found in §4.5). Following [31], we set $\lambda=25$ in Eq. (3).

Clustering Analysis based Training (§3.3). Our clustering analysis based training strategy enforces the point feature space to better respect the discovered data structures. Following the common practice in contrastive learning [76, 99], we set the scalar temperature τ in Eqs. (6-7) as 0.1. For the cluster-wise memory bank, we sample 10 point features per-cluster from each scene and store all the sampled features of all the training point clouds $\{\mathcal{P}^k\}_k$. For the training loss \mathcal{J} (Eq. (8)), the coefficient is set as $\alpha=1$ (we empirically find our algorithm is insensitive to α when $\alpha \in [0, 1]$).

Point Cloud Segmentation Network $h \circ \varphi$. Our algorithm is a general supervised learning scheme for point cloud segmentation. In principle, it can be applied to any segmentation networks that can learn point-wise features. In our experiments, we approach our algorithm on four typical segmentation networks, including voxel-based [16], point-based [25], Transformer-based [26], and NAS-based [32].

Inference. Our training algorithm does not cause extra inference cost or network architectural modification during model deployment. The $M \times |\mathcal{C}|$ cluster centers and $M \times |\mathcal{C}|$ memory banks are directly discarded after network training.

4. Experiment

We first report our 3D segmentation results on static point clouds of urban scenes and indoor environments in §4.1 and §4.2, respectively. Then we assess our performance on 4D segmentation of outdoor point cloud sequences in §4.3. For thorough evaluation, in §4.4, we extend our algorithm to 3D object detection setting and conduct experiments. The hyperparameters mentioned in §3.4 are used for all the above experiments. Finally, in §4.5, we provide ablative analyses on the core components of our training algorithm.

Base Segmentation Networks. For thorough examination, we apply our training algorithm to Cylinder3D [16] (voxel-based), KPConv [25] (point-based), PTV1 [26] (Transformer-based), and SPVNAS [32] (NAS-based), which are representative for current mainstream network architectures in point cloud segmentation and with publicly accessible implementations. For fair comparison, we adopt their default implementation settings, including hyper-parameters and augmentation recipes.

4.1. 3D Segmentation on Static Urban Point Clouds

Dataset. SemanticKITTI [33] is a large-scale driving-scene dataset for point cloud segmentation. It has 43,000 scans with point-wise annotation, collected from 22 sequences. According to the official setting, we use sequences 00 to 10 for train (but 08 is left for val), and 11 to 21 for test. In single-scan challenge for static segmentation, 19 classes are used and mean intersection-over-union (mIoU) is reported.

Quantitative Result. Table 1 reports comparison results on SemanticKITTI single-scan challenge test. As seen, our algorithm improves the performance of the base seg-

Table 1: **Quantitative 3D segmentation results on SemanticKITTI [33] single-scan challenge test (§4.1).** For clarity, IoUs on 6 of 19 classes are given (c_1 : sidewalk, c_2 : parking, c_3 : building, c_4 : truck, c_5 : bicycle, c_6 : motorcyclist).

Method	mIoU(%)	c_1 (%)	c_2 (%)	c_3 (%)	c_4 (%)	c_5 (%)	c_6 (%)
PointASNL _(CVPR20) [52]	46.8	74.3	24.3	83.1	39.0	0.0	0.0
PolarNet _(CVPR20) [9]	54.3	74.4	61.7	90.0	22.9	40.3	5.6
RandLA-Net _(CVPR20) [21]	55.9	74.0	61.8	89.7	43.9	29.8	9.4
SqueezeSegV3 _(ECCV20) [8]	55.9	74.8	63.4	89.0	29.6	38.7	20.1
SalsaNext _(ISVC20) [10]	59.5	75.8	63.7	90.2	38.9	48.3	19.4
FusionNet _(ECCV20) [46]	61.3	77.1	68.8	92.5	41.8	47.5	11.9
JS3C-Net _(AAAI21) [107]	66.0	72.1	61.9	92.5	54.3	59.3	39.9
AF2S3Net _(CVPR21) [108]	69.7	72.5	68.8	87.9	39.2	65.4	74.3
RPVNet _(ECCV21) [109]	70.3	80.7	70.3	93.5	44.2	68.4	43.4
PVKD _(CVPR21) [110]	71.4	77.5	70.9	92.4	53.5	67.9	50.5
KPConv _(ICCV19) [25]	58.8	72.7	61.3	90.5	33.4	30.2	11.8
KPConv + Ours	61.0 $\uparrow 2.2$	75.0	63.4	91.4	49.0	45.0	36.4
SPVNAS _{10.8M} _(ECCV20) [32]	62.3	73.8	63.2	90.9	50.9	40.6	21.8
SPVNAS _{10.8M} + Ours	64.3 $\uparrow 2.0$	73.9	64.0	91.4	48.0	48.9	23.2
Cylinder3D _(CVPR21) [16]	67.8	75.5	65.1	91.0	50.8	67.6	36.0
Cylinder3D + Ours	70.4 $\uparrow 2.6$	77.2	66.1	92.3	51.9	68.4	54.6

Table 2: **Quantitative 3D segmentation results on S3DIS [34] Area-5 (§4.2).** For clarity, IoUs on 5 of 13 classes are given (c_1 : wall, c_2 : column, c_3 : window, c_4 : door, c_5 : board).

Method	mIoU(%)	mAcc(%)	c_1 (%)	c_2 (%)	c_3 (%)	c_4 (%)	c_5 (%)
HPEIN _(ICCV19) [51]	61.9	68.3	81.4	23.3	65.3	40.0	65.6
PAT _(CVPR19) [20]	60.1	70.8	72.3	41.5	85.1	38.2	61.3
PointWeb _(CVPR19) [14]	60.3	66.6	79.4	21.1	59.7	34.8	64.9
MinkowskiNet _(CVPR19) [6]	65.4	71.7	86.2	34.1	48.9	62.4	74.4
SCF-Net _(CVPR21) [53]	63.8	-	-	-	-	-	-
BAAF-Net _(CVPR21) [111]	65.4	73.1	-	-	-	-	-
CGA-Net _(CVPR21) [112]	68.6	-	83.0	25.3	59.6	71.0	69.5
PTV1+CBL _(CVPR21) [113]	71.6	77.9	-	-	-	-	-
Stratified Trans. _(CVPR21) [114]	72.0	78.1	-	-	-	-	-
PTV2 _(NeurIPS21) [115]	72.6	78.0	-	-	-	-	-
KPConv _(ICCV19) [25]	67.1	72.8	82.4	23.9	58.0	69.0	66.7
KPConv + Ours	69.0 $\uparrow 1.9$	76.2 $\uparrow 3.4$	84.0	30.7	66.7	77.6	63.0
PTV1 _(ICCV21) [26]	70.4	76.5	86.3	38.0	63.4	74.3	76.0
PTV1 + Ours	72.2 $\uparrow 1.8$	79.6 $\uparrow 3.1$	88.1	49.3	65.3	79.4	81.0

mentation networks by solid margins. Concretely, it yields **2.2%**, **2.6%**, and **2.0%** mIoU gains over point-based KPConv [25], voxel-based Cylinder3D [16], and SPVNAS [32], respectively. Our algorithm also obtains consistent performance improvements across most classes. These results illustrate the wide potential benefit of our algorithm. Moreover, ‘‘Cylinder3D + Ours’’ reaches comparable results with published competitors. This is particularly impressive considering the fact that the improvement is solely achieved by our training scheme, without any network architectural modification and inference speed delay.

Qualitative Result. As shown in the top row of Fig. 3, our method can reduce errors over both small nature objects (such as *trunk*) and widely distributed classes (like *sidewalk*).

4.2. 3D Segmentation on Static Indoor Point Clouds

Dataset. S3DIS [34] is a famous 3D indoor parsing dataset. It contains 273M points collected from six areas and labeled with 13 classes. Following [1, 25], we use Area-5 as test scene to better test the generalization ability. We report two

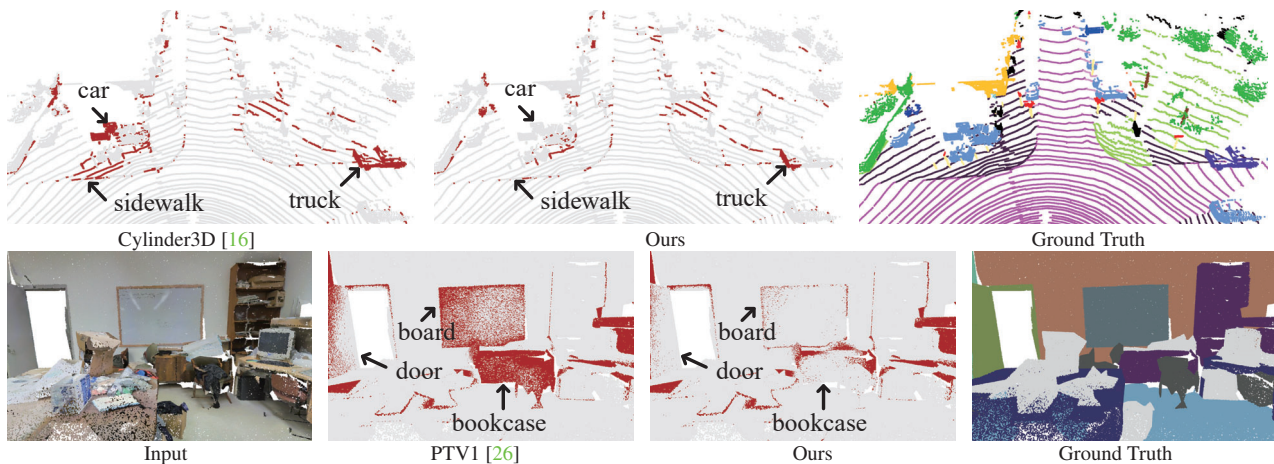


Figure 3: Error maps on SemanticKITTI [33] single-scan challenge val (top), and S3DIS [34] Area-5 (bottom). The differences are illustrated by arrows.

Table 3: **Quantitative 4D segmentation results** on SemanticKITTI [33] multi-scan challenge test (§4.3). IoUs on 6 of 25 classes are reported (c_1 : sidewalk, c_2 : moving car, c_3 : moving truck, c_4 : bicycle, c_5 : motorcyclist, c_6 : traffic-sign).

Method	mIoU(%)	c_1 (%)	c_2 (%)	c_3 (%)	c_4 (%)	c_5 (%)	c_6 (%)
TangentConv _[CVPR18] [38]	34.1	64.0	40.3	1.1	2.0	0.0	31.2
DarkNet53 _[ICCV19] [33]	41.6	75.3	61.5	14.1	30.4	0.0	31.2
TemporalLidarSeg _[3DV20] [71]	47.0	75.8	68.2	2.1	47.7	0.0	60.4
SpSeqnet _[CVPR20] [72]	43.1	73.9	53.2	41.2	24.0	0.0	48.7
KPConv _[ICCV19] [25]	51.2	70.5	69.4	5.8	44.9	0.0	53.9
KPConv+ Ours	53.2 ± 2.0	75.2	75.2	4.1	67.2	9.9	64.6
Cylinder3D _[CVPR21] [16]	52.5	74.5	74.9	0.0	67.6	0.2	61.4
Cylinder3D+ Ours	54.7 ± 2.2	76.9	81.7	11.9	55.9	3.0	68.0

metrics: mIoU and mean of class-wise accuracy (mAcc).

Quantitative Result. Table 2 summarizes the comparison results on S3DIS, showing our training algorithm also works well on large-scale challenging indoor point clouds. In particular, our algorithm brings impressive gains over KPConv, *i.e.*, 67.1% \rightarrow 69.0% and 72.8% \rightarrow 76.2%, in terms of mIoU and mAcc. Notably, with PTV1 as the backbone, our approach attains mIoU/mAcc of 72.2%/79.6%, outperforming PTV1+CBL (71.6%/77.9%).

Qualitative Result. As shown in the bottom row of Fig. 3, our method significantly reduces the errors of PTV1 [26] in an indoor environment of S3DIS [34] Area-5.

4.3. 4D Segmentation on Urban Point Sequences

Dataset. SemanticKITTI [33] multi-scan challenge is devoted to 4D point cloud segmentation. It involves six more classes to distinguish between moving objects and stationary ones for *car*, *truck*, *bicyclist*, *other-vehicle*, *person*, and *motorcyclist* categories. mIoU is adopted as the evaluation metric.

Quantitative Result. Table 3 reports our comparison results on SemanticKITTI [33] multi-scan challenge test. Our algorithm, again, leads to improvements over backbones, *i.e.*, 2.0% and 2.2% mIoU gain compared with KPConv [25] and Cylinder3D [16], respectively. This confirms our algorithm

Table 4: **Quantitative 3D detection results** on KITTI [37] challenge val (§4.4).

Difficulty	Method	mAP(%)	Car(%)	Pedestrian(%)	Cyclist(%)
Easy	Second _[SENSORS18] [35]	75.25	88.61	56.55	80.59
	Second+ Ours	78.60 ± 3.35	89.13	58.50	88.16
	PointPillar _[CVPR19] [116]	74.76	86.46	57.75	80.06
	PointPillar+ Ours	76.82 ± 2.06	88.34	58.19	83.92
Moderate	Second _[SENSORS18] [35]	66.25	78.62	52.98	67.16
	Second+ Ours	69.67 ± 3.42	82.97	55.64	70.39
	PointPillar _[CVPR19] [116]	64.08	77.28	52.29	62.68
	PointPillar+ Ours	66.07 ± 1.99	78.43	53.31	66.47
Hard	Second _[SENSORS18] [35]	62.69	77.22	47.73	63.11
	Second+ Ours	65.36 ± 2.67	78.55	50.91	66.61
	PointPillar _[CVPR19] [116]	60.76	74.65	47.91	59.71
	PointPillar+ Ours	62.96 ± 2.20	77.14	49.15	62.61

is also applicable in point cloud sequences. Our algorithm also obtains superior performance for vehicle categories with moving patterns, such as *moving car*, *moving truck*, *moving other-vehicle*, *etc.* We attribute this to our capacity of capturing complex patterns and variations, which improves the robustness in dynamic scenes.

4.4. 3D Detection on Static Urban Point Clouds

To fully reveal the power of our idea, we conduct additional experiments on 3D object detection.

Algorithmic Modification. To apply our algorithm to the 3D object detection task and minimize the modification effort, we view the bounding box annotations as a form of coarse segmentation labels. For each labeled bounding box with semantic class $c \in \mathcal{C}$, we simply treat all the points within the bounding box as data examples of class c , which are used in our clustering analysis based representation learning (*cf.* Eqs. 6-7). Note that there is no change to the base 3D detection network, including the detection head.

Dataset. KITTI [37] is a standard benchmark for 3D object detection. We split 3712 scans for train and 3769 scans for val, with 3D bounding box annotations of vehicles, pedestrians and cyclists. Detection outcomes are evaluated

Table 5: Study of proposed training strategy on S3DIS [34] Area-5 and SemanticKITTI [33] multi-scan val set (§4.5).

	\mathcal{J}_{PPC} (Eq. (6))	\mathcal{J}_{PCC} (Eq. (7))	S3DIS mIoU(%)	S-KITTI mIoU(%)	Training Speed (sec/epoch)
Baseline (w/o clustering analysis)			67.1	53.3	281.46
Point-Point Contrast	✓		68.0	54.4	310.20
Point-Center Contrast		✓	68.4	54.7	310.28
Point-Point + Point-Center Contrast	✓	✓	69.0	55.7	311.71

Table 6: Curve of CE Loss on SemanticKITTI [33] single-scan challenge train (left) and val (right).

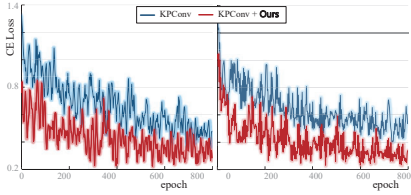


Table 7: Parameter studies on S3DIS [34] Area-5 and SemanticKITTI [33] (S-KITTI) multi-scan val set (§4.5). (mIoU(%) is reported.)

# Cluster	S3DIS S-KITTI		Memory Capacity		S3DIS S-KITTI		Coefficient μ		S3DIS S-KITTI	
	S3DIS	S-KITTI	S3DIS	S-KITTI	S3DIS	S-KITTI	S3DIS	S-KITTI	S3DIS	S-KITTI
$M = 1$	67.5	53.7	Mini-Batch (w/o memory)	68.0	54.4	$\mu = 0$	67.7	53.6		
$M = 10$	68.0	54.5				$\mu = 0.9$	68.0	54.0		
$M = 20$	68.5	55.2	$5 \times \#scene$	68.6	55.0	$\mu = 0.99$	68.5	54.7		
$M = 40$	69.0	55.7	$10 \times \#scene$	69.0	55.7	$\mu = 0.999$	68.6	55.3		
$M = 60$	68.9	55.2	$15 \times \#scene$	68.8	55.7	$\mu = 0.9999$	69.0	55.7		
$M = 80$	68.7	55.5	$20 \times \#scene$	68.7	55.6	$\mu = 0.99999$	68.8	55.5		

(a) Per-class cluster Num (b) Per-cluster memory (c) Momentum coefficient

under three regimes: *easy*, *moderate*, *hard*, defined according to occlusion and truncation levels of objects. Average precisions are reported with IoU thresholds of 0.7, 0.5, and 0.5, respectively for *car*, *pedestrian*, and *cyclist* classes.

Base Detection Networks. We apply our algorithm to two famous 3D detectors, *i.e.*, Second [35] and PointPillar [36].

Quantitative Result. Table 4 reports the experimental results on KITTI val. We can observe that, for both Second and PointPillar, our training algorithm brings notable performance gains, across different classes and under different regimes. This proves the high versatility of our algorithm.

4.5. Diagnostic Experiment

To test the efficacy of our core algorithm designs, we conduct a series of ablative studies on S3DIS [34] Area-5 and SemanticKITTI [33] multi-scan challenge val. We adopt KPConv [25] as our base segmentation network. The results are reported without post-processing or test-time augmentation.

Clustering Analysis based Network Training. We first test the efficacy of our core idea of clustering analysis based point representation learning. As shown in Table 5, the baseline model, trained in the standard strategy, gains 67.1% and 53.3% mIoU, on S3DIS and SemanticKITTI, respectively. Additionally considering point-point contrast \mathcal{J}_{PPC} (Eq. (6)) or point-center contrast \mathcal{J}_{PCC} (Eq. (7)) can lead to better performance. However, combining these two training objectives yields the best results, *i.e.*, 69.0% and 55.7%. These results verify that mining latent data structures can benefit detailed analysis of point cloud. Table 5 also gives comparisons for training speed. Our algorithm only brings negligible delay (~ 30 s for each epoch), confirming its high efficiency.

Per-Class Cluster Number M . We next investigate the impact of the cluster number M of each class. The results are summarized in Table 7a. $M = 1$ means that directly treating each class as a single cluster. This baseline obtains 67.5% and 53.7% mIoU, on S3DIS and SemanticKITTI, respectively. After clustering based fine-grained pattern mining,

we observe consistent improvements, *e.g.*, 67.5% \rightarrow 69.0% on S3DIS when $M = 40$. This verifies that **i**) there indeed exist some latent patterns in point clouds, and **ii**) these latent patterns are valuable for point cloud parsing. When $M > 40$, further increasing M gives marginal performance gains even worse results. We speculate this is because the model is distracted by some trivial patterns due to over-clustering.

Memory Bank. Then we study the influence of our memory bank in Table 7b. “Mini-Batch (w/o memory)” means that only computing contrast within each mini-batch, without the memory; it earns 68.0% and 54.4% mIoU, on S3DIS and SemanticKITTI, respectively. We then provision this baseline with class-wise memory bank with different capacities. When storing 10 point features per scene for each cluster, the best performance is achieved, *i.e.*, 69.0% and 55.7%.

Momentum Coefficient μ . Table 7c gives the performance with regard to the momentum coefficient μ (*cf.* Eq. 5), which controls the evolution speed of cluster centers. The model performs better with a relatively large coefficient (*i.e.*, $\mu = 0.9999$), showing that slow update is more favored. Moreover, at the extreme case of $\mu = 0$, the performance drops considerably, evidencing that simply approximating the cluster centers with per-batch cluster means is not a sound solution.

5. Conclusion and Discussion

We devise a clustering based supervised training scheme for point cloud analysis, which discovers and respects latent data structures during point representation learning. Rather than simply minimizing the point recognition error, we iteratively perform 1) unsupervised, within-class clustering based subclass pattern mining, and 2) clustering assignment based point embedding space optimization. Our algorithm is general and shows outstanding performance over various tasks and datasets. It also brings some new challenges, including the extension in instance-aware segmentation setting, and automatic estimation of the cluster number.

Acknowledgments: This work was supported in part by the Australian Research Council (ARC) under Grant DP200100938. This work was also partially supported by a China Scholarship Council (CSC) scholarship.

References

- [1] Lyne Tchaptmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. Segcloud: Semantic segmentation of 3d point clouds. In *3DV*, 2017. 1, 6
- [2] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *ICRA*, 2018. 2
- [3] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3D semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. 2
- [4] Hsien-Yu Meng, Lin Gao, Yu-Kun Lai, and Dinesh Manocha. Vv-net: Voxel vae net with group convolutions for point cloud segmentation. In *ICCV*, 2019. 2
- [5] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *IROS*, 2019. 2
- [6] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 2, 6
- [7] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *ICRA*, 2019. 2
- [8] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *ECCV*, 2020. 2, 6
- [9] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *CVPR*, 2020. 2, 6
- [10] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: fast, uncertainty-aware semantic segmentation of lidar point clouds for autonomous driving. *arXiv preprint arXiv:2003.03653*, 2020. 2, 6
- [11] Zeyu Hu, Xuyang Bai, Jiayang Shang, Runze Zhang, Jiayu Dong, Xin Wang, Guangyuan Sun, Hongbo Fu, and Chiew-Lan Tai. Vmnet: Voxel-mesh network for geodesic-aware 3d semantic segmentation. In *ICCV*, 2021. 1
- [12] Francis Engelmann, Theodora Kontogianni, and Bastian Leibe. Dilated point convolutions: On the receptive field size of point convolutions on 3d point clouds. In *ICRA*, 2020. 1
- [13] Binh-Son Hua, Minh-Khoi Tran, and Sai-Kit Yeung. Pointwise convolutional neural networks. In *CVPR*, 2018. 2
- [14] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. Pointweb: Enhancing local neighborhood features for point cloud processing. In *CVPR*, 2019. 2, 6
- [15] Zhiyuan Zhang, Binh-Son Hua, and Sai-Kit Yeung. Shellnet: Efficient point cloud convolutional neural networks using concentric shells statistics. In *ICCV*, 2019. 2
- [16] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *CVPR*, 2021. 1, 2, 6, 7
- [17] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 2017. 1, 2
- [18] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017. 2
- [19] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *CVPR*, 2019. 2
- [20] Jiancheng Yang, Qiang Zhang, Bingbing Ni, Linguo Li, Jinxian Liu, Mengdie Zhou, and Qi Tian. Modeling point clouds with self-attention and gumbel subset sampling. In *CVPR*, 2019. 2, 6
- [21] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *CVPR*, 2020. 1, 2, 6
- [22] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *CVPR*, 2018. 1, 2
- [23] Shenlong Wang, Simon Suo, Wei-Chiu Ma, Andrei Pokrovsky, and Raquel Urtasun. Deep parametric continuous convolutional neural networks. In *CVPR*, 2018. 1
- [24] Benjamin Ummerhofer, Lukas Prantl, Nils Thuerey, and Vladlen Koltun. Lagrangian fluid simulation with continuous convolutions. In *ICLR*, 2019.
- [25] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegeui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, 2019. 1, 2, 6, 7, 8
- [26] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021. 1, 2, 6, 7
- [27] Kirill Mazur and Victor Lempitsky. Cloud transformers: A universal approach to point cloud processing tasks. In *ICCV*, 2021.
- [28] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *CVPR*, 2021. 1, 2
- [29] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008. 2
- [30] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013. 2, 4
- [31] YM Asano, C Rupprecht, and A Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2019. 2, 4, 6
- [32] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *ECCV*, 2020. 2, 6
- [33] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, 2019. 2, 6, 7, 8
- [34] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. 2, 6,

- 7, 8
- [35] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 2, 7, 8
- [36] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019. 2, 8
- [37] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2, 7
- [38] Maxim Tatarchenko, Jaesik Park, Vladlen Koltun, and Qian-Yi Zhou. Tangent convolutions for dense prediction in 3d. In *CVPR*, 2018. 2, 7
- [39] Yecheng Lyu, Xinming Huang, and Ziming Zhang. Learning to segment 3d point clouds in 2d image space. In *CVPR*, 2020.
- [40] Yuqi Yang, Shilin Liu, Hao Pan, Yang Liu, and Xin Tong. Pfcnn: convolutional neural networks on 3d surfaces using parallel frames. In *CVPR*, 2020. 2
- [41] Daniel Maturana and Sebastian Scherer. Voxnet: A 3D convolutional neural network for real-time object recognition. In *IROS*, 2015. 2
- [42] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *CVPR*, 2017.
- [43] Dario Roth, Johanna Wald, Jurgen Sturm, Nassir Navab, and Federico Tombari. Fully-convolutional point networks for large-scale point clouds. In *ECCV*, 2018.
- [44] Truc Le and Ye Duan. Pointgrid: A deep network for 3d shape understanding. In *CVPR*, 2018.
- [45] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. In *NeurIPS*, 2019.
- [46] Feihu Zhang, Jin Fang, Benjamin Wah, and Philip Torr. Deep fusionnet for point cloud semantic segmentation. In *ECCV*, 2020. 2, 6
- [47] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *NeurIPS*, 2018. 2
- [48] Jiaxin Li, Ben M Chen, and Gim Hee Lee. So-net: Self-organizing network for point cloud analysis. In *CVPR*, 2018.
- [49] Francis Engelmann, Theodora Kontogianni, Jonas Schult, and Bastian Leibe. Know what your neighbors do: 3d semantic segmentation of point clouds. In *ECCV*, 2018.
- [50] Qiangui Huang, Weiye Wang, and Ulrich Neumann. Recurrent slice networks for 3d segmentation of point clouds. In *CVPR*, 2018.
- [51] Li Jiang, Hengshuang Zhao, Shu Liu, Xiaoyong Shen, Chi-Wing Fu, and Jiaya Jia. Hierarchical point-edge interaction network for point cloud semantic segmentation. In *ICCV*, 2019. 2, 6
- [52] Xu Yan, Chaoda Zheng, Zhen Li, Sheng Wang, and Shuguang Cui. Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In *CVPR*, 2020. 6
- [53] Siqi Fan, Qiulei Dong, Fenghua Zhu, Yisheng Lv, Peijun Ye, and Fei-Yue Wang. Scf-net: Learning spatial contextual features for large-scale point cloud segmentation. In *CVPR*, 2021. 6
- [54] Yanchao Lian, Tuo Feng, Jinliu Zhou, Meixia Jia, Aijin Li, Zhaoyang Wu, Licheng Jiao, Myron Brown, Gregory Hager, Naoto Yokoya, et al. Large-scale semantic 3-d reconstruction: Outcome of the 2019 ieee grss data fusion contest—part b. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:1158–1170, 2020. 2
- [55] Martin Simonovsky and Nikos Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *CVPR*, 2017. 2
- [56] Yiru Shen, Chen Feng, Yaoqing Yang, and Dong Tian. Mining point cloud local structures by kernel correlation and graph pooling. In *CVPR*, 2018.
- [57] Chu Wang, Babak Samari, and Kaleem Siddiqi. Local spectral graph convolution for point set feature learning. In *ECCV*, 2018.
- [58] Loic Landrieu and Mohamed Boussaha. Point cloud over-segmentation with graph-structured deep metric learning. In *CVPR*, 2019. 2
- [59] Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, and Jie Shan. Graph attention convolution for point cloud semantic segmentation. In *CVPR*, 2019.
- [60] Chao Chen, Guanbin Li, Ruijia Xu, Tianshui Chen, Meng Wang, and Liang Lin. Clusternet: Deep hierarchical cluster network with rigorously rotation-invariant representation for point cloud analysis. In *CVPR*, 2019.
- [61] Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. Deepgcn: Can gcn go as deep as cnns? In *ICCV*, 2019.
- [62] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *IEEE TOG*, 38(5):1–12, 2019. 2
- [63] Hang Su, Varun Jampani, Deqing Sun, Subhansu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. Splatnet: Sparse lattice networks for point cloud processing. In *CVPR*, 2018. 2
- [64] Huan Lei, Naveed Akhtar, and Ajmal Mian. Octree guided cnn with spherical kernels for 3d point clouds. In *CVPR*, 2019.
- [65] Artem Komarichev, Zichun Zhong, and Jing Hua. A-cnn: Annularly convolutional neural networks on point clouds. In *CVPR*, 2019.
- [66] Shiyi Lan, Ruichi Yu, Gang Yu, and Larry S Davis. Modeling local geometric structure of 3d point clouds using geo-cnn. In *CVPR*, 2019.
- [67] Jiageng Mao, Xiaogang Wang, and Hongsheng Li. Interpolated convolutional networks for 3d point cloud understanding. In *ICCV*, 2019. 2
- [68] Saining Xie, Sainan Liu, Zeyu Chen, and Zhuowen Tu. Attentional shapecontextnet for point cloud recognition. In *CVPR*, 2018. 2
- [69] Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges. In *CVPR*, 2021. 2
- [70] Hehe Fan, Xin Yu, Yuhang Ding, Yi Yang, and Mohan Kankanhalli. Pstnet: Point spatio-temporal convolution on point cloud sequences. In *ICLR*, 2021. 2
- [71] Fabian Duerr, Mario Pfaffler, Hendrik Weigel, and Jürgen

- Beyerer. Lidar-based recurrent 3d semantic segmentation with temporal memory alignment. In *3DV*, 2020. 2, 7
- [72] Hanyu Shi, Guosheng Lin, Hao Wang, Tzu-Yi Hung, and Zhenhua Wang. Spsequencenet: Semantic segmentation network on 4d point clouds. In *CVPR*, 2020. 7
- [73] Jiazhao Zhang, Chenyang Zhu, Lintao Zheng, and Kai Xu. Fusion-aware point convolution for online semantic 3d scene segmentation. In *CVPR*, 2020.
- [74] Yunsong Zhou, Hongzi Zhu, Chunqin Li, Tiankai Cui, Shan Chang, and Minyi Guo. Tempnet: Online semantic segmentation on large-scale point cloud series. In *ICCV*, 2021. 2
- [75] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 5
- [76] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 6
- [77] Yi Yang, Yueting Zhuang, and Yunhe Pan. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Frontiers of Information Technology & Electronic Engineering*, 22(12):1551–1558, 2021.
- [78] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 5
- [79] Junbo Yin, Dingfu Zhou, Liangjun Zhang, Jin Fang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Proposal-contrast: Unsupervised pre-training for lidar-based 3d object detection. In *ECCV*, 2022. 2
- [80] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE TPAMI*, 38(9):1734–1747, 2015. 2
- [81] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 2, 4
- [82] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010. 2
- [83] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2
- [84] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 2
- [85] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *ECCV*, 2020. 2
- [86] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *CVPR*, 2021.
- [87] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021.
- [88] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, 2021. 4
- [89] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *CVPR*, 2022.
- [90] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. Gmmseg: Gaussian mixture based generative semantic segmentation models. In *NeurIPS*, 2022.
- [91] Li Jiang, Shaoshuai Shi, Zhuotao Tian, Xin Lai, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Guided point contrastive learning for semi-supervised point cloud semantic segmentation. In *ICCV*, 2021.
- [92] Junbo Yin, Jin Fang, Dingfu Zhou, Liangjun Zhang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Semi-supervised 3d object detection with proficient teachers. In *ECCV*, 2022.
- [93] Qinghao Meng, Wenguan Wang, Tianfei Zhou, Jianbing Shen, Yunde Jia, and Luc Van Gool. Towards a weakly supervised framework for 3d point cloud object detection and annotation. *IEEE TPAMI*, 2021. 2
- [94] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *ICML*, 2016. 2
- [95] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *CVPR*, 2016.
- [96] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018. 4
- [97] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, 2019. 5
- [98] Tianfei Zhou, Liulei Li, Xueyi Li, Chun-Mei Feng, Jianwu Li, and Ling Shao. Group-wise learning for weakly supervised semantic segmentation. *IEEE Transactions on Image Processing*, 31:799–811, 2021.
- [99] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2020. 6
- [100] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *AAAI*, 2021.
- [101] James Liang, Tianfei Zhou, Dongfang Liu, and Wenguan Wang. Clustseg: Clustering for universal segmentation. *ICML*, 2023.
- [102] Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. Deep hierarchical semantic segmentation. In *CVPR*, 2022.
- [103] Wenguan Wang, Guolei Sun, and Luc Van Gool. Looking beyond single images for weakly supervised semantic segmentation learning. *IEEE TPAMI*, 2022. 2
- [104] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, 2018. 3
- [105] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *CVPR*, 2020. 5
- [106] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop*, pages 1–5, 2015. 5

- [107] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *AAAI*, 2021. 6
- [108] Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, and Bingbing Liu. (af)2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In *CVPR*, 2021. 6
- [109] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In *ICCV*, 2021. 6
- [110] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *CVPR*, 2022. 6
- [111] Shi Qiu, Saeed Anwar, and Nick Barnes. Semantic segmentation for real point cloud scenes via bilateral augmentation and adaptive fusion. In *CVPR*, 2021. 6
- [112] Tao Lu, Limin Wang, and Gangshan Wu. Cga-net: Category guided aggregation for point cloud semantic segmentation. In *CVPR*, 2021. 6
- [113] Liyao Tang, Yibing Zhan, Zhe Chen, Baosheng Yu, and Dacheng Tao. Contrastive boundary learning for point cloud segmentation. In *CVPR*, 2022. 6
- [114] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *CVPR*, 2022. 6
- [115] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *NeurIPS*, 2022. 6
- [116] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019. 7