

Learning and Memorizing Representative Prototypes for 3D Point Cloud Semantic and Instance Segmentation

Tong He * Dong Gong * Zhi Tian Chunhua Shen †
 The University of Adelaide

Abstract

3D point cloud semantic and instance segmentation is crucial and fundamental for 3D scene understanding. Due to the complex structure, point sets are distributed off balance and diversely, which appears as both category imbalance and pattern imbalance. As a result, deep networks can easily forget the non-dominant cases during the learning process, resulting in unsatisfactory performance. Although re-weighting can reduce the influence of the well-classified examples, they cannot handle the non-dominant patterns during the dynamic training. In this paper, we propose a memory-augmented network to learn and memorize the representative prototypes that cover diverse samples universally. Specifically, a memory module is introduced to alleviate the forgetting issue by recording the patterns seen in mini-batch training. The learned memory items consistently reflect the interpretable and meaningful information for both dominant and non-dominant categories and cases. The distorted observations and rare cases can thus be augmented by retrieving the stored prototypes, leading to better performances and generalization. Exhaustive experiments on the benchmarks, i.e. S3DIS and ScanNetV2, reflect the superiority of our method on both effectiveness and efficiency. Not only the overall accuracy but also non-dominant classes have improved substantially.

1. Introduction

The recent development of rapid and practical 3D sensors has provided easier ways to acquire 3D point cloud data, one of the widely used types of geometric data due to its simplicity [23]. 3D scene understanding is critically important and fundamental for various applications, such as robotics, autonomous driving, and virtual reality. The core tasks include semantic segmentation and instance segmentation on point clouds, *i.e.* assigning semantic labels and instance indication label for each point, respectively. Com-

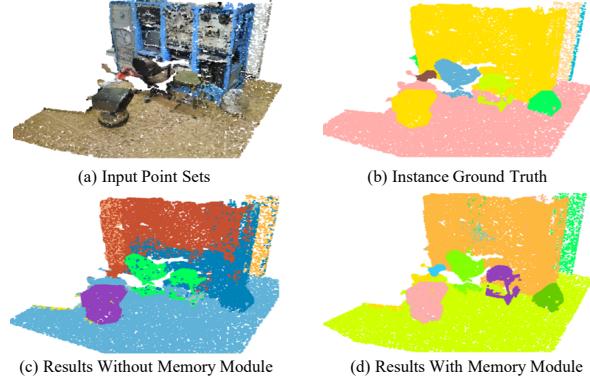


Figure 1 – Comparison of instance segmentation results between the proposed method with and without memory module. The performance of our method shows strong robustness against non-dominant cases.

paring to the studies on 2D images [12, 2, 5], semantic and instance on 3D point clouds lags far behind and have just started recently [35, 36, 41, 40, 13, 14].

Based on the pioneering works PointNet [23] and PointNet++ [25], directly processing point sets becomes simpler, more memory-efficient and flexible than handling the volumetric grids with 3D convolution [13, 39, 19]. Some following approaches [35, 36, 40, 41] propose to handle semantic and instance segmentation in an end-to-end network jointly for fine-grained description of the scene. Specifically, discriminative instance embeddings are learned to measure the instance-level clustering patterns of the points [36, 22].

Although existing methods have achieved some impressive results, we still can observe performance bottlenecks on different benchmark datasets [1, 3], especially on the non-dominant classes with less samples (see Figure 5). Suffering from the *catastrophic forgetting* issue [20, 32], deep networks can forget the non-dominant rare cases easily while learning on a dataset distributed off balance and diversely. On point cloud data, imbalance issue usually appears as the *category imbalance* and *pattern imbalance*, which is severer than that on 2D images [40]. Firstly, discrepancy among the proportions of different categories are significant. In an indoor scene (see Figure 1 and 6), most

*The first two authors contribute equally.

†Corresponding author: chunhua.shen@adelaide.edu.au.

points belong to the background (*e.g.* ground, ceiling, and wall), whereas the proportions of the objects (*e.g.* chairs, desks and monitors) are much smaller. For example, in S3DIS [1], the total amount of ceiling points is 50 times larger than chair. Secondly, the patterns of the points are imbalanced and distributed diversely, which are often caused by the complex geometric informations, such as positions, shapes and relative relationships among instances. Some rare instance cases only have limited examples across the whole dataset. For example, chairs are usually placed neatly in a conference room, while can also be placed in arbitrary positions (*e.g.*, stacking and back-to-back) in an office room, as is shown in Figure 1. Conventional methods [40] ignore this issue or simply resort to the focal loss [17], by down weighing the well learned samples during training. However, they cannot directly handle the non-dominant patterns, which can be easily overwhelmed and forgotten.

To address the above issues, we propose to learn and memorize the discriminative and representative prototypes covering all the samples, which is implemented as a memory-augmented network, referred to as MPNet. The proposed MPNet includes two branches for predicting point-level semantic labels and obtaining per-point embedding for instance grouping, respectively. As shown in Figure 2, the two branches access a shared compact memory via two separate memory readers, which dynamically calibrate the per-point features with the memory items and associate the two tasks via the shared memory. Given an input, MPNet retrieves the most relevant items in the memory for the extracted per-point features and feeds only retrieved features to the following segmentation tasks. Thus, driven by the task-specific training objectives, the compact memory is pushed to record the representative prototypes seen in mini-batches and associate them with newly seen patterns, alleviating the forgetting issues and strengthening the generalization.

In the proposed MPNet, the memory is maintained as a dictionary of the representative features and a semantic summarization, as shown in Figure 2. Since the memory is trained to represent all the instances compactly, the learned prototypes can express a shared understanding of various instances. We observe that the learned memory items (*i.e.* dictionary bases) can reflect interpretable and meaningful informations, such as position and structure (see Figure 3). Benefiting from the associative memory, the rare cases and distorted observations can be augmented by retrieving the stored prototypes, leading to better robustness and generalization. Additionally, different from previous methods relying on either pairwise relations computing [35] or KNN based feature aggregation [36], the proposed MPNet is free from complex and time-consuming operations, which is more efficient.

The main contributions are summarized as:

- We propose a memory-augmented network for point cloud instance segmentation (*i.e.* MPNet), which is trained to explicitly record the prototypes of the per-point features in a compact memory. The proposed MPNet is more effective and efficient than previous methods.
- The learned prototypes can consistently represent interpretable and meaningful concepts of various instances, including dominant and non-dominant cases.
- Our proposed MPNet can boost the performance by a large margin with limited consumptions on computation and memory. State-of-the-art performance is achieved, showing the superiority on both effectiveness and efficiency.

2. Related Work

Deep Learning for 3D Point Cloud Existing methods for extracting features for 3D point cloud can be roughly categorized into three groups, including voxel-based [39, 19], multi-view based [4, 13, 24, 30] and point-based [23, 25, 15, 26, 31]. [19, 39] are the pioneering works to transfer irregular points to regular volumetric grids, aiming to efficiently extract feature representation with 3D convolution. To reduce irrelevant operation on void places and save runtime memory usage, many works are proposed [27, 10]. Multi-view based methods extract features in both 2D and 3D domain. [30] is one of the pioneering multi-view based method, which apply view-pooling over the 2D predictions. 3D-SIS [13], proposed by Hou *et al.*, combine features from 2D and 3D via explicit spatial mapping in an end-to-end trainable network. PointNet [23] is the first deep-learning-based work to operate directly on point sets, which uses shared MLP (multi-layer perceptron) to extract per-point feature. PointNet++ [25] improves the performance by extracting a hierarchical representation. Many following works [15, 34, 31, 38, 16] have been proposed to get a better representation of local context. Due to its simplicity, we select PointNet++ as our backbone and leave the choices of other backbones for future work.

Instance Segmentation on Point Cloud Deep-learning-based instance segmentation for 3D point cloud is rarely studied until huge application potential has been discovered recently. SGPN [35] is the first deep learning based method working on this field. It first splits the whole scene into separate blocks. For every single block, per-point grouping candidates are proposed by predicting a similarity matrix that reflects affinity between each pair of points. A block merging algorithm is conducted for post-processing by taking segmentation results of the overlapped area into consideration. However, huge memory is needed for storing the pair-wise matrix, which makes it memory-consuming for post-processing. In order to solve this, Wang *et al.* proposed ASIS [36], which utilized a discriminative loss func-

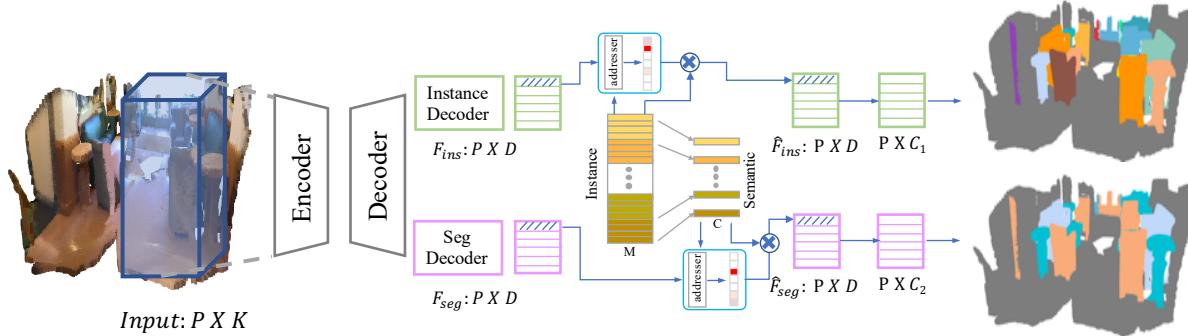


Figure 2 – The framework of our proposed MPNet, which contains two parallel branches with a shared encoder. A memory module is proposed to memorize representative prototypes that are shared by all samples. The maintained memory module is shared with all instances across different categories. Both distorted and rare cases can be augmented by retrieving the stored prototypes.

tion [2] to encourage points belonging to the same instance are mapped to a metric space with close distances. Moreover, in order to make the two tasks take advantage of each other, convolution and KNN search are applied for mutual feature aggregation of the two tasks, making it inefficient and time-consuming.

Memory Networks Memory based approaches have been discussed for solving various problems. NTM [11] is proposed to improve the generalization ability of the network by introducing an attention-based memory module. Gong *et al.* [9] proposed a memory augmented auto-encoder for detecting anomaly. Anomaly is detected by represented the input with prototypical elements of the normal data maintained in a memory module. Prototypical Network [29] maintains a category-wise templates for the problem of few-shot classification. Liu [18] proposed an OLTR algorithm to solve the open-ended and long-tail problem by associating a memory feature that can be transferred to both head and tail classes adaptively.

3. The Proposed Method

3.1. Overview of the Proposed MPNet

We propose to tackle the imbalance issue in point cloud semantic and instance segmentation by learning and memorizing prototypes of the cases seen during training. The discriminative and representative prototypes are stored in a memory module and can be accessed via specific readers. As shown in Figure 2, the proposed memory-augmented network (*i.e.* MPNet) adopts an encoder-decoder architecture, which is free from the specific design of the encoder and decoder. In the proposed MPNet, we use PointNet++ [25] to implement the encoder for per-point feature extraction. Two parallel decoders for instance segmentation and semantic segmentation are built upon the shared encoder. As described in the following, the memory is implemented as a dictionary to record the prototypes as bases. For the

both branches, given a per-point feature, two specifically designed memory readers are applied to generate addressing weights to access the memory, respectively, via soft-attention. The retrieved items from the memory are then applied for the following semantic labeling and instance grouping tasks. For any input sample, the relevant memory items are retrieved for the two tasks and also updated via prorogations driven by the task objectives and specifically designed instance regularizer (described in Section 3.3).

Given a set of input points $\{\mathbf{p}_i\}_{i=1}^P$ with $\mathbf{p}_i \in \mathbb{R}^K$, we can formulate the input of the network as a matrix $\mathbf{P} \in \mathbb{R}^{P \times K}$, where K denotes the input feature dimension and P denotes the total number of input points. Input features of each points may consist of both geometry and appearance information, *i.e.* 3D coordinate (x, y, z) and RGB values. The two branches produce features $\mathbf{F}_{\text{seg}} \in \mathbb{R}^{N \times D}$ and $\mathbf{F}_{\text{ins}} \in \mathbb{R}^{N \times D}$, respectively, where D denotes the dimension of features. Instead of directly using \mathbf{F}_{seg} and \mathbf{F}_{ins} to perform semantic and instance segmentation tasks, respectively, MPNet applies them as queries to retrieve the prototypes in the memory and then obtain alternative features $\hat{\mathbf{F}}_{\text{seg}}$ and $\hat{\mathbf{F}}_{\text{ins}}$, which are delivered to the following semantic classifier and instance embedding module. The memory is randomly initialized and updated during training. The two branches access the memory with specifically designed read heads,

3.2. Memory Representation for Prototypes

The *prototypes memory* is designed as a matrix $\mathbf{M} \in \mathbb{R}^{N \times D}$, where N is a hyper-parameter that defines the number of memory slots and D is the feature dimension that is identical with the outputs from the two branches. The N memory slots are used to restore the prototypes shared by all the instances across all categories. To easily represent the semantic characteristics, we define a *semantic memory* $\mathbf{C} \in \mathbb{R}^{C \times D}$ of \mathbf{M} , where C denotes the number of categories for the semantic segmentation task and each row of

\mathbf{C} represents the summary of a class. Although the memory slots in prototypes memory \mathbf{M} are shared to represent universal concepts of the all instances, to generate semantic summary \mathbf{C} from \mathbf{M} , we equally associate the N memory slots in \mathbf{M} with C categories and thus define $N = N_c \times C$, where N_c is denoted as the number per-category prototypes. As shown in Figure 2, the i -th row in \mathbf{C} , i.e. \mathbf{c}_i , can be seen as a average of the i -th subsegment in \mathbf{M} , i.e. rows in \mathbf{M} from $(i-1) \times N_c + 1$ to $i \times N_c$. Specifically, we obtain \mathbf{c}_i by averaging the submatrix \mathbf{M}_i :

$$\mathbf{c}_i = \frac{1}{N_c} \sum_{j=(i-1) \times N_c + 1}^{i \times N_c} \mathbf{m}_j, \quad (1)$$

where \mathbf{m}_j denotes the j -th row vector of \mathbf{M} .

Given the query features \mathbf{F}_{ins} and \mathbf{F}_{seg} , the instance grouping branch directly addresses the prototypes memory \mathbf{M} and the semantic labeling branch accesses the semantic summary \mathbf{C} , with two specifically designed readers. \mathbf{M} can be seen as an dictionary to restore the representative bases shared by all instances, since the instances cross different categories can share some common basic components and characteristics. As the semantic memory \mathbf{C} is a re-parameterization of \mathbf{M} , the two tasks are naturally associated together, without computation-consuming operations as [36]. Due to the supervision from both tasks, the learned and memorized prototypes are discriminative not only for grouping instances but also for semantic classification.

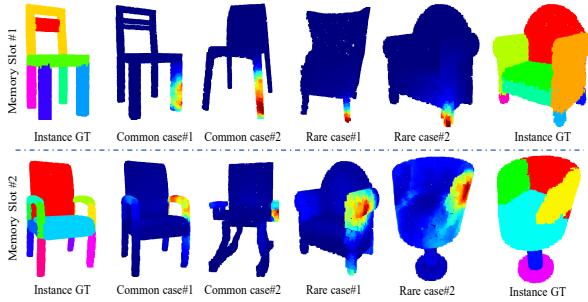


Figure 3 – Visualization of the memory representation. The goal is to find out part instance segmentation within an object, i.e. four chair legs in each example are four different instances. Even with various external shapes appearances, the instance memory successfully capture consistent representation for both dominant cases and rare cases.

3.3. Memory-augmented Instance Embedding

Memory reader for instance segmentation Given the i -th instance query feature $\mathbf{f}_{\text{ins},i}$ from \mathbf{F}_{ins} , an attention-based reader is proposed to address the most relevant prototypes from \mathbf{M} . The soft addressing weights w is calculated as follows:

$$w_{ij} = \frac{\exp(d(\mathbf{f}_{\text{ins},i}, \mathbf{m}_j))}{\sum_{j=1}^N \exp(d(\mathbf{f}_{\text{ins},i}, \mathbf{m}_j))}, \quad (2)$$

where \mathbf{m}_j is the j -th row vector of \mathbf{M} and $d(\cdot, \cdot)$ is function for measuring similarity of the i -th query item and the j -th prototype item. In MPNet, we utilize cosine distance. The alternated i -th instance feature $\hat{\mathbf{f}}_{\text{ins},i}$ from $\hat{\mathbf{F}}_{\text{ins}}$ can be calculated through: $\hat{\mathbf{f}}_{\text{ins},i} = \sum_{j=1}^N w_{ij} \mathbf{m}_j$.

To have a better understanding of the learned memory prototypes, we select the category of ‘Chair’ in PartNet [21] for training and visualization, as shown in Figure 3. Each chair is a testing sample and the goal is to find out instance part of object. For example, the four chair legs from a chair are noted as different instances. We select two memory prototypes \mathbf{m}_i and \mathbf{m}_j , and different colors in a chair refer to the addressing weights of the i -th and j -th columns in w (see Eq. (2)). For each memory item, the points that are addressing it have consistent geometric meaning, as shown in Figure 3. The consistency of the learned prototypes allows it to capture discriminative representation for both dominant and rare cases.

Instance-aware regularization The prototypes memory are updated via propagation driven by the training objectives. To make it effective, specifically designed regularization term R_{ins} is proposed, defined as follows:

$$R_{\text{ins}} = \frac{1}{K} \sum_{k=1}^K \frac{1}{N_k} \sum_{n=1}^{N_k} \|G(\hat{\mathbf{f}}_{\text{ins},n}) - GT_k\|^2, \quad (3)$$

where K is the instance number, N_k is the point number of k -th instance, $G(\cdot)$ is a simple MLP that predicts geometric centroid of the k -th instance. GT_k is the corresponding geometrical ground truth.

3.4. Memory-augmented Semantic Labeling

Memory reader for semantic segmentation Similar to the instance reader, the segmentation reader outputs an addressing weights for retrieving the most relevant categories. Given the i -th segmentation query $\mathbf{f}_{\text{seg},i}$ from \mathbf{F}_{seg} and j -th memory item \mathbf{c}_j , the feature for semantic segmentation is calibrated by: $\hat{\mathbf{f}}_{\text{seg},i} = \alpha^T \mathbf{C} = \sum_{j=1}^C \alpha_{ij} \mathbf{c}_j$. where \mathbf{c}_j is the j -th centroid for segmentation and α_{ij} is the similarity coefficients between $\mathbf{f}_{\text{seg},i}$ and \mathbf{c}_j , similar to Eq. (2).

Semantic memory regularization To force the centroids of different classes, i.e. the semantic summarization \mathbf{C} , to keep a separable distance, R_{seg} is proposed to regularize the large margin of the inter-class and the compactness of inner-class. Given the i -th calibrated feature $\hat{\mathbf{f}}_{\text{seg},i}$ and its semantic label y_i , the regularization term R_{seg} is calculated as:

$$R_{\text{seg}} = \max(0, \sum_{j=y_i} \|\hat{\mathbf{f}}_{\text{seg},i} - \mathbf{c}_j\| - \sum_{j \neq y_i} \|\hat{\mathbf{f}}_{\text{seg},i} - \mathbf{c}_j\| + m), \quad (4)$$

where m is the relaxation margin, which is set to 5 in all our experiments. Each \mathbf{c}_j performs like an anchor point and

pull the features with identical semantic labels close to it and push the features with different semantic labels away from it.

3.5. Loss Functions

Classification loss We use cross entropy loss L_{CE} for the semantic segmentation task. Instead of using softmax for normalization, we found squashing function proposed by [28, 18] provides a little higher performance on the accuracy of semantic segmentation. To summarize, classification loss is defined as:

$$L_{CE} = \frac{1}{P} \sum_{n=1}^P \text{CE}\left(\frac{\|fc(\hat{\mathbf{f}}_{\text{seg},i})\|^2}{1 + \|fc(\hat{\mathbf{f}}_{\text{seg},i})\|^2} \cdot \frac{fc(\hat{\mathbf{f}}_{\text{seg},i})}{\|fc(\hat{\mathbf{f}}_{\text{seg},i})\|}, y_n\right), \quad (5)$$

where CE refers to the cross entropy loss and P is the total number of examples. $fc(\cdot)$ is a fully convolution operation that projects the calibrated $\hat{\mathbf{f}}_{\text{seg},i}$ to the classification space.

Instance discriminative loss Similar to [36], given the retrieved instance features $\{\hat{\mathbf{f}}_{\text{ins},i}\}_{i=1}^P$, a simple layer perceptron is utilized to project the feature to the embedding space $\{\mathbf{e}_{\text{ins},i} \in \mathbb{R}^{c'}\}_{i=1}^P$ with feature dimension c' (we set $c' = 5$ in all our experiments). The loss is formulated as follows:

$$\begin{aligned} L_{\text{dis}} &= \frac{1}{K} \sum_{k=1}^K \frac{1}{N_k} \sum_{n=1}^{N_k} [\|\mathbf{e}_{\text{ins},n} - \boldsymbol{\mu}_k\| - \sigma_v]_+^2 \\ &\quad + \frac{1}{K(K-1)} \sum_{i=1}^K \sum_{\substack{j=1 \\ i \neq j}}^K [2\sigma_d - \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|]_+^2, \end{aligned} \quad (6)$$

where K is the instance amount and N_k is the number of k -th instance and $\boldsymbol{\mu}_k$ is the average embedding of the k -th instance, which is calculated by $\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N_k} \mathbf{e}_{\text{ins},n}$. σ_v and σ_d in Eq. (6) are respectively the margins for the variance and distance loss terms as defined in [2, 36].

Training objective As all operations are differentiable, prototypes memory module can be updated through back-propagation in an end-to-end manner. By combining the four losses discussed above, the training objective is formulated as:

$$L = L_{CE} + L_{\text{dis}} + R_{\text{seg}} + \lambda R_{\text{ins}}, \quad (7)$$

we found R_{ins} is sensitive to the learning rate and we set it 0.1 and maintain the others to 1.0 in all our experiments.

4. Experiments

To validate the effectiveness of our proposed method, both qualitative and quantitative experiments are conducted on two public datasets: Stanford 3D Indoor Semantic Dataset (S3DIS) [1] and ScanNetV2 [3].

4.1. Datasets

S3DIS dataset [1] covers more than $6000 m^2$ and is collected in 6 large-scale indoor areas. It includes 272 rooms and more than 215million points, each of which contains both instance and semantic annotations out of 13 classes. ScanNetV2 is another large-scale dataset for point cloud instance segmentation, which consists of 1613 indoor scans from 40 categories. The dataset is split into 1201, 312 and 100 for training, validating and testing, respectively.

4.2. Evaluation

Following [36] on the S3DIS dataset, the performance on Area-5 and k-fold cross-validation are reported in our experiments. For semantic segmentation, we present the overall accuracy (oAcc), which measures point-level accuracy, mean class accuracy (mAcc), which calculates average category-level accuracy and mean intersection-over-union (mIoU), which provides a measure for matching predicted segmentation results and the ground truth across all categories. For instance segmentation, four evaluation metrics are calculated, namely, $mConv$, $mWConv$, $mPrec$ and $mRec$. $mConv$ is defined as the mean instance-wise matching IoU score between ground truth and prediction. Instead of treating every instance equally, $mWConv$ is weighted by the size of each instance object. Moreover, traditional $mPrec$ and $mRec$ represents mean precision and mean recall with IoU threshold 0.5, which are widely used in the 2D image object detection task.

4.3. Implementation Details

For the S3DIS and ScanNetV2, similar to PointNet [23], each room is divided into $1m \times 1m$ blocks with a stride of $0.5m$. 4096 points are randomly sampled as input from each block during the training process. The feature for each point is consist of both color and geometric information, *i.e.* $R, G, B, X, Y, Z \dots$. Without special notation, all experiments are conducted using vanilla PointNet++ [25] as backbone (without introducing any multi-scale grouping operation). We use ADAM optimizer with initial learning rate of 1e-2, momentum of 0.9 and batch size of 16. The learning rate is divided by 2 for every 3×10^5 iterations. The hyper-parameters for metric learning are selected to be the same with [36], namely, $\sigma_v = 0.5$, $\sigma_d = 1.5$. The number for memory slots is set to 150 per-category. The loss weight for L is set to 0.01, which has a significant influence to the final performance. The whole network is trained end-to-end for 100 epochs in total.

During inference time, blocks within each room are merged in a snack pattern by utilizing the segmentation and instance results of the overlapped region. Detailed settings of the algorithm are identical with [35].

Table 1 – Ablation study on the S3DIS dataset Area-5 set with vanilla Pointnet++ as backbone. **FL** refers to focal loss. **InsMem** means the memory is updated by instance information. **SegMem** means the memory is updated by semantic segmentation supervision. **Regul** refers to the regularizations used in learning the prototypes memory. Both instance segmentation and semantic segmentation results are provided.

Method	FL	InsMem	SegMem	Regul	mPre	mRec	oAcc
Baseline	✓				52.3	41.4	86.2
					55.2	43.0	86.9
Ours	✓				58.9	47.0	87.7
		✓			60.2	47.2	88.1
		✓	✓	✓	62.5	49.0	88.2

4.4. Ablation Study

In this section, we describe the influence of each integration of aforementioned components. All the results are tested on S3DIS Area-5 for fair comparison. We first build a strong baseline which is similar to ASIS vanilla [36]. The framework has two independent decoders which are responsible for semantic segmentation and instance metric grouping, respectively. Using Pointnet++ as backbone, the baseline model achieves 52.3 *mPrec* and 41.4 *mRec* on S3DIS Area-5, which is 16.3 and 12.7 higher than SGPN [35], respectively. Built upon the strong baseline, our MPNet surpass it by a large margin via memorizing representative prototypes. In the following section, provide detailed analysis on different parts.

Focal Loss. The discrepancy among different categories are significant in 3D point cloud. Focal loss [17] has been widely used in different kinds of vision tasks due to the imbalance of data distribution. It addresses the problem by down-weighting the well-classified samples. However, it only alleviates the category imbalance to some extent and fail to solve the diverse distributed patterns. As shown in Table 1, focal loss can only improve the mean precision and mean recall by 2.9 and 2.4, respectively. Compared with Focal Loss, our method is more powerful to solve both data imbalance and pattern imbalance by recording and memorizing the prototypical patterns.

Prototypes Memory M and C. The representative and consistent prototypes are maintained in a memory module **M**, which is shared to represent universal concepts of all instances. Besides, a semantic memory **C** is served as a prototypes summary to efficiently represent the semantic characteristics. As shown in Table 1, using instance memory **M** alone can boost *mPre* from 52.3% to 58.9% and *mRec* from 41.4% to 47.0%. On the other hand, using segmentation memory **C** can bring another 1.3% and 0.5% improvement with the metric of *mPrec* and *oAcc*. As two tasks are highly correlated due to the shared encoder backbone, utilizing **M** can also brings about 1.5% improvement for se-

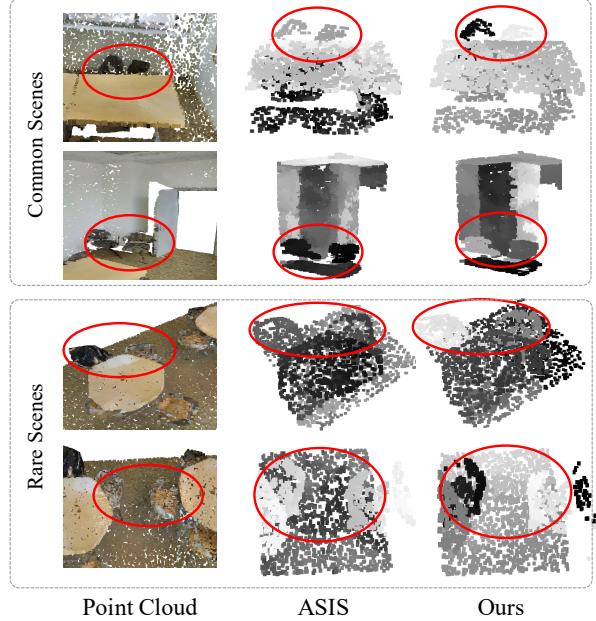


Figure 4 – Barnes-Hut t-SNE [33] visualization of our instance embedding on S3DIS Area-5 set (Best viewed when zoom in). The embedding feature is projected to 1-D and the distances is normalized to unit length so that the gap of gray-scale between different instances reflects the distances in the embedding space.

Table 2 – Instance Segmentation results on S3DIS dataset. Both Area-5 and 6-fold results are reported. All our results are achieved based on a vanilla PointNet++ backbone (without multi-scale grouping) for fair comparison.

Method	Year	mCov	mWCov	mPrec	mRec
Test on Area 5					
SGPN [35]	2018	32.7	35.5	36.0	28.7
ASIS [36]	2019	44.6	47.8	55.3	42.4
3D-BoNet [40]	2019	-	-	57.5	40.2
Ours	-	50.1	53.2	62.5	49.0
Test on 6-fold					
SGPN [35]	2018	37.9	40.8	31.2	38.2
MT-PNet [22]	2019	-	-	24.9	-
MV-CRF [22]	2019	-	-	36.3	-
ASIS [36]	2019	51.2	55.1	63.6	47.5
3D-BoNet [40]	2019	-	-	65.6	47.6
PartNet [21]	2019	-	-	56.4	43.4
Ours	-	55.8	59.7	68.4	53.7

mantic segmentation in terms of *oAcc*.

Regularization Loss. To effectively learn representative and discriminative prototypes, regularization losses are proposed in Eq. (4) and Eq. (3). The first one is to keep large-margin between different categories from memory **C**. The second one is designed for forcing the calibrated instance

Table 3 – Comparison per-class performance of our proposed method with state-of-the-art on S3DIS semantic segmentation task, tested on all areas. Our result utilize vanilla pointnet++ [25] without multi-scale group. Even with a simple baseline, the proposed method surpassed the graph based method by more than 1% with mIOU.

Method	OA	mIOU	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookcase	board	clutter
PointNet [23]	78.5	47.6	88.0	88.7	69.3	42.4	23.1	47.5	51.6	54.1	42.0	9.6	38.2	29.4	35.2
MS+CU [7]	79.2	47.8	88.6	95.8	67.3	36.9	24.9	48.6	52.3	51.9	45.1	10.6	36.8	24.7	37.5
G+RCU [7]	81.1	49.7	90.3	92.1	67.9	44.7	24.2	52.3	51.2	58.1	47.4	6.9	39.0	30.0	41.9
PointNet++ [25]	-	53.2	90.2	91.7	73.1	42.7	21.2	49.7	42.3	62.7	59.0	19.6	45.8	48.2	45.6
PointNeighbor [8]	-	58.3	92.1	90.4	78.5	37.8	35.7	51.2	65.4	64.0	61.6	25.6	51.6	49.9	53.7
DGCNN [37]	84.1	56.1	-	-	-	-	-	-	-	-	-	-	-	-	-
ResGCN-28 [15]	85.9	60.0	93.1	95.3	78.2	33.9	37.4	56.1	68.2	64.9	61.0	34.6	51.5	51.1	54.4
Ours PointNet++	86.8	61.3	94.0	94.1	76.6	53.4	33.6	54.2	62.7	70.2	60.2	36.6	53.4	54.3	53.5

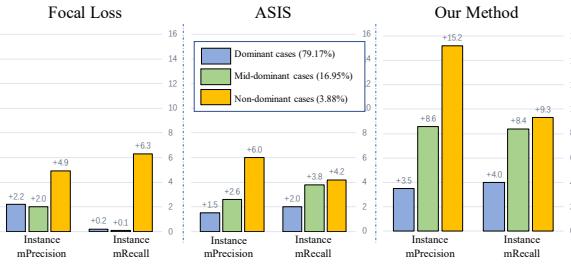


Figure 5 – The comparison of improvements between our proposed method and baseline model with focal loss [17] and ASIS [36]. Both mean precision and mean recall of instance are reported.

embeddings to have identical geometric output. As shown in Table 1, these regularizations can boost the *mPre* and *mRec* for about 1.7 and 1.8, respectively. **The Impact of Memory Size.** We study the influence of the memory size to the final performance. We set three values of N_c with 100, 150, 200 as the number of per-category prototypes. The *mPrec* on S3DIS Area-5 are 60.4, 62.7 and 62.5 respectively. The results show that the performance increases as N_c grows, and become stable after 200. In all our experiments, N_c is set to 150.

Visualization of the Memory Representation. Given the input features, the most relevant prototypes are retrieved to calibrate the features. In Figure 4, we visualize the embedding features with and without the memory module. Both common and rare scenes, *i.e.* office and lobby, are selected, according to the amount of training samples. The embeddings are projected to 1-D with the help pf Barnes-Hut t-SNE [33]. In both situations, our MPNet generate more discriminative embedding features, which is critical for separate different instance.

4.5. Comparison with the State-of-the-art

Performance on non-dominant cases. We first compare the performance of our proposed MPNet with state-of-the-art method ASIS [36] on non-dominant cases. We first sort

the 13 categories on S3DIS according to the total amount of training samples, and split the dataset into three levels: dominant cases (the first 4 classes), mid-dominant cases (the mid 5 classes) and non-dominant cases (the last 4 classes). The amount proportions of the three levels are 79.17%, 16.95% and 3.88%, respectively. As shown in Figure 5, we report the improvement with two metrics: *mPrec*, *mRec*. Our method can not only boost the performance on dominant cases, but surpass focal loss and ASIS [36] by a large margin on non-dominant cases.

Performance on S3DIS. We first compare the instance segmentation performance on both Area-5 and 6-fold. The results are presented in Table 2. Our proposed MPNet achieve promising results and surpass the previous state-of-the-art approaches substantially by a large margin. The large improvement is mainly beneficial from the strong ability of the proposed prototypes memory. Qualitative results is show in Figure 6. In addition to instance segmentation, we also report the results of semantic segmentation and compare it with other methods. The performance is tested on all areas (6-fold), as shown in Table 3. Although based on a simple PointNet++, we achieve even better quantitative results than other methods which are based on graph neural networks [15, 37].

Performance on ScanNetV2. In addition to S3DIS, we conduct experiments on ScanNetV2 [3]. The instance segmentation results are reported in Table 4, which is tested on the validation set. To make fair comparison, we select the methods that are based on PointNet or PointNet++. Our proposed MPNet outperforms previous methods over all overlap thresholds and dominant in many categories.

Speed Analysis. We compare the inference speed with other two methods: SGPN [35] and ASIS [36]. The whole evaluation process includes two parts: network forward and instance grouping. The first part is to get per-point semantic labeling and instance embedding. The second part utilizes a grouping algorithm to find out instance groups. SGPN, which is based on PointNet, predicts a pair-wise affinity matrix to group points into instance clusters. Due to the

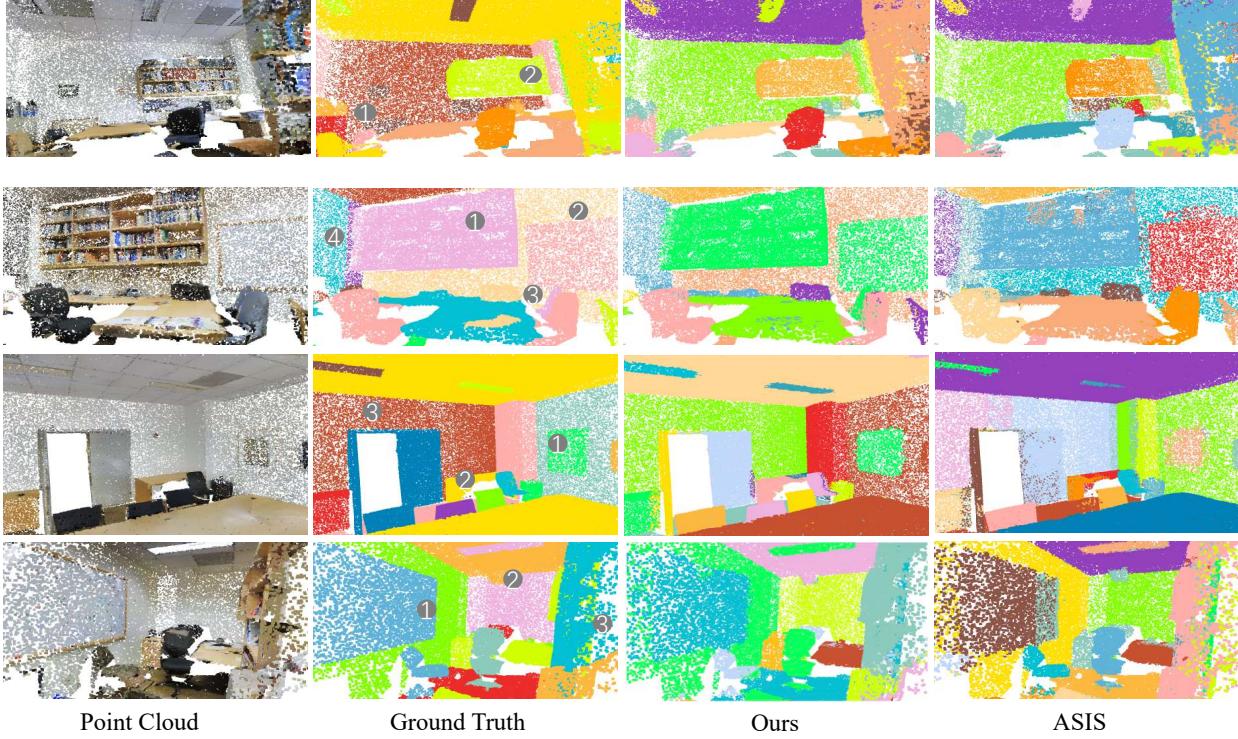


Figure 6 – Qualitative results of our method on S3DIS dataset. From left to right are: input point cloud, instance segmentation ground truth, the results of our method and the results of [36]. Note that different instance are shown with different colors, and the same instance are not necessarily have the same color in ground truth and prediction presentation.

Table 4 – Instance segmentation results on ScannetV2 benchmark (validation set). Both results of mAP@0.25 and mAP@0.5 are reported. All methods except [8] are based on PointNet or PointNet++ (3D-BEVIS [6] is multi-view based method).

Method	Year	mAP @0.25	mAP @0.5	bathtub	bed	shelf	cabinet	chair	counter	curtain	desk	door	other	picture	refrig	shCur	sink	sofa	table	toilet	window
MaskRCNN [12]	2017	26.1	5.8	33.3	0.2	0.0	5.3	0.2	0.2	2.1	0.0	4.5	2.4	23.8	6.5	0.0	1.4	10.7	2.0	11.0	0.6
SGPN [35]	2019	35.1	14.3	20.8	39.0	16.9	6.5	27.5	2.9	6.9	0.0	8.7	4.3	1.4	2.7	0.0	11.2	35.1	16.8	43.8	13.8
3D-BEVIS [6]	2019	-	24.8	66.7	56.6	7.6	3.5	39.4	2.7	3.5	9.8	9.9	3.0	2.5	9.8	37.5	12.6	60.4	18.1	85.4	17.1
R-PointNet [41]	2019	40.0	23.5	51.3	52.3	12.5	15.2	61.8	0.0	1.5	7.6	29.0	11.7	14.7	25.0	3.7	14.0	34.5	18.1	53.0	16.1
ASIS [36]	2019	41.5	24.0	29.9	50.5	0.0	16.7	57.7	0.0	18.4	7.8	14.8	12.9	1.8	12.4	38.0	10.2	36.9	37.4	71.7	14.5
Ours	-	49.3	31.0	69.4	59.8	2.7	23.7	71.1	4.5	8.4	18.3	11.6	17.3	4.8	21.8	57.0	13.4	27.7	41.8	87.3	18.3

Table 5 – Inferencing time comparison on S3DIS Area-5 set. Forward time is network running time on GPU, whereas Postprocessing time is the BlockMerging algorithm introduced in [35]. ASIS is 45% slower than our method in the forward process due to the usage of KNN, which is extremely time consuming. Reported time is running on a single 1080ti GPU with 4096 input points.

Method	Backbone	Inference Time (ms)			mPre	mRec
		Overall	Forward	Post		
SGPN[35]	PointNet	730	22	708	36.0	28.7
ASIS[36]	PointNet2	183	58	125	55.3	42.4
Ours	PointNet2	165	40	125	62.5	49.0

large size of input point cloud, a huge memory is required. Meanwhile, as lots of heuristic parameters are introduced,

the whole time for post-processing is much slower than our proposed method. Different from SGPN, ASIS utilize mean-shift for clustering embeddings to instance groups. Meanwhile, ASIS applies KNN for fusing semantic context from a fixed number of neighboring points, which is used on every input point. This operation is extremely time-consuming and fail to take fully advantage of computational resources. Compared with the above two approaches, our proposed MPNet is free from complex and time-consuming operations, showing the superiority in both effectiveness and efficiency.

5. Conclusion

In this paper, we propose a memory-augmented network to handle both category and pattern imbalance in point cloud instance segmentation. A memory module is introduced to alleviate the forgetting issue during the training process. The performance on the benchmarks shows the superiority of our method in both effectiveness and efficiency.

6. Appendix

In this supplementary material, we provide more detailed experimental results, including:

- Both qualitative and quantitative results on the “Chair” category in PartNet [21];
- More visualization of our approach on S3DIS [1] and ScanNetV2 [3].

6.1. Experimental Results on PartNet dataset [21]

In Figure 3 in the main paper, to better understand the learned memory prototypes, we do visualization relying on the category of “Chair” in PartNet [21]. PartNet [21] is a consistent dataset of 3D objects with fine-grained and hierarchical 3D part annotations. In this section, we report the quantitative results in Table 6. **Level-1** refers to the coarsest annotation and **Level-3** refers to the most fine-grained annotation as defined in [21]. For fair comparison, all results are evaluated with the same backbone PointNet++ [25]. Our method outperforms the previous methods by a large margin, showing the flexibility of our method to handle various types of input data. Moreover, visualization examples of the results are shown in Figure 7, indicating that our method can handle both rare and common cases well.

Table 6 – Comparison of the per-level performance of our method with the state-of-the-art methods on “Chair” category in PartNet [21]. The performance is evaluated using part-category mAP, with IoU threshold of 0.5. All the results are achieved with the same backbone: PointNet++ [25].

Method	Year	Level-1	Level-2	Level-3
SGPN [35]	2019	72.4	25.4	19.4
PartNet [21]	2019	74.4	35.5	29.0
GSPN [41]	2019	-	-	26.8
Ours	-	79.9	41.2	32.5

6.2. More Visualization Results

In the main paper, we illustrate the quantitative results on S3DIS [1] and ScanNetV2 [3] datasets in Table 2 and 4, respectively. Visualization examples of both semantic

and instance segmentation results on S3DIS and ScanNetV2 datasets are shown in Figure 8 in the following.

References

- [1] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016. [1](#), [2](#), [5](#), [9](#), [10](#)
- [2] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic Instance Segmentation with a Discriminative Loss Function . In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. [1](#), [2](#), [5](#)
- [3] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. [1](#), [5](#), [7](#), [9](#), [10](#)
- [4] Angela Dai and Matthias Nießner. 3DMV: Joint 3D-Multi-View Prediction for 3D Semantic Scene Segmentation. In *Proc. Eur. Conf. Comp. Vis.*, 2018. [2](#)
- [5] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic Segmentation via Multi-task Network Cascades . In *Proc. Eur. Conf. Comp. Vis.*, 2016. [1](#)
- [6] Cathrin Elich, Francis Engelmann, Theodora Kontogianni, and Bastian Leibe. 3D-BEVIS: Bird’s-Eye-View Instance Segmentation . *arXiv preprint arXiv:1904.02199*, 2019. [8](#)
- [7] Francis Engelmann, Theodora Kontogianni, Alexander Hermans, and B. Leibe. Exploring Spatial Context for 3D Semantic Segmentation of Point Clouds . In *Proc. IEEE Int. Conf. Comp. Vis. Workshops*, 2017. [7](#)
- [8] Francis Engelmann, Theodora Kontogianni, Jonas Schult, and Bastian Leibe. Know What Your Neighbors Do: 3D Semantic Segmentation of Point Clouds . *arXiv:1810.01151*, 2018. [7](#), [8](#)
- [9] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing Normality to Detect Anomaly: Memory-augmented Deep Autoencoder for Unsupervised Anomaly Detection. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2019. [3](#)
- [10] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018. [2](#)
- [11] Alex Graves, Greg Wayne, and Ivo Danihelk. Neural Turing Machines . *arXiv preprint arXiv:1410.5401*, 2014. [3](#)
- [12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN . In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017. [1](#), [8](#)
- [13] Ji Hou, Angela Dai, and Matthias Nießner. 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019. [1](#), [2](#)
- [14] Jean Lahoud, Bernard Ghanem, Marc Pollefeys, and Martin R. Oswald. 3D Instance Segmentation via Multi-task Metric Learning. *arXiv preprint arXiv:1906.08650*, 2019. [1](#)

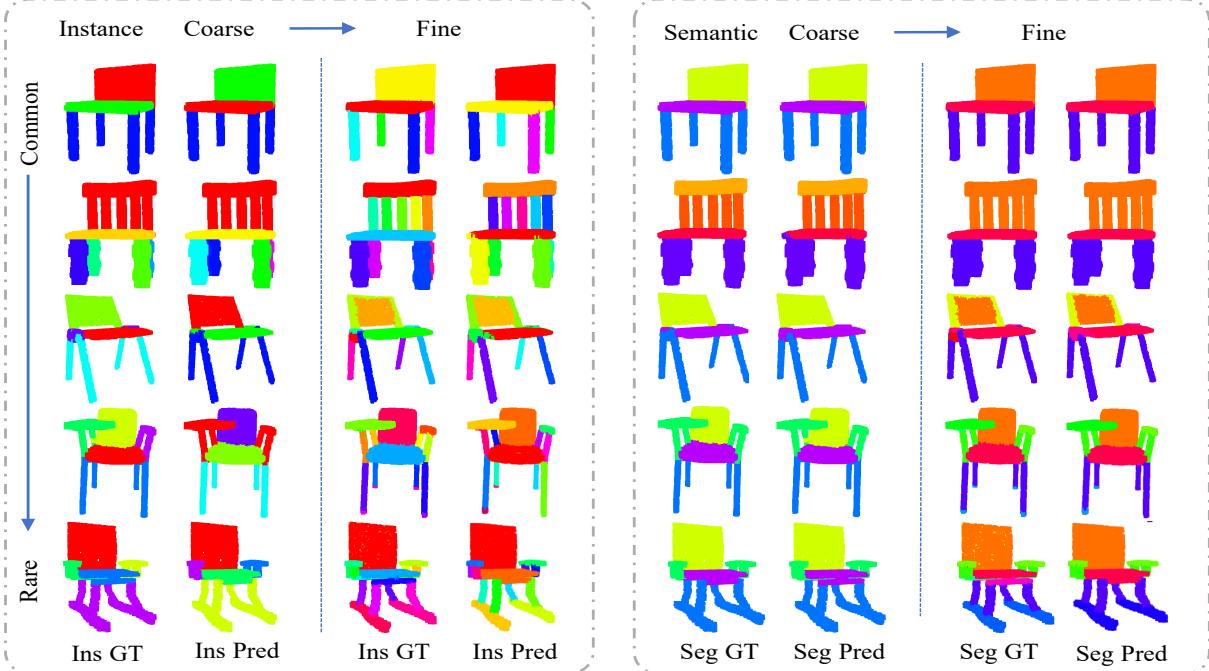


Figure 7 – Visualization of the performance of on PartNet [21]. Both coarse and fine-grained results are provided. Note that different instance are shown with different colors, and the same instance are not necessarily have the same color in ground truth and prediction presentation.

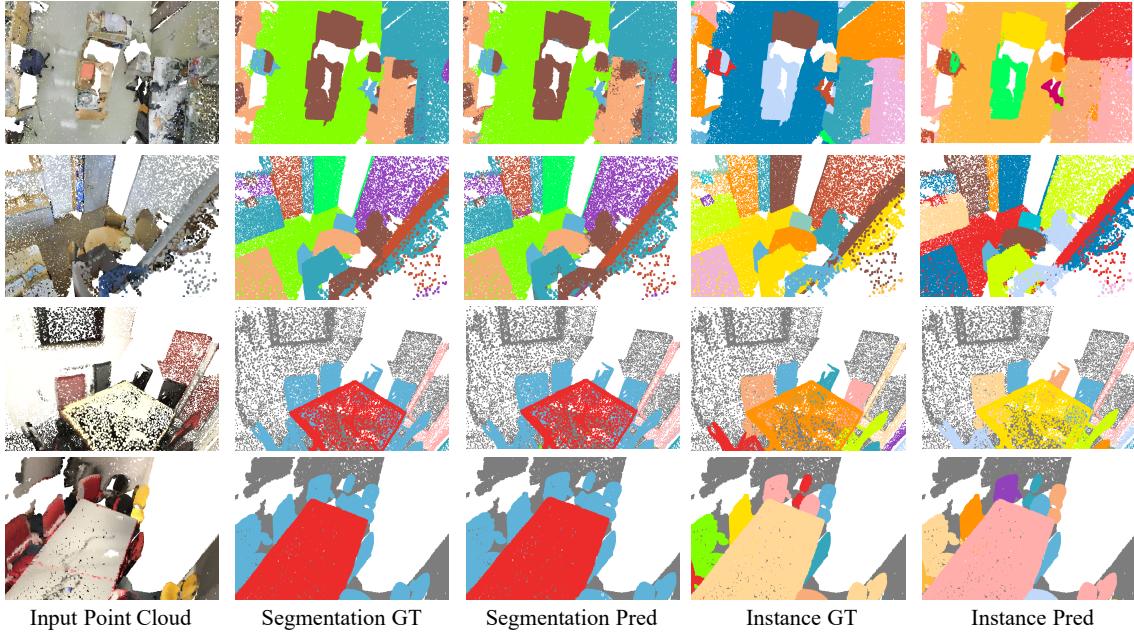


Figure 8 – Visualization of the performance of on S3DIS [1] and ScanNetV2 [3]. Both instance and semantic segmentation results are provided. Note that different instance are shown with different colors, and the same instance are not necessarily have the same color in ground truth and prediction presentation.

- [15] Guohao Li, Matthias Mller, Ali Thabet, and Bernard Ghanem. DeepGCNs: Can GCNs Go as Deep as CNNs? In *Proc. IEEE Int. Conf. Comp. Vis.*, 2019. [2](#), [7](#)
- [16] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhua Di, and Baoquan Chen. PointCNN: Convolution On X-

- Transformed Points. In *Proc. Advances in Neural Inf. Process. Syst.*, 2018. [2](#)
- [17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017. [2](#), [6](#), [7](#)

- [18] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-Scale Long-Tailed Recognition in an Open World. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019. [3](#), [5](#)
- [19] Daniel Maturana and Sebastian Scherer. VoxNet: A 3D Convolutional Neural Network for real-time object recognition. In *Proc. IEEE Int. Conf. Intelligent Robots Syst.*, 2015. [1](#), [2](#)
- [20] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. [1](#)
- [21] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A Large-scale Benchmark for Fine-grained and Hierarchical Part-level 3D Object Understanding. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019. [4](#), [6](#), [9](#), [10](#)
- [22] Quang-Hieu Pham, Duc Thanh Nguyen, Binh-Son Hua, Gemma Roig, and Sai-Kit Yeung. JSIS3D: Joint semantic-instance segmentation of 3d point clouds with multi-task pointwise networks and multi-value conditional random fields. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019. [1](#), [6](#)
- [23] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. [1](#), [2](#), [5](#), [7](#)
- [24] Charles R. Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J. Guibas. Volumetric and Multi-View CNNs for Object Classification on 3D Data . In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016. [2](#)
- [25] Charles R Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Proc. Advances in Neural Inf. Process. Syst.*, 2017. [1](#), [2](#), [3](#), [5](#), [7](#), [9](#)
- [26] Charles R. Qi, Litany Or, Kaiming He, and Leonidas J. Guibas. Deep Hough Voting for 3D Object Detection in Point Clouds. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2019. [2](#)
- [27] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. OctNet: Learning Deep 3D Representations at High Resolutions . *arXiv preprint arXiv:1611.05009*, 2016. [2](#)
- [28] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic Routing Between Capsules . In *Proc. Advances in Neural Inf. Process. Syst.* 2017. [5](#)
- [29] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical Networks for Few-shot Learning . In *Proc. Advances in Neural Inf. Process. Syst.* 2017. [3](#)
- [30] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view Convolutional Neural Networks for 3D Shape Recognition. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2015. [2](#)
- [31] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. KPConv: Flexible and Deformable Convolution for Point Clouds. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2019. [2](#)
- [32] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geofrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018. [1](#)
- [33] Laurens van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.*, 15:3221–3245, 2014. [6](#), [7](#)
- [34] Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, and Jie Shan. Graph Attention Convolution for Point Cloud Semantic Segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019. [2](#)
- [35] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. SGPN: Similarity Group Proposal Network for 3D Point Cloud Instance Segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018. [1](#), [2](#), [5](#), [6](#), [7](#), [8](#), [9](#)
- [36] Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, and Jiaya Jia. Associatively Segmenting Instances and Semantics in Point Clouds. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019. [1](#), [2](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [37] 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 Trans. On. Graphic*, 2019. [7](#)
- [38] Wenxuan Wu, Zhongang Qi, and Li Fuxin. PointConv: Deep Convolutional Networks on 3D Point Clouds. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019. [2](#)
- [39] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Lin-guang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A Deep Representation for Volumetric Shapes . In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015. [1](#), [2](#)
- [40] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning Object Bounding Boxes for 3D Instance Segmentation on Point Clouds. In *Proc. Advances in Neural Inf. Process. Syst.*, 2019. [1](#), [2](#), [6](#)
- [41] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J. Guibas. GSPN: Generative Shape Proposal Network for 3d Instance Segmentation in Point Cloud. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018. [1](#), [8](#), [9](#)