U3DS^3: Unsupervised 3D Semantic Scene Segmentation

Jiaxu Liu[1] Zhengdi Yu[1] Toby P. Breckon[1,2] Hubert P. H. Shum[1]
Department of {Computer Science[1] | Engineering[2]}, Durham University, UK
{jiaxu.liu, zhengdi.yu, toby.breckon, hubert.shum}@durham.ac.uk

Abstract

Contemporary point cloud segmentation approaches largely rely on richly annotated 3D training data. However, it is both time-consuming and challenging to obtain consistently accurate annotations for such 3D scene data. Moreover, there is still a lack of investigation into fully unsupervised scene segmentation for point clouds, especially for holistic 3D scenes. This paper presents U3DS^3, as a step towards completely unsupervised point cloud segmentation for any holistic 3D scenes. To achieve this, U3DS^3 leverages a generalized unsupervised segmentation method for both object and background across both indoor and outdoor static 3D point clouds with no requirement for model pre-training, by leveraging only the inherent information of the point cloud to achieve full 3D scene segmentation. The initial step of our proposed approach involves generating superpoints based on the geometric characteristics of each scene. Subsequently, it undergoes a learning process through a spatial clustering-based methodology, followed by iterative training using pseudo-labels generated in accordance with the cluster centroids. Moreover, by leveraging the invariance and equivariance of the volumetric representations, we apply the geometric transformation on voxelized features to provide two sets of descriptors for robust representation learning. Finally, our evaluation provides state-of-the-art results on the ScanNet and SemanticKITTI, and competitive results on the S3DIS, benchmark datasets.

1. Introduction

As a crucial task in 3D computer vision, there has been increasing attention paid to point cloud segmentation in recent years due to its broad applicability to many real-world applications such as autonomous driving, virtual reality, robotics, and human-computer interaction. However, owing to the unordered and unstructured nature of point clouds, it is a non-trivial exercise to undertake segmentation upon them. In recent years, supervised point cloud segmentation approaches have made significant progress [1–7] against several benchmark datasets [8–11]. However, these approaches rely heavily on copious fully-annotated training data, in the form of labeled 3D point clouds. It is both time-consuming and labour-intensive to obtain such annotations accurately and consistently - especially for dense and complex 3D scenes. An alternative body of work leverages semi-supervised [12] and weakly-supervised [13–16] approaches to mitigate the labelled data requirements, but still require labour-intensive annotation at some level and lack of being readily scalable and adaptable to new datasets. Our work aims to characterize 3D features without any explicit guidance allowing it to learn from the intrinsic structure of the data, and offer independence from erroneous, bias or inconsistent annotations, which significantly differ from prior weakly-supervised methods. To date, there are only a handful of prior works trying to address fully unsupervised segmentation for point clouds [17–20]. However, these approaches essentially focus on object-level segmentation or co-segmentation and cannot recover the full 3D scene labels without extra scene priors [18–20] and only a recent work [17] has attempted to address fully unsupervised semantic segmentation for 3D scenes. Our proposed new U3DS^3 approach performs full holistic segmentation for the entire 3D scene in a scene-agnostic manner, spanning both indoor and outdoor scenarios across differing metric scales and achieving superior results on ScanNet [9] and SemanticKITTI [11] when compared to [17].

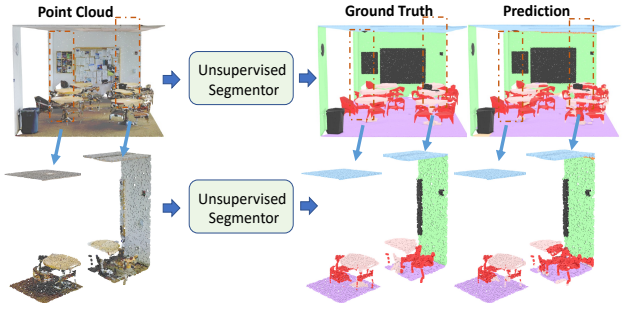

Figure 1. U3DS^3 unsupervised point cloud semantic segmentation (illustrated on S3DIS dataset [8]). Left to right: real scene, ground truth and U3DS^3 segmentation results for the full scene (upper), for a single point cloud block input (lower).

Despite the growth of unsupervised learning on 2D image segmentation [21–25], there is a lack of in-depth investigation into any 3D point cloud equivalent. Although some achievements in unsupervised segmentation learning have addressed 3D point cloud data via domain adaptation [26, 27], our work does not rely upon transfer learning. OGC [19] leverages the dynamic motion pattern of a

(LiDAR derived) point cloud sequence to acquire dynamic tracks and achieve competitive results for object-level segmentation. Similarly, Yang et al. [18] successfully apply unsupervised learning for object co-segmentation in point clouds. [17] made the first attempt towards unsupervised 3D semantic segmentation via region growing to generate high-quality over-segmentation, but their method does not fully leverage the intrinsic geometric information of the point clouds and tends to predict over-smooth segmentations with more background (e.g. floor, wall) and overlook detailed object categories of the scene.

Traditional clustering methods, like k-means [28] and DBSCAN [29], can be beneficial in establishing unsupervised semantic segmentation baselines. However, these methods still exhibit notable drawbacks. k-means [28], for instance, struggles to converge effectively with non-convex datasets, exhibits weaknesses in handling uneven data distributions, and struggles to form coherent clusters in the presence of outliers and data noise. Interestingly, some existing unsupervised approaches [17, 21] also incorporate k-means as a component of their algorithms. On the other hand, DBSCAN [29] encounters challenges when dealing with categorical features, often fails to identify clusters with varying densities, requires a drop in density to identify boundaries, and experiences decreased performance in high-dimensional scenarios.

The goal of our approach is to enable a generalized method that is able to perform semantic segmentation for large-scale indoor and outdoor 3D scenes without utilizing any human labels or dynamic information between LiDAR frames. This paper takes a new step towards scene-level unsupervised semantic segmentation with a novel strategy. Specifically, we first apply voxel cloud connectivity segmentation (VCCS) [30] to generate the initial superpoint and merge them according to the distance and normals of the superpoints. Following this, we propose the baseline method by applying mini-batch k-means [31] on the features of a 3D point cloud to generate and update the clustering centroids, and subsequently calculate the distance between features and clustering centroids to assign labels for each point as pseudo-labels under the guidance of the superpoint. After that, we train the network with the pseudo-labels to provide new network parameters for the next iteration of clustering. Subsequently, we apply a non-parametric classifier that operates solely on the feature space distance. Finally, by leveraging the invariance and equivariance of the volumetric representations, we are able to apply differing volumetric transformations on the point cloud input and a subsequent voxelized reverse geometric transformation on these feature representations.

In this manner, our network is capable of producing several variant feature representations from the same data source. This transformation operation is derived from a very intuitive sense that the same inputs should result in similar predictions even under geometric transformation due to the principle of invariance. Fundamentally, we learn a feature representation that maximizes the effective semantic class separation. We provide two pathways to enforce color invariance and geometric equivariance that each provide our underlying inductive bias for semantic consistency and geometric structure by way of consistent clustering assignment across the two pathways. This is performed via iterative optimization of the clustering loss, which enforces a discriminative feature space capable of high-level visual similarity disambiguation. Finally, we train our voxel-based method in an end-to-end manner. Furthermore, our evaluation illustrates promising results across both indoor and outdoor datasets, S3DIS [8], ScanNet [9] and SemanticKITTI [11], demonstrating the effectiveness and practicality of our method and providing an initial reference performance for completely unsupervised 3D semantic scene segmentation. Overall, we propose a simple yet effective framework that makes the new approach towards the task of unsupervised point cloud segmentation for holistic 3D scenes, named **U3DS**^{3}. Fig. 1 illustrates an initial qualitative result of our approach. Our key contributions are summarized as:

- We propose a novel unsupervised semantic segmentation method to leverage the invariance and equivariance through geometric transformation for both 3D indoor and outdoor holistic scenes.
- We analyze and compare existing clustering approaches and the concurrent state-of-the-art, demonstrating the advantages and superiority of our method for efficient unsupervised learning on large-scale point clouds of holistic 3D scenes with faster convergence.
- We conduct extensive experiments and ablation studies to demonstrate significant improvement over standard baselines, across the S3DIS [8] ScanNet [9] and SemanticKITTI [11] benchmark datasets, and illustrate both the practicability of the proposed framework and justify the intuition behind our design.

2. Related Work

In this section, we briefly summarize the prior literature on 3D Semantic Segmentation (Sec. 2.1) and Unsupervised Segmentation Learning (Sec. 2.2).

2.1. 3D Semantic Segmentation

To learn per-point semantics for 3D point clouds, many deep learning based approaches tackle 3D point cloud semantic segmentation tasks. PointNet [1] is a pioneering work and the first one to leverage the point-based encoding strategy, which is able to directly learn point features from the raw points and extract local information embedded in the neighbouring points. Following this work, more point-based methods [2–4] have been proposed. KPConv

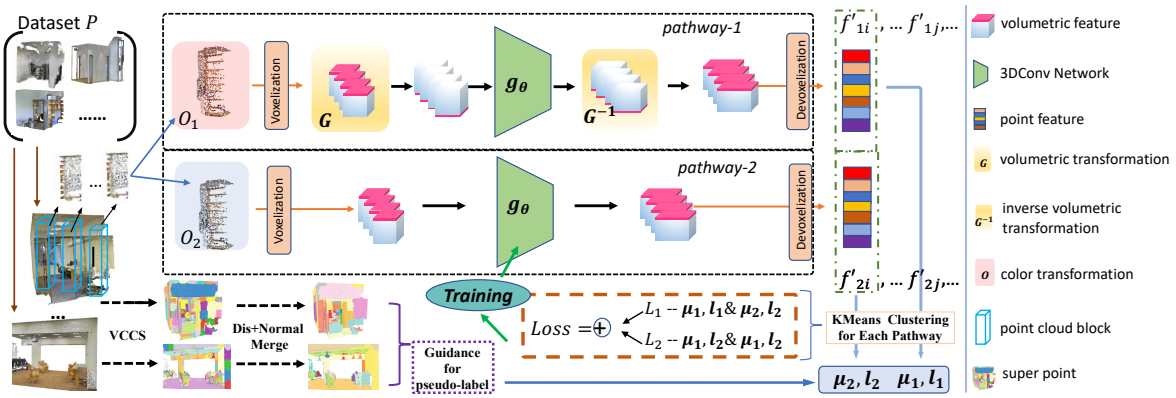

Figure 2. The illustration of the proposed unsupervised semantic segmentation method. Each input point cloud is assigned to two pathways and gives two groups of clustering centroids and labels for training. The point cloud is initially calculated to form a superpoint. This superpoint is then merged to produce a refined superpoint, which guides the generation of pseudo-labels. The pink part of the volumetric feature here denotes reverse the input tensor along the z axis.

[3] designs a kernel function for operating convolution in 3D space to tackle local geometric structures. RandLA-Net [4] proposes a more efficient framework by replacing a complicated point selection strategy with random sampling. On the other hand, voxel-based approaches [32, 33] typically employ a 3D convolutional neural network by converting the point cloud from uneven distribution to regular voxel grids. Some works [34, 35] explore the more efficient voxel-based method by conducting sparse convolution. PVCNN [36] proposes a point-voxel corporate method and utilizes trilinear de-voxelization for voxelized features for fine-grained feature extraction. Our method also leverages a 3D convolutional neural network and follows the trilinear de-voxelization approach from [36] to avoid extracting identical features for the points that lie in the same voxel grid. Moreover, [37,38] bring in graph convolutions to learn point features. However, all of these fully supervised methods require richly annotated training data, which are time-consuming and labour-expensive to obtain. To address the issue, Jiang et al. [12] propose a semi-supervised contrastive learning method for alleviating the tedious labelling cost. Zhang et al. [13] utilizes perturbed self-distillation to employ a weakly supervised method for point cloud semantic segmentation and reducing human annotations. In terms of unsupervised manner, KMeans [28] and DBSCAN [29] are classic methods and have no requirement for labelled data. However, these methods can only deal with simple object-level segmentation and lack of robustness under non-convex and uneven data distribution.

2.2. Unsupervised Segmentation Learning

The exploration of unsupervised 2D semantic segmentation shows more maturity compared to 3D. DeepCluster [39] clusters the feature vector of the entire dataset using k-means [28] to assign pseudo-labels and subsequently trains its encoder. Our method shares the common idea that iterative optimization of clustering can improve feature representation learning. Abdal et al. [25] propose an unsupervised segmentation framework that enables foreground and background separation for raw input images and segments class-specific Style-GAN images. Liu et al. [40] study the segmentation for object parts by means of semantic consistency of object parts, where the segmentation regions of the same part should be semantically consistent across object instances and robust to appearance and shape changes. Gansbeke et al. [24] put forward a two-step framework that adopts a predetermined mid-level prior in a contrastive optimization objective to learn pixel embedding for semantic segmentation. PICIE [21] notably proposes a pixel-level semantic segmentation method that can incorporate geometric consistency as an inductive bias to learn invariance and equivariance for photometric and geometric variations.

As for the 3D domain, some work already shows notable performance on unsupervised segmentation, but not especially for scene-level semantic segmentation. Yang et al. [18] employs object sampler and background sampler to tackle unsupervised point cloud co-segmentation for object-level segmentation by co-contrastive learning and mutual attention sampling. However, the co-segmentation method only works for objects and it is limited by the groups of common 3D objects. Some other works [26, 27] focus on unsupervised domain adaptation for point cloud semantic segmentation. OGC [19] can simultaneously identify multiple 3D objects in a single forward pass, without any human annotations, which leverages the dynamic motion patterns over (LiDAR captured) sequential point clouds as supervision signals to automatically discover rigid objects. However, OGC [19] needs the dynamic information of continuous point cloud frames as an input prior. Poux *et al.* [41] leverages the region growing method for indoor

unsupervised object-level segmentation, but the segmentation is only for generating larger object segment parts. GrowSP [17] is the only unsupervised 3D semantic segmentation that employs a progressively region-growing scheme to generate high-quality over-segmentation, however, their method does not fully leverage the intrinsic geometric information of the point clouds and tends to predict over-smooth segmentation and lose accuracy in intricate scenes. In contrast, our work aims to investigate unsupervised 3D semantic segmentation leveraging the intrinsic geometric information of the point clouds for holistic 3D scenes without any dynamic information or transfer learning prior.

3. U3DS^3 Methodology

This work formulates the task of unsupervised point cloud semantic segmentation as point-level segmentation, where every point within the point cloud needs to be assigned a label of a fixed number of semantic class labels.

To state formally, given a point cloud set $\boldsymbol{P}$ without labels, let $\boldsymbol{c} = \{c_i\}$ and $\boldsymbol{f} = \{f_i\}$ denote the point coordinates and the corresponding features from $\boldsymbol{P} \in \mathbb{R}^{N\times 3}$, $\boldsymbol{F} \in \mathbb{R}^{N\times d}$, where N is the number of points of the input point cloud, and d denotes the feature size, which contains coordinates, colours, and normalized positional information. Hence, the goal of this work is to learn a semantic segmentation function $\boldsymbol{g}_\theta$, which is able to predict per-point labels in an unsupervised way for $\boldsymbol{P}$ using only $\boldsymbol{c}$ and $\boldsymbol{f}$.

As shown in Fig. 2, for each input point region, we first apply two different colour transformations and afterwards convert them to the volumetric domain. For *pathway-1* in the top row, we implement a geometric transformation before the voxelized features are fed into the model. After the forward pass, we operate a corresponding inverse geometric transformation to the output features to ensure this representation shares the same properties with the non-transformed *pathway-2*. Subsequently, we cluster features from the different point cloud blocks and produce two groups of clustering centroids and labels, which can be used for further training and loss assembled from different pathways.

3.1. Superpoint

For all point clouds $P_1, P_2, P_3, \ldots$ within a point cloud set $\boldsymbol{P}$, we adhere to the VCCS [30] method to obtain initial superpoints for each point cloud. These can be denoted as $\left\{\left\{SP_1^1, SP_1^2, SP_1^3, \ldots\right\}, \left\{SP_2^1, SP_2^2, SP_2^3, \ldots\right\}, \ldots\right\}$, where SP_j^i represents the i-th superpoint in the j-th point cloud. The initial superpoints may vary across different point clouds. Subsequently, we employ a straightforward strategy to merge the superpoints within each scene: 1) Identify the smallest superpoint SP^i along with its two closest neighboring superpoints SP^{k1}, SP^{k2}; 2) Compute the vector addition of points within each superpoint and calculate the cosine similarity, here simply noted as $cos[SP^i, SP^{k1}]$; 3) Merge the smallest superpoint with the one that exhibits higher cosine similarity; 4) Repeatedly execute the above three steps until the superpoints reach a predetermined number. This simplistic approach is based on the principle that similar semantic objects possess comparable normals. Ultimately, the updated superpoints become $\left\{\left\{SP_1^{n1}, SP_1^{n2}, \ldots\right\}, \left\{SP_2^{n1}, SP_2^{n2}, \ldots\right\}, \ldots\right\}$, ensuring that the points within the same superpoint are assigned identical labels. We define the final superpoint count as a parameter, represented by γ_{sp}. For all datasets, the optimal value is empirically found as $\gamma_{sp} = 40$.

3.2. Voxelization and Devoxelization

We produce different representations for the input point cloud via the geometric transformation on the volumetric domain, where a voxel-based architecture is naturally adopted for such representation. Here, using voxelization and devoxelization in the pipeline, we present a simple yet effective network which contains only 3D convolutional layers with batch normalization without any additional component (details in Sec. 3.3).

Given the input points coordinate $\boldsymbol{c}$ with corresponding features $\boldsymbol{f}$ in the input blocks, we normalize the coordinates $\boldsymbol{c}$ before voxelizing the original points to gain scale invariance. Specifically, we normalize the coordinate $\boldsymbol{c}$ into [0,1] and denoted by $\boldsymbol{c}^* = \{c_i^*\}$. In this process, the point features (including the coordinates) do not change, and the normalized coordinates are only used for converting the feature to the proper volumetric space.

When transferring the features $\boldsymbol{f}$ with normalized coordinates $\boldsymbol{c^*} = \{\boldsymbol{x^*}, \boldsymbol{y^*}, \boldsymbol{z^*}\}$ into the voxel grids $\{\boldsymbol{V}_{m,p,q}\}$, the interpolated feature f_i for the voxel grid is calculated as the mean value of the features of points located in the grid.

$$\begin{aligned}\boldsymbol{V}_{m,p,q} = \frac{1}{K_{m,p,q}} \sum_{i=1}^{n} \boldsymbol{I}[floor(x_i^* \times r) = m, \\ floor(y_i^* \times r) = p, floor(z_i^* \times r) = q] \times f_i\end{aligned} \tag{1}$$

where r denotes the voxel resolution and $\boldsymbol{I}$ is an indicator function that indicates whether coordinates c_i belong to the voxel grid $\{m, p, q\}$. $K_{m,p,q}$ represents the count of points falling within the grid $\{m, p, q\}$, and $floor(\cdot)$ is floor function that outputs the greatest integer less than or equal to the input.

In terms of the per-point clustering, we need to devoxelize the voxel-based features output from the model g_θ to point-based features. We follow the trilinear interpolation of PVCNN [36] instead of the traditional nearest neighbor interpolation to ensure that nearby points are not assigned identical features.

3.3. Baseline: Clustering and Iteration

U3DS^3 applies a clustering-based method iteratively to generate pseudo-labels and train our baseline method, as in-

spired by DeepCluster [39]. Adapting [39] to the 3D domain is non-trivial due to the irregular nature and varying sparsity of point clouds. We present a simple yet effective strategy: switching between generating pseudo-labels via clustering with the current feature representations, and training new feature representations with the generated pseudo-labels. Different from [21, 39], the segmentation function g_θ should be able to produce per-point features, and we replace the parametric classifier with a non-parametric distance metric. Specifically, we denote the voxelization and devoxelization operations as $\boldsymbol{Z}$ and $\boldsymbol{Z}^{-1}$. The voxelized feature is $\boldsymbol{v} = \{v_i\} = \{\boldsymbol{Z}(f_i, c_i^*)\}$, and the output voxelized feature of the 3D convolutional function is $\boldsymbol{v}^{out} = g_\theta(\boldsymbol{v})$. Finally, the features for clustering can be denoted as $\boldsymbol{f}^{'} = \left\{ f_i^{'} \right\} = \left\{ \boldsymbol{Z}^{-1}(v_i^{out}, c_i^*) \right\}$. The main procedure can be separated as two parts:

(1) Using the current embeddings and k-means to cluster the points with superpoints guidance in the point cloud:

$$\min_{\boldsymbol{l},\boldsymbol{\mu}} \sum_i \left\| f_i^{'} - \mu_{l_i^{sp}} \right\|^2 \tag{2}$$

where l_i^{sp} denotes the cluster label of point c_i with the constraint of superpoint, and μ_k denotes the k-th cluster centroid. Note the features $f_i^{'}$ and the centroids μ_k have the same dimension.

(2) Using the class labels as pseudo-labels, we train a classifier via cross-entropy loss, which is shown in the point cloud setting as:

$$\min_{\theta,W} \sum_i L_{CE}\left(g_W(f_i^{'}), l_i^{sp}, \boldsymbol{\mu} \right) \tag{3}$$

where g_W denotes the parametric classifier. Under the unsupervised setting, it will be very challenging to train a classifier jointly with constantly changing pseudo-labels. We therefore choose to label points only based on their cosine distance from to the clustering centroids in feature space. Specifically, the loss function shows the following format:

$$\min_{\theta} \sum_i L_{cluster}\left(f_i^{'}, l_i^{sp}, \boldsymbol{\mu} \right) \tag{4}$$

$$L_{cluster}\left(f_i^{'}, l_i^{sp}, \boldsymbol{\mu} \right) = -log\left(\frac{e^{-d(f_i^{'}, \mu_{l_i^{sp}})}}{\sum_t e^{-d(f_i^{'}, \mu_t)}} \right) \tag{5}$$

where $d(\cdot,\cdot)$ denotes the cosine distance.

3.4. Volumetric Transformations

To improve robustness in the unsupervised setting under different scenarios, we leverage the invariance and equivariance of volumetric representations of point clouds. Invariance means that the labelling should not change after applying different transformations such as colour jittering. Equivariance in the volumetric domain means when we apply a geometric transformation to the point cloud, the corresponding 3D convolutional feature should be similarly transformed, and the corresponding labels are also wrapped according to this transformation.

For simplicity, we name the two pipelines processing the two representations as *pathway-1* and *pathway-2*. To produce two different representations for an individual input block, we apply a geometric transformation before volumetric feature extraction and then perform a corresponding inverse transformation on the final voxelized features.

Specifically, let $\boldsymbol{G}$ and $\boldsymbol{G}^{-1}$ denote the voxelized feature geometric transformation and its reverse transformation respectively, and $\boldsymbol{O}$ is the colour transformation. For point $\boldsymbol{c}$ with its feature $\boldsymbol{f}$, we apply different colour transformations for original features $\boldsymbol{f}$:

$$\boldsymbol{f}_1 = \boldsymbol{O}_1(\boldsymbol{f}), \boldsymbol{f}_2 = \boldsymbol{O}_2(\boldsymbol{f}) \tag{6}$$

Next, we transform these two features into the voxel grid, noting that $\boldsymbol{c}_1^*$ is actually equal to $\boldsymbol{c}_2^*$:

$$\boldsymbol{v}_1 = \boldsymbol{Z}(\boldsymbol{f}_1, \boldsymbol{c}_1^*), \boldsymbol{v}_2 = \boldsymbol{Z}(\boldsymbol{f}_2, \boldsymbol{c}_2^*) \tag{7}$$

After that, the voxelized feature transformations are applied to the volumetric domain: only the features of *pathway-1* are transformed whilst the other remains unchanged. The geometric transformations operate on the voxelized feature $\boldsymbol{v}$ and the corresponding reverse geometric transformations operate on the output voxel feature $\boldsymbol{v}^{out}$:

$$\boldsymbol{v}_1^{out} = \boldsymbol{G}^{-1}\left\{ g_\theta[\boldsymbol{G}(\boldsymbol{v}_1)] \right\}, \boldsymbol{v}_2^{out} = g_\theta(\boldsymbol{v}_2) \tag{8}$$

Subsequently, we perform de-voxelization to get the features for clustering:

$$\boldsymbol{f}_1^{'} = \boldsymbol{Z}^{-1}(\boldsymbol{v}_1^{out}, \boldsymbol{c}_1^*), \boldsymbol{f}_2^{'} = \boldsymbol{Z}^{-1}(\boldsymbol{v}_2^{out}, \boldsymbol{c}_2^*) \tag{9}$$

3.5. Losses and Labelling Scheme

Given input clouds $\boldsymbol{c}$ with features $\boldsymbol{f}$, according to the colour and geometric transformations introduced in Sec. 3.3, two different feature representations, $\boldsymbol{f}_1^{'}, \boldsymbol{f}_2^{'}$, can be produced. By leveraging these two features, we cluster the two representations separately to get two groups of centroids and pseudo-labels:

$$\boldsymbol{l^{(1)}}, \boldsymbol{\mu^{(1)}} = arg\min_{\boldsymbol{l},\boldsymbol{\mu}} \sum_i \left\| f_{1i}^{'} - \mu_{l_i^{sp}} \right\|^2 \tag{10}$$

$$\boldsymbol{l^{(2)}}, \boldsymbol{\mu^{(2)}} = arg\min_{\boldsymbol{l},\boldsymbol{\mu}} \sum_i \left\| f_{2i}^{'} - \mu_{l_i^{sp}} \right\|^2 \tag{11}$$

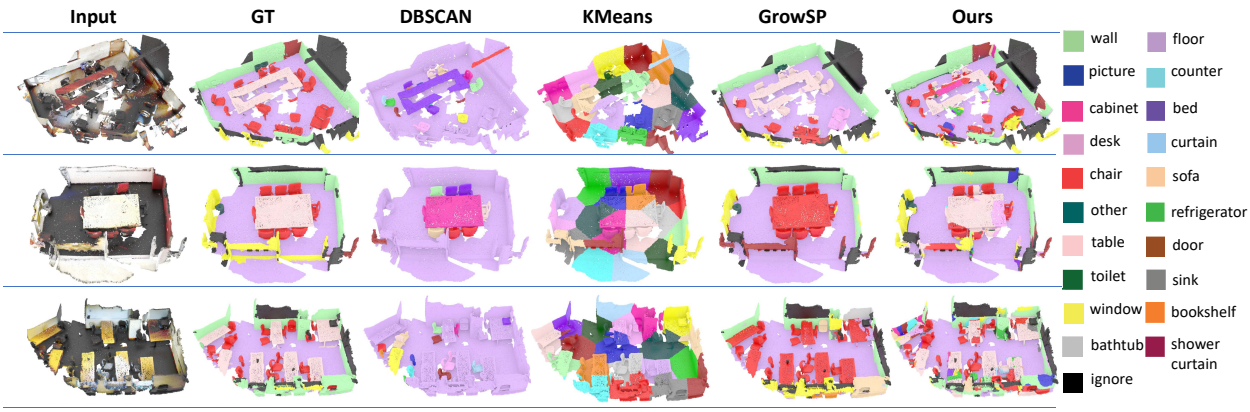

Figure 3. Qualitative results on ScanNet [9]. Each class label is assigned a colour (as per legend, right). This illustration shows superior segmentation performance compared to the baselines.

We then set two loss functions. Firstly, the feature representation should match the pseudo-labels produced by the same pathway:

$$L_1 = \sum_i L_{cluster}\left(f'_{1i}, l_i^{sp(1)}, \boldsymbol{\mu^{(1)}}\right) + \sum_i L_{cluster}\left(f'_{2i}, l_i^{sp(2)}, \boldsymbol{\mu^{(2)}}\right) \quad (12)$$

Similarly, the feature representation should whilst match the pseudo-labels produced by the different pathway:

$$L_2 = \sum_i L_{cluster}\left(f'_{1i}, l_i^{sp(2)}, \boldsymbol{\mu^{(2)}}\right) + \sum_i L_{cluster}\left(f'_{2i}, l_i^{sp(1)}, \boldsymbol{\mu^{(1)}}\right) \quad (13)$$

The final training objective is their summation:

$$L_{final} = L_1 + L_2 \quad (14)$$

The loss encourages the feature from one pathway to adhere to labels generated by another pathway, which encourages the network to label similarly to feature representations from different pathways.

Hungarian Algorithm: To match the clustering labels with the real labels, we utilize the Hungarian algorithm [42] accross the whole dataset every epoch. Specifically, where C is categories, P is the predicted set and G is the ground truth (GT) set. $S^{C\times C}$ is the matching matrix, where S_{ij} denotes the matching degree between i^{th} predicted category and j^{th} GT category. Criterion: finding bijection $\mathbf{f}: i \rightarrow j$ to maximize $\sum_{i=1}^{C} S_{i,f(i)}$.

4. Experiments

Implementation Details: We implement a simple yet effective framework with 8 layers 3D convolution, where each layer employs a 3D batch normalization and leaky rectified linear activation function (ReLU). The input point cloud contains 12D features, i.e., the point coordinates (x, y, z) in the normalized block coordinate system, colour information (R, G, B), per-point normals and normalized raw coordinates in the original scene coordinate system. Note that no colour information is provided in SemanticKITTI [11].

Method	Level of Supervision	mIoU	mAcc	oAcc
KMeans [28]	unsupervised	3.4	10.4	10.2
DBSCAN [29]	unsupervised	6.1	10.1	15.3
GrowSP [17]	unsupervised	25.4	44.2	57.3
U3DS^3 (ours)	unsupervised	**27.3**	**46.8**	**60.1**

Table 1. Semantic segmentation results on ScanNet dataset. We evaluate 20 categories on validation set

Training: We use a batch size of 4 with 4096 points per batch for all datasets. The chosen optimizer is stochastic gradient descent (SGD) with a learning rate of $1e-4$ and a weight decay of $1e-5$. We train our network for 10 epochs. For the geometric transformation in the volumetric domain, we reverse the order of tensors along the given x, y and z-axis respectively. The colour transformation comprises random contrast and random brightness adjustment. The output feature dimension from the model and the clustering feature dimension is set to 128. The resolution of the voxel grid is set to 32. Besides, we use the FAISS library [43] on GPU to compute the cluster centroids via employing a mini-batch k-means approach [31].

Evaluation: For evaluation and comparison with other methods, we choose two classical unsupervised clustering methods, k-means [28], DBSCAN [29], and the only unsupervised semantic segmentation method GrowSP [17] as baselines. Our method is evaluated with three metrics: overall accuracy (oAcc), mean accuracy (mAcc) and the mean intersection of union (mIoU) on all datasets. All experiments are performed on a single NVIDIA RTX 2080Ti GPU.

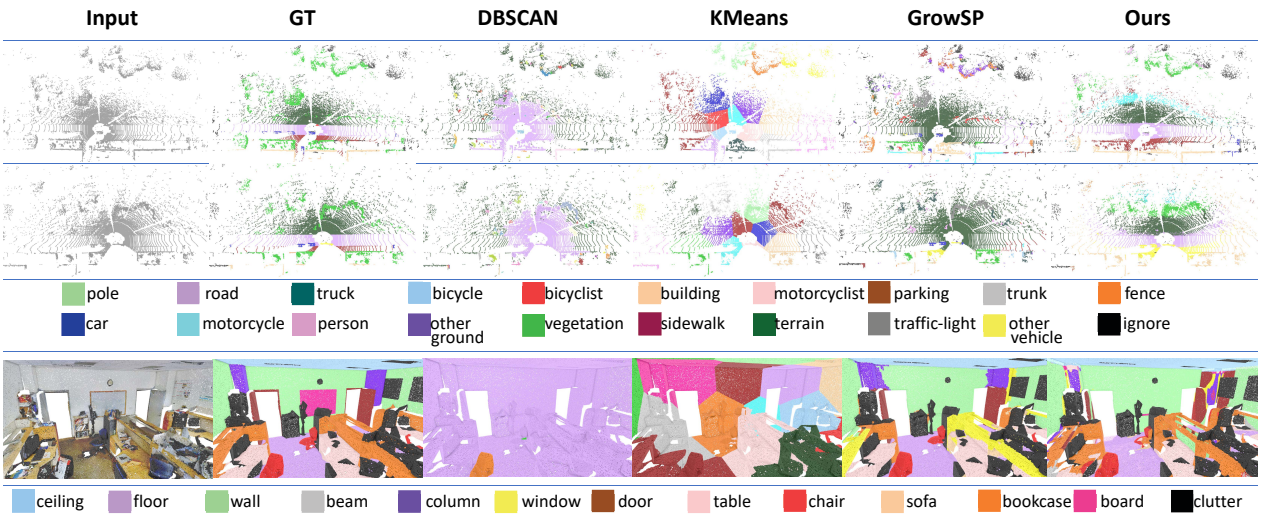

Figure 4. Qualitative results on SemanticKITTI [11] (top 2 rows) and S3DIS [8] (bottom row). Our method draws more versatile results compared with DBSCAN [29] and is more stable than k-means [28], which shows promising segmentation results.

Method	Level of Supervision	mIoU	mAcc	oAcc
KMeans [28]	unsupervised	2.5	8.1	8.2
DBSCAN [29]	unsupervised	6.8	7.5	17.8
GrowSP [17]	unsupervised	13.2	19.7	**38.3**
U3DS^3 (ours)	unsupervised	**14.2**	**23.1**	34.8

Table 2. Semantic segmentation results on SemanticKITTI dataset. We evaluate 19 categories on validation set

Method	Level of Supervision	mIoU	mAcc	oAcc
PTv2 [46]	fully supervised	72.6	78.0	91.6
KPConv [3]	fully supervised	67.1	72.8	-
SSP+SPG [47]	fully supervised	61.7	68.2	87.9
PointNet [1]	fully supervised	41.4	-	-
Jiang et al. [12]	semi-supervised (10%)	57.7	-	69.1
MT [48]	weakly supervised (1pt)	44.4	-	-
Zhang et al. [13]	weakly supervised (1pt)	48.2	-	-
KMeans [28]	unsupervised	9.4	21.2	22.1
DBSCAN [29]	unsupervised	9.2	19.8	17.5
GrowSP(12) [17]	unsupervised	**44.6**	**57.2**	**78.4**
U3DS^3 (ours)(12)	unsupervised	42.8	55.8	75.5
U3DS^3 (ours)	unsupervised	40.1	52.9	72.3

Table 3. Semantic segmentation results on S3DIS Area-5 are compared using mIoU, mAcc and oAcc across various methods. Where (12) indicates the exclusion of clutter, while the results without (12) are tested with 13 classes.

4.1. Datasets

We evaluate U3DS^3 on two indoor and one outdoor benchmark: S3DIS [8], ScanNet [9] and SemanticKITTI [11].

S3DIS [8] is a large-scale indoor scenes dataset which consists of 271 point cloud rooms in six areas. The annotations of each point in the point cloud scene belong to 13 semantic categories. We train the model in areas 1, 2, 3, 4, 6 and test it in area 5 following [1, 3, 44]. We exclude clutter and test with 12 classes for a fair comparison with GrowSP [17], nevertheless, we also test with 13 categories to compare with the existing supervised, weakly, and semi-supervised methods.

ScanNet-v2 [9] is an RGB-D real-world indoor dataset. It contains 1201 scenes for training, 312 for validation, and 100 for online testing. For scene semantic segmentation, it has 40 classes and one unlabelled class for training and 20 classes and for testing. We compare with existing clustering and unsupervised methods on the validation set.

SemanticKITTI [11]: is a large-scale outdoor dataset that is based on the KITTI Vision Odometry Benchmark. For the semantic segmentation task, it provides 22 sequences with point-wise annotation of 19 classes. Each sequence contains a number of scene scans collected by the complete 360 field-of-view of the employed automotive LIDAR, where sequences 11-21 are used for online testing, 08 is the validation set and the others are training sets.

Data Preparation: For all datasets, we choose $\gamma_{sp} = 40$ as the superpoint number for each scene. We first apply uniform downsampling to S3DIS [8] and ScanNet [9] with the sub-grid size 0.03 and subsequently follow PointCNN [44] to sample point clouds into blocks to ensure that each data sample in the batch has the same number of points. For S3DIS [8] and ScanNet [9], the block size is 1.5×1.5 on xy plane, and each block contains 4096 points. For SemanticKITTI [11], we set each block size as 5×5 on xy plane with 4096 points. For each point cloud, we utilize VCCS [30] to derive the initial superpoint. This is then merged for enhanced segmentation, as detailed in Sec. 3.1. Furthermore, due to the characteristics and predominance of roads in outdoor SemanticKITTI [11] datasets, we apply RANSAC [45] to fit a plane as the road for improved generation of superpoints. Note this process will not be utilized elsewhere.

4.2. Results and Comparison on Benchmarks

To thoroughly evaluate our U3DS^3, we test our methods on the indoor S3DIS [8], ScanNet [9] and outdoor SemanticKITTI [11] benchmarks. Tabs. 1 to 3 respectively shows the semantic segmentation results on the ScanNet, SemanticKITTI and S3DIS dataset. Not surprisingly, fully supervised methods provide the best performance. From Tab. 3, our method significantly outperforms the existing clustering methods, where it achieves 75.5% overall accuracy and 42.8 mIoU on the S3DIS dataset. Moreover, our method is even close to the performance reported by the fully supervised method [1] and some up-to-date weakly supervised methods [48, 49], which is a big step forward for unsupervised semantic 3D scene segmentation.

Moreover, we outperform GrowSP [17] on both the ScanNet and SemanticKITTI datasets. Specifically, as displayed in Tab. 1, our method achieves a superiority of +1.9 mIoU and +2.6 mAcc over their results. Additionally, Tab. 2 demonstrates that our method achieves 1 mIoU and 3.4 mAcc higher than GrowSP [17], despite having a slightly lower oAcc. Fig. 3 shows the qualitative comparison on S3DIS, which further demonstrates the superiority of our method.

4.3. Ablation Study

To showcase the effectiveness of each module and the different volumetric transformations. We conduct eight groups of experiments on the S3DIS [8] dataset: (1) the baseline approach proposed in Sec. 3.3, (2) adding colour transformation on the basis of the control group (1), (3) adding voxelized feature transformation on the basis of the control group (1), and (4) full model without prior superpoint, (5)-(8) different final prior superpoints as guidance. As shown in Tab. 4, our full model clearly outperforms the baseline on all of the evaluation metrics, benefiting from the delicate volumetric transformation design and superpoint prior. Groups (3) and (4) outperform by **+5** mIoU and **8** OA compared to the baseline. More interestingly, the improvement of adding the geometric transformation for equivariance is more significant than that of the invariance transformations, which is different from prior unsupervised learning work in the 2D domain [24, 25, 40]. It is known that point clouds essentially present much stronger geometric priors than 2D images with explicit 3D structures, which we believe can significantly help the 3D representations to be more robust and consistent cross-view and less sensitive to light changes and jittering. Moreover, the employment of superpoints can significantly enhance the overall performance. This enhancement is a result of the more abundant information of superpoints, which facilitates the pre-segmentation of the scene into higher-level semantic classes. Additional results are available in the supplementary material.

Baseline	Eqv	Inv	γ_{sp}	mIoU	mAcc	oAcc
✓				29.8	42.5	55.3
✓		✓		30.7	43.5	57.2
✓	✓			33.9	45.9	61.4
✓	✓	✓		34.8	46.3	63.2
✓	✓	✓	80	38.8	49.7	68.7
✓	✓	✓	60	41.0	52.6	72.4
✓	✓	✓	40	**42.8**	**55.8**	**75.5**
✓	✓	✓	20	41.9	53.9	74.3

Table 4. Ablation study on S3DIS Area-5: Eqv denotes equivariant voxelized feature transformation; Inv denotes invariant colour transformation. γ_{sp} denotes the final superpoint number.

4.4. Analysis

Our U3DS^3 approach demonstrates a promising level of performance on both indoor and outdoor datasets when compared to existing baselines. In contrast to GrowSP [17], our method achieves superior results on ScanNet [17] and SemanticKITTI [11]. As the scene complexity increases, the quality of GrowSP [17] superpoints tends to degrade. In contrast, our approach not only incorporates pre-segmentation but also employs a two-pathways training algorithm, leveraging the concepts of invariance and equivariance.

Nonetheless, slight performance degradation can occur in practical scenarios. To address this, we have implemented three strategies: (i) splitting the largest cluster when another cluster in the set reaches zero entities; (ii) applying mild centroid perturbation during updates; and (iii) reweighting for loss balancing using per-class pseudo-label ratios at each epoch. Additionally, our two-pathways approach expedites the convergence time during training. For instance, while training with only one pathway necessitates around 8 epochs to achieve convergence, the two-pathways approach accomplishes convergence in just 2-3 epochs.

5. Conclusion and Discussion

We propose a novel generalized unsupervised semantic segmentation method for both indoor and outdoor 3D scenes with objects and the background. Our method leverages a simple yet effective framework via clustering and iterative generation leveraging the invariance and equivariance of the volumetric representations with the assistance of superpoint. Experiments show promising performance on S3DIS, ScanNet and SemanticKITTI datasets which proves the superiority of our approach beyond all the existing baselines. This work aims to provide more insight for 3D unsupervised learning. Future work will explore improved point sampling strategies and an extension to point- or graph-based representations, benefiting other areas related to unsupervised learning, metric learning and 3D representation learning.

Acknowledgement: EPSRC NortHFutures (ref: EP/X031012/1).

References

[1] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 652–660. 1, 2, 7, 8

[2] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems (NIPS)*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc. 1, 2

[3] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, Francois Goulette, and Leonidas J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *Int. Conf. Comput. Vis. (ICCV)*, October 2019. 1, 2, 3, 7

[4] 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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2020. 1, 2, 3

[5] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics (TOG)*, 2019. 1

[6] Li Li, Hubert P. H. Shum, and Toby P. Breckon, "Less is more: Reducing task and model complexity for 3d point cloud semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2023, pp. 9361–9371. 1

[7] Ziyi Chang, George Alex Koulieris, and Hubert P. H. Shum, "On the design fundamentals of diffusion models: A survey," *arXiv*, 2023. 1

[8] 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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2016. 1, 2, 7, 8

[9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017. 1, 2, 6, 7, 8

[10] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu, "ShapeNet: An Information-Rich 3D Model Repository," Tech. Rep. arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 1

[11] 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 *Int. Conf. Comput. Vis. (ICCV)*, October 2019. 1, 2, 6, 7, 8

[12] 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 *Int. Conf. Comput. Vis. (ICCV)*, 2021. 1, 3, 7

[13] Yachao Zhang, Yanyun Qu, Yuan Xie, Zonghao Li, Shanshan Zheng, and Cuihua Li, "Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation," in *Int. Conf. Comput. Vis. (ICCV)*, October 2021, pp. 15520–15528. 1, 3, 7

[14] Jiahui Lei, Congyue Deng, Karl Schmeckpeper, Leonidas Guibas, and Kostas Daniilidis, "Efem: Equivariant neural field expectation maximization for 3d object segmentation without scene supervision," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 1

[15] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Hua Xiansheng, and Lei Zhang, "Box-supervised instance segmentation with level set evolution," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 1–18. 1

[16] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie, "Exploring data-efficient 3d scene understanding with contrastive scene contexts," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 15587–15597. 1

[17] Zihui Zhang, Bo Yang, Bing Wang, and Bo Li, "Growsp: Unsupervised semantic segmentation of 3d point clouds," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2023, pp. 17619–17629. 1, 2, 4, 6, 7, 8

[18] Cheng-Kun Yang, Yung-Yu Chuang, and Yen-Yu Lin, "Unsupervised point cloud object co-segmentation by co-contrastive learning and mutual attention sampling," in *Int. Conf. Comput. Vis. (ICCV)*, October 2021, pp. 7335–7344. 1, 2, 3

[19] Ziyang Song and Bo Yang, "Ogc: Unsupervised 3d object segmentation from rigid dynamics of point clouds," in *Advances in Neural Information Processing Systems (NIPS)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. 2022, vol. 35, pp. 30798–30812, Curran Associates, Inc. 1, 3

[20] Changfeng Ma, Yang Yang, Jie Guo, Fei Pan, Chongjun Wang, and Yanwen Guo, "Unsupervised point cloud completion and segmentation by generative adversarial autoencoding network," in *Advances in Neural Information Processing Systems (NIPS)*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, Eds., 2022. 1

[21] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan, "Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2021, pp. 16794–16804. 1, 2, 3, 5

[22] Xu Ji, Joao F. Henriques, and Andrea Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *Int. Conf. Comput. Vis. (ICCV)*, October 2019. 1

[23] Yassine Ouali, Céline Hudelot, and Myriam Tami, "Autoregressive unsupervised image segmentation," in *Eur.*

Conf. Comput. Vis. (ECCV), Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds., Cham, 2020, pp. 142–158, Springer International Publishing. 1

[24] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool, "Unsupervised semantic segmentation by contrasting object mask proposals," in *Int. Conf. Comput. Vis. (ICCV)*, October 2021, pp. 10052–10062. 1, 3, 8

[25] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka, "Labels4free: Unsupervised segmentation using stylegan," in *Int. Conf. Comput. Vis. (ICCV)*, October 2021, pp. 13970–13979. 1, 3, 8

[26] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Perez, "xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2020. 1, 3

[27] Wei Liu and Fulin Su, "Unsupervised adversarial domain adaptation network for semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 11, pp. 1978–1982, 2020. 1, 3

[28] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982. 2, 3, 6, 7

[29] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al., "A density-based algorithm for discovering clusters in large spatial databases with noise," 1996, vol. 96, pp. 226–231. 2, 3, 6, 7

[30] Jeremie Papon, Alexey Abramov, Markus Schoeler, and Florentin Worgotter, "Voxel cloud connectivity segmentation - supervoxels for point clouds," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2013. 2, 4, 7

[31] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th International Conference on World Wide Web*, New York, NY, USA, 2010, WWW '10, p. 1177–1178, Association for Computing Machinery. 2, 6

[32] Dario Rethage, Johanna Wald, Jurgen Sturm, Nassir Navab, and Federico Tombari, "Fully-convolutional point networks for large-scale point clouds," in *Eur. Conf. Comput. Vis. (ECCV)*, September 2018. 3

[33] Daniel Maturana and Sebastian Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015, pp. 922–928. 3

[34] Christopher Choy, JunYoung Gwak, and Silvio Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019, pp. 3075–3084. 3

[35] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018. 3

[36] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han, "Point-voxel cnn for efficient 3d deep learning," in *Advances in Neural Information Processing Systems (NIPS)*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc. 3, 4

[37] Martin Simonovsky and Nikos Komodakis, "Dynamic edge-conditioned filters in convolutional neural networks on graphs," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017. 3

[38] Loic Landrieu and Martin Simonovsky, "Large-scale point cloud semantic segmentation with superpoint graphs," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2018. 3

[39] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze, "Deep clustering for unsupervised learning of visual features," in *Eur. Conf. Comput. Vis. (ECCV)*, September 2018. 3, 5

[40] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu, "Unsupervised part segmentation through disentangling appearance and shape," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2021, pp. 8355–8364. 3, 8

[41] F. Poux, C. Mattes, and L. Kobbelt, "Unsupervised segmentation of indoor 3d point cloud: Application to object-based classification," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS)*, pp. 111–118, 2020. 3

[42] Harold W Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955. 6

[43] Jeff Johnson, Matthijs Douze, and Hervé Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2021. 6

[44] Wenxuan Wu, Zhongang Qi, and Li Fuxin, "Pointconv: Deep convolutional networks on 3d point clouds," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2019. 7

[45] Ondřej Chum, Jiri Matas, and Josef Kittler, "Locally optimized ransac," in *DAGM-Symposium*, 2003. 7

[46] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao, "Point transformer v2: Grouped vector attention and partition-based pooling," in *Advances in Neural Information Processing Systems (NIPS)*, 2022. 7

[47] Loic Landrieu and Mohamed Boussaha, "Point cloud oversegmentation with graph-structured deep metric learning," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2019. 7

[48] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems (NIPS)*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc. 7, 8

[49] Yachao Zhang, Zonghao Li, Yuan Xie, Yanyun Qu, Cuihua Li, and Tao Mei, "Weakly supervised semantic segmentation for large-scale point cloud," in *AAAI*, 2021, vol. 35, pp. 3421–3429. 8

U3DS^3: Unsupervised 3D Semantic Scene Segmentation Supplementary Material

Jiaxu Liu[1] Zhengdi Yu[1] Toby P. Breckon[1,2] Hubert P. H. Shum[1]
Department of {Computer Science[1] | Engineering[2]}, Durham University, UK
{jiaxu.liu, zhengdi.yu, toby.breckon, hubert.shum}@durham.ac.uk

1. Introduction

In this supplementary, we provide more ablation studies on ScanNet [1] and SemanticKITTI [2]. Moreover, we demonstrate the two-pathway approach facilitates quicker convergence. Additionally, we provide more qualitative results for different scenes to validate the effectiveness and generalization ability of our method on both indoor (S3DIS [3], ScanNet [1]) and outdoor(SemanticKITTI [2]) datasets. Furthermore, we provide a demo video for improved visualization (https://www.youtube.com/watch?v=X_NLmoh5Q0c.). Tables S1 and S2 present the ablation studies on ScanNet [1] and SemanticKITTI [2] datasets, respectively, demonstrating the efficacy of our method. Both tables confirm that each component of our approach performs effectively and that the choice of superpoint number significantly influences the final outcomes. Figure S1 illustrates that two-pathway training not only enhances performance but also expedites convergence. Figures S2 to S4 showcase the qualitative results on the SemanticKITTI [2], ScanNet [1] and S3DIS [3] datasets. We compare our results with two classical clustering methods and the current existing work GrowSP [4].

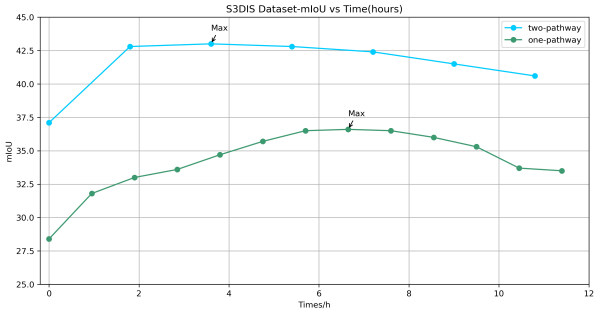

Figure S1. Curve of mIoU(%) changes over time(hour). The figure illustrates the expedited convergence achieved by the two-pathway approach.

Baseline	Eqv	Inv	γ_{sp}	mIoU	mAcc	oAcc
✓				11.2	25.5	36.3
✓		✓		12.0	26.3	37.5
✓	✓			14.3	29.3	40.4
✓	✓	✓		15.6	30.7	41.5
✓	✓	✓	80	22.7	41.5	52.8
✓	✓	✓	60	25.2	44.3	55.6
✓	✓	✓	40	**27.3**	**46.8**	**60.1**
✓	✓	✓	20	26.2	45.2	58.2

Table S1. Ablation study on ScanNet: Eqv denotes equivariant voxelized feature transformation; Inv denotes invariant colour transformation. γ_{sp} denotes the final superpoint number.

Baseline	Eqv	γ_{sp}	mIoU	mAcc	oAcc
✓			6.9	13.2	18.9
✓	✓		8.1	15.4	21.2
✓	✓	80	11.4	19.8	29.5
✓	✓	60	13.5	21.9	32.8
✓	✓	40	**14.2**	**23.1**	**34.8**
✓	✓	20	13.2	22.1	33.6

Table S2. Ablation study on SemanticKITTI: Eqv denotes equivariant voxelized feature transformation. However, there is no Inv due to the lack of color information in this dataset. γ_{sp} denotes the final superpoint number.

References

[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017. 1, 4

[2] 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 *Int. Conf. Comput. Vis. (ICCV)*, October 2019. 1, 2, 3

[3] 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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2016. 1, 5

[4] Zihui Zhang, Bo Yang, Bing Wang, and Bo Li, "Growsp: Unsupervised semantic segmentation of 3d point clouds," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2023, pp. 17619–17629. 1

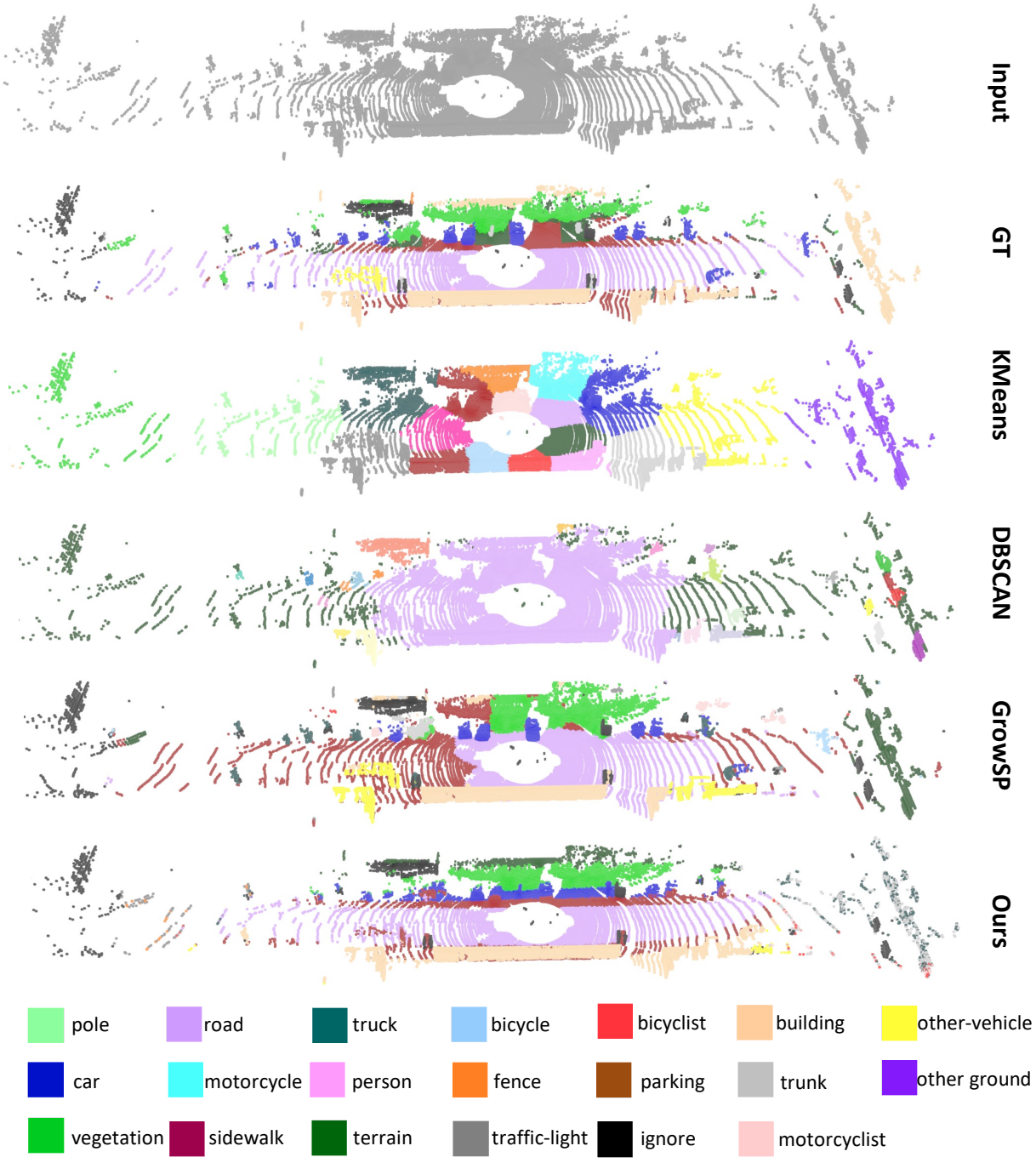

Figure S2. Qualitative example (a) on SemanticKitti [2]. Where each class is assigned to a colour (as per legend, bottom). This illustration shows superior performance compared to the baseline

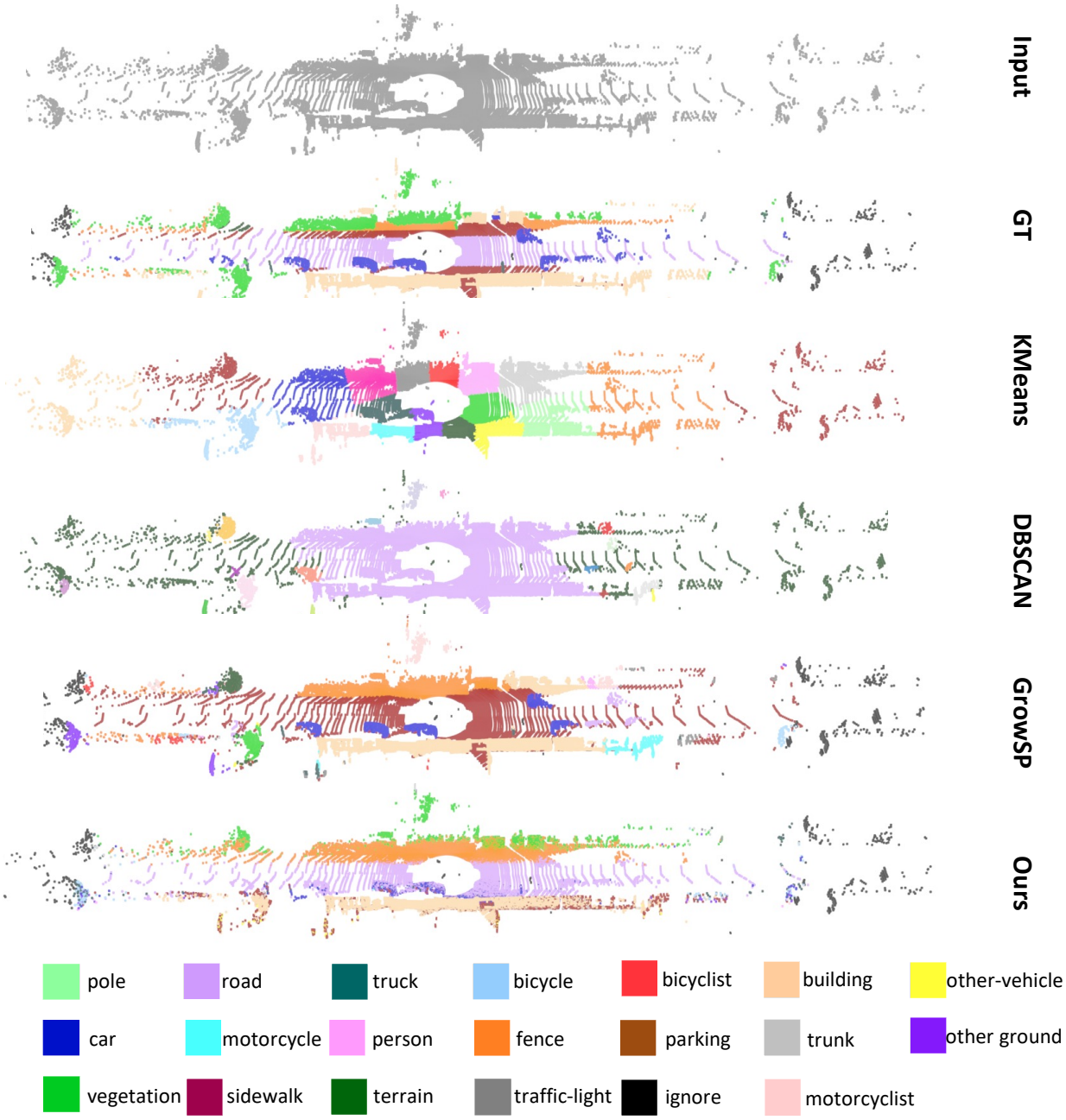

Figure S2. Qualitative example (b) on SemanticKitti [2]. Where each class is assigned to a colour (as per legend, bottom). This illustration shows superior performance compared to the baseline

Figure S3. Qualitative results on Scannet [1]. Evaluated with 20 categories exclude the ignored label, and each label is assigned to a colour.

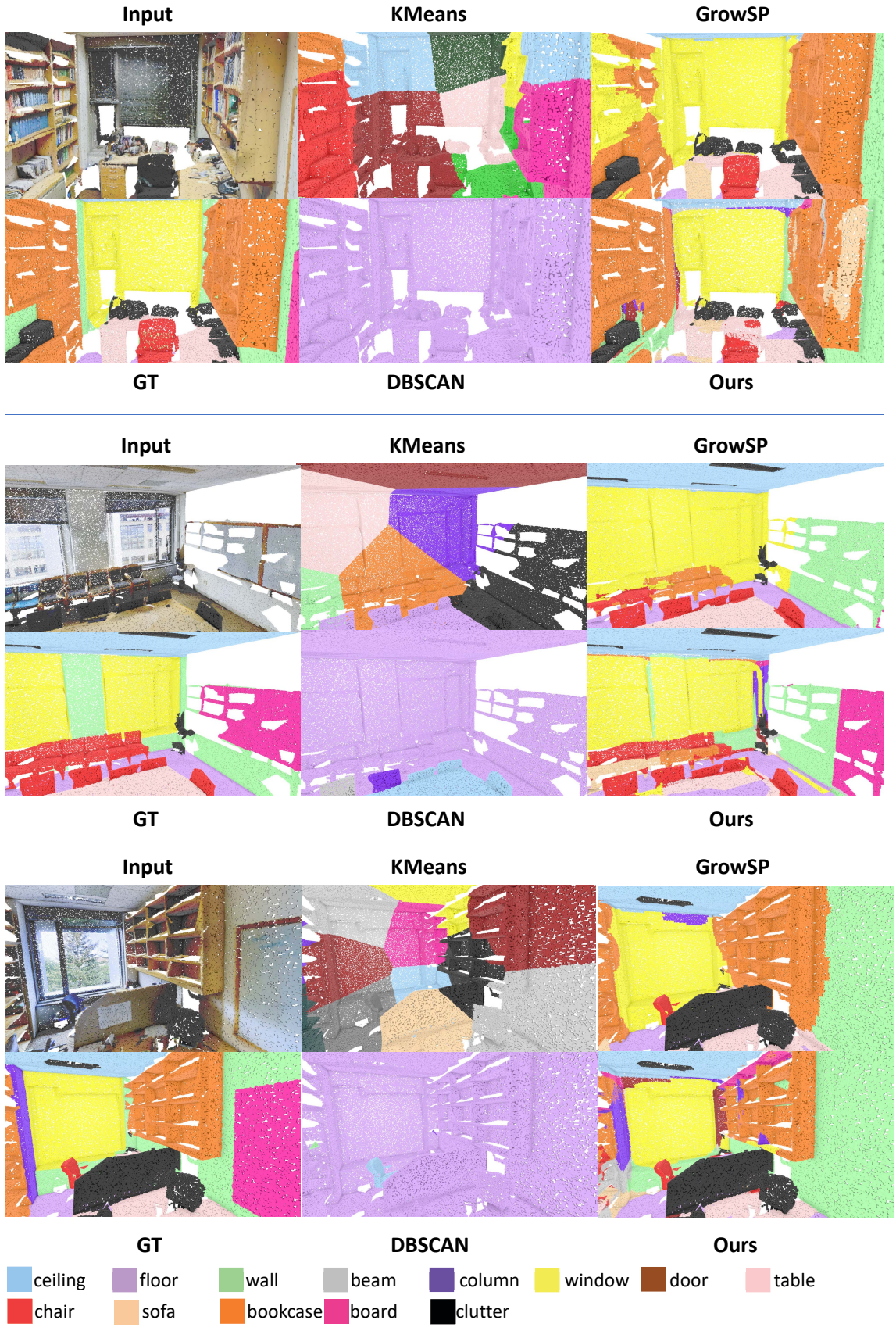

Figure S4. Qualitative results on S3DIS [3]. Each label at the bottom denotes one class, and this figure shows promising results.