

SoftGroup for 3D Instance Segmentation on Point Clouds

Thang Vu Kookhoi Kim Tung M. Luu Thanh Nguyen Chang D. Yoo
Korea Advanced Institute of Science and Technology (KAIST)
 {thangvubk, rlarnrghlapz, tungluu2203, thanhnguyen, cd-yoo}@kaist.ac.kr

Abstract

Existing state-of-the-art 3D instance segmentation methods perform semantic segmentation followed by grouping. The hard predictions are made when performing semantic segmentation such that each point is associated with a single class. However, the errors stemming from hard decision propagate into grouping that results in (1) low overlaps between the predicted instance with the ground truth and (2) substantial false positives. To address the aforementioned problems, this paper proposes a 3D instance segmentation method referred to as SoftGroup by performing bottom-up soft grouping followed by top-down refinement. SoftGroup allows each point to be associated with multiple classes to mitigate the problems stemming from semantic prediction errors and suppresses false positive instances by learning to categorize them as background. Experimental results on different datasets and multiple evaluation metrics demonstrate the efficacy of SoftGroup. Its performance surpasses the strongest prior method by a significant margin of +6.2% on the ScanNet v2 hidden test set and +6.8% on S3DIS Area 5 in terms of AP_{50} . SoftGroup is also fast, running at 345ms per scan with a single Titan X on ScanNet v2 dataset. The source code and trained models for both datasets are available at <https://github.com/thangvubk/SoftGroup.git>.

1. Introduction

Scene understanding on 3D data has received increasing attention for the rapid development of 3D sensors and availability of large-scale 3D datasets. Instance segmentation on point clouds is a 3D perception task, serving as the foundation for a wide range of applications such as autonomous driving, virtual reality, and robot navigation. Instance segmentation processes the point clouds to output a category and an instance mask for each detected object.

State-of-the-art methods [4, 15, 20] consider 3D instance segmentation as a bottom-up pipeline. They learn the point-wise semantic labels and center offset vectors and then group points of the same labels with small geomet-

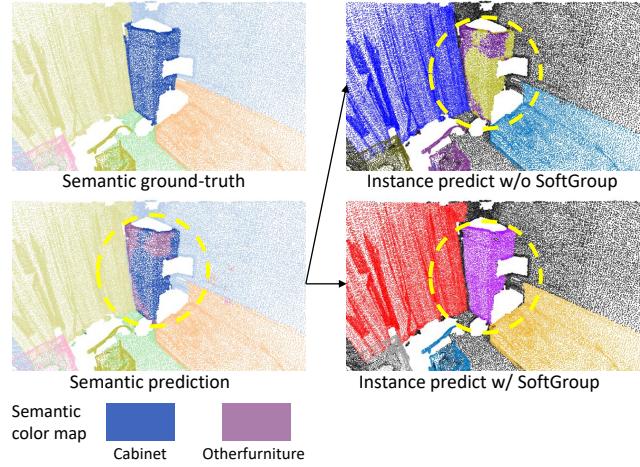


Figure 1. Instance segmentation with and without SoftGroup from the same semantic prediction results. The last row shows the palette for semantic predictions only. Instance predictions are illustrated by different random colors for different objects. In the semantic prediction results, some regions of cabinet are wrongly predicted as other furniture. Without SoftGroup, these errors are propagated to instance prediction. SoftGroup addresses this problem and produces more accurate instance masks.

ric distances into instances. These grouping algorithms are performed on the hard semantic prediction, where a point is associated with a single class. In many cases, objects are locally ambiguous, the output semantic predictions show different categories for different parts, and thus using hard semantic predictions for instance grouping leads to two problems: (1) low overlap between predicted instance and the ground-truth and (2) extra false-positive instances from wrong semantic regions. Figure 1 shows a visualization example. Here, in the semantic prediction results, some parts of cabinet is wrongly predicted as other furniture. When hard semantic predictions are used to perform grouping, the semantic prediction error is propagated to instance prediction. As a result, the predicted cabinet instance has low overlap with the ground truth, and the other furniture instance is a false positive.

This paper proposes SoftGroup to address these prob-

lems by considering soft semantic scores to perform grouping instead of hard one-hot semantic predictions. The intuition of SoftGroup is illustrated in Figure 2. Our finding is that the object parts with wrong semantic predictions still have reasonable scores for the true semantic class. SoftGroup relies on a score threshold to determine which category the object belongs instead of the argument max values. Grouping on the soft semantic scores produces for accurate instance on true semantic class. The instance with wrong semantic prediction will be suppressed by learning to categorize it as background. To this end, we treat an instance proposal as either a positive or negative sample depending on the maximum Intersection over Union (IoU) with the ground truth, then construct a top-down refinement stage to refine the positive sample and suppress the negative one. As shown in Figure 1, SoftGroup is able to produce accurate instance masks from imperfect semantic prediction.

SoftGroup is conceptually simple and easy to implement. Experiments on the ScanNet v2 [6] and S3DIS [1] benchmark datasets show the efficacy of our method. Notably, SoftGroup outperforms the previous state-of-the-art method by a significant margin of +6.2% on the ScanNet hidden test set and +6.8% on S3DIS Area 5 in terms of AP₅₀. SoftGroup is fast, requiring 345ms to process a ScanNet scene. In summary, our contribution is threefold.

- We propose SoftGroup that performs grouping on soft semantic scores to avoid error propagation from hard semantic predictions to instance segmentation.
- We propose a top-down refinement stage to correct, refine the positive samples and suppress false positives introduced by wrong semantic predictions.
- We report extensive experiments on multiple datasets with different evaluation metrics, showing significant improvements over existing state-of-the-art methods.

2. Related work

Deep Learning on 3D Point Clouds. Point cloud representation is a common data format for 3D scene understanding. To process point clouds, early methods [2, 3, 36, 37] extract hand-crafted features based on statistical properties of points. Recent deep learning methods learn to extract features from points. PointNet-based methods [32, 33] propose to process points through shared Multi-Layer Perceptron (MLP) and then aggregate regional and global features from symmetric function, such as max-pooling. Convolution methods are actively explored for point clouds processing. Continuous convolution methods [23, 40, 44, 45] learn the kernels which are associated to the spatial distribution of local points. Discrete convolution methods [5, 8, 13, 19, 25, 34] learn the kernels which are regular grids obtaining from point quantization. Transformers [18, 50] and graph-based methods [38, 39, 43] are also proposed to address the data irregularity of point clouds.

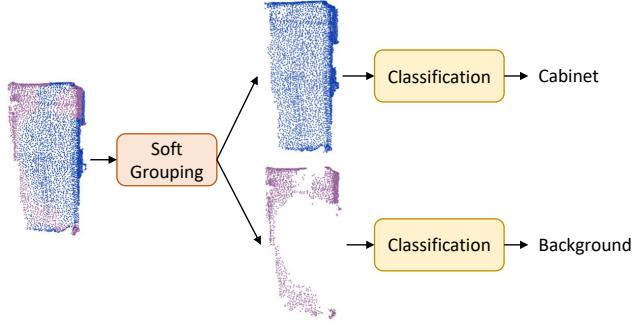


Figure 2. The cabinet in Figure 1 is extracted to illustrate the high-level pipeline of our method. The soft grouping module based on soft semantic scores to output more accurate instance (the upper one). The classifier processes each instance and suppress the instance from wrong semantic prediction (the lower one).

Proposal-based Instance Segmentation. Proposal-based methods consider a top-down strategy that generates region proposals and then segments the object within each proposal. Existing proposal-based methods for 3D point clouds are highly influenced by the success of Mask-R CNN for 2D images. To handle data irregularity of point clouds, Li *et al.* [47] propose GSPN, which takes an analysis-by-synthesis strategy to generate high-objectness 3D proposals, which are refined by a region-based PointNet. Hou *et al.* [12] present 3DSIS that combines multi-view RGB input with 3D geometry to predict bounding boxes and instance masks. Yang *et al.* [46] propose 3D-BoNet which directly outputs a set of bounding boxes without anchor generation and non-maximum suppression, then segments the object by a pointwise binary classifier. Liu *et al.* [22] present GICN to approximate the instance center of each object as a Gaussian distribution, which is sampled to get object candidates then produce corresponding bounding boxes and instance masks.

Grouping-based Instance Segmentation. Grouping-based methods rely on a bottom-up pipeline that produces per-point predictions (such as semantic maps, and geometric shifts, or latent features) then groups points into instances. Wang *et al.* [41] propose SGPN to construct a feature similarity matrix for all points and then group points of similar features into instances. Pham *et al.* [29] present JSIS3D that incorporates the semantic and instance labels by a multi-value conditional random field model and jointly optimizes the labels to obtain object instances. Lahoud *et al.* [17] propose MTML to learn feature and directional embedding, then perform mean-shift clustering on the feature embedding to generate object segments which are scored according to their direction feature consistency. Han *et al.* [9] introduce OccuSeg that performs graph-based

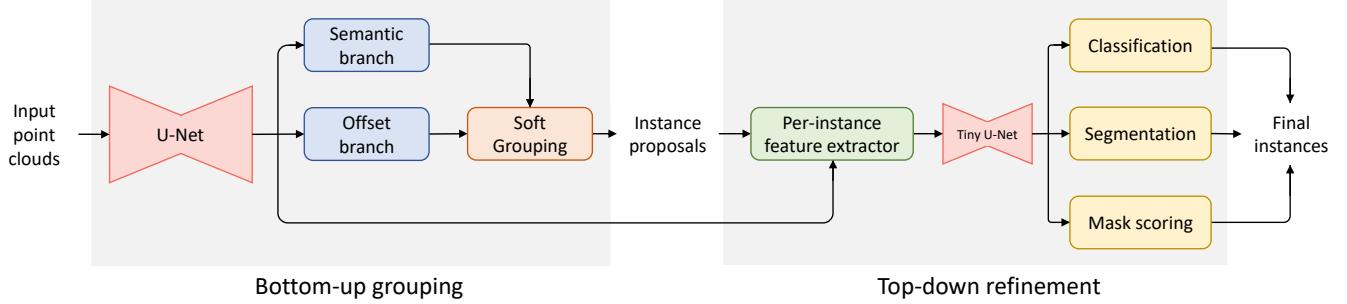


Figure 3. The architecture of the proposed method consists of bottom-up grouping and top-down refinement stages. From the input point clouds, the U-Net backbone extracts the point features. Then semantic and offset branches predict the semantic scores and offset vectors, followed by a soft grouping module to generate instance proposal. The feature extractor layer extracts backbone features from instance proposals. The features for each proposal are fed into a tiny U-Net followed by the classification, segmentation, and mask scoring branches to get the final instances.

clustering guided by object occupancy signal for more accurate segmentation outputs. Zhang *et al.* [48] consider a probabilistic approach that represents each point as a tri-variate normal distribution followed by a clustering step to obtain object instances. Jiang *et al.* [15] propose PointGroup to segment objects on original and offset-shifted point sets, relying on a simple yet effective algorithm that groups nearby points of the same label and expands the group progressively. Chen *et al.* [4] extend PointGroup and propose HAIS that further absorbs surrounding fragments of instances and then refines the instances based on intra-instance prediction. Liang *et al.* [20] SSTNet to construct a tree network from pre-computed superpoints then traverse the tree and split nodes to get object instances.

The common proposal-based and grouping-based methods have their advantages and drawbacks. Proposal-based methods process each object proposal independently that is not interfered with by other instances. Grouping-based methods process the whole scene without proposal generation, enabling fast inference. However, proposal-based methods have difficulties in generating high-quality proposals since the point only exists on the object surface. Grouping-based methods highly depend on semantic segmentation such that the errors in semantic predictions are propagated to instance predictions. The proposed method leverages the advantages and address the limitations of both approaches. Our method is constructed as a two-stage pipeline, where the bottom-up stage generates high-quality object proposals by grouping on soft semantic scores, and then the top-down stage process each proposal to refine positive samples and suppress negative ones.

3. Method

The overall architecture of SoftGroup is depicted in Figure 3, which is divided into two stages. In the bottom-up grouping stage, the point-wise prediction network (Sec.

3.1) takes point clouds the input and produces point-wise semantic labels and offset vectors. The soft grouping module (Sec. 3.2) processes these outputs to produce preliminary instance proposals. In the top-down refinement stage, based on the proposals, the corresponding features from the backbone are extracted and used to predict classes, instance masks, and mask scores as the final results.

3.1. Point-wise Prediction Network

The input of the point-wise prediction network is a set of N points, each of which is represented by its coordinate and color. The point set is voxelized to convert unordered points to ordered volumetric grids, which are fed into a U-Net style backbone [35] to obtain point features. The Submanifold Sparse Convolution [8] is adopted to implement the U-Net for 3D point clouds. From the point features, two branches are constructed to output the point-wise semantic scores and offset vectors.

Semantic Branch. A semantic branch is constructed from a two-layer MLP and learns to output semantic scores $S = \{s_1, \dots, s_N\} \in \mathbb{R}^{N \times N_{\text{class}}}$ for N points over N_{class} classes. Different from existing methods [4, 15], we directly perform grouping on semantic scores without converting the semantic scores to one-hot semantic predictions.

Offset Branch. In parallel with the semantic branch, we apply a two-layer MLP to learn the offset vectors $O = \{o_1, \dots, o_N\} \in \mathbb{R}^{N \times 3}$, which represents the vector from each point to the geometric center of the instance the point belongs. Based on the learned offset vectors, we shift the points to the center of the corresponding instance to perform grouping more effectively.

The cross-entropy loss and ℓ_1 regression loss are used to train the semantic and offset branches, respectively.

$$L_{\text{semantic}} = \frac{1}{N} \sum_{i=1}^N \text{CE}(s_i, s_i^*), \quad (1)$$

$$L_{\text{offset}} = \frac{1}{\sum_{i=1}^N \mathbb{1}_{\{p_i\}}} \sum_{i=1}^N \mathbb{1}_{\{p_i\}} \|\mathbf{o}_i - \mathbf{o}_i^*\|_1, \quad (2)$$

where s^* is the semantic label, \mathbf{o}^* is offset label representing the vector from a point to the geometric center of the instance that the point belongs to (analogous to [4, 15, 20]), and $\mathbb{1}_{\{p_i\}}$ is the indicator function indicating whether the point p_i belongs to any instance.

3.2. Soft Grouping

The soft grouping module receives the semantic scores and offset vectors as the input and produces instance proposals. First, the offset vectors are used to shift points toward the corresponding instance centers. To perform grouping using the semantic scores, we define a score threshold τ to determine which semantic classes a point belongs to, allowing the point to be associated with multiple classes. Given semantic scores $S \in \mathbb{R}^{N \times N_{\text{class}}}$, we iterate through N_{class} classes, and at each class index we slice a point subset of the whole scene that has the score (w.r.t. the class index) higher than the threshold τ . We follow [4, 15] to perform grouping on each point subset. Since all points in each subset belong to the same class, we simply traverse all the points in the subset and create the links between points having a geometric distance smaller than a grouping bandwidth b to get the instance proposals. For each iteration, the grouping is performed on a point subset of the whole scan, ensuring fast inference. The overall instance proposals are the union of the proposals from all subsets.

We note that existing proposal-based methods [12, 22, 46] commonly consider bounding boxes as object proposals then perform segmentation within each proposal. Intuitively, the bounding box with high overlap with the instance should have the center close to the object center. However, generating high-quality bounding box proposals in 3D point clouds is challenging since the point only exists on object surfaces. Instead, SoftGroup relies on point-level proposals which are more accurate and naturally inherit the scattered property of point clouds.

Since the quality of instance proposals from grouping highly depend on the quality of semantic segmentation, we quantitatively analyze the impact of τ on the recall and precision of semantic predictions. The recall and precision for class j is defined as follows.

$$\begin{aligned} \text{recall}_j &= \sum_{i=1}^N \frac{(s_{ij} > \tau) \wedge (s_i^* = j)}{s_i^* = j}, \\ \text{precision}_j &= \sum_{i=1}^N \frac{(s_{ij} > \tau) \wedge (s_i^* = j)}{s_{ij} > \tau}. \end{aligned} \quad (3)$$

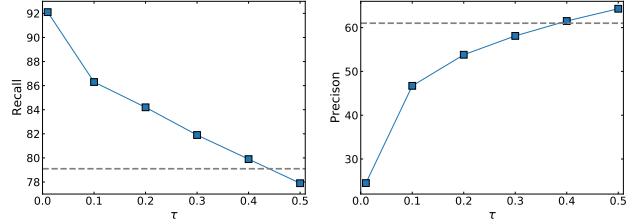


Figure 4. The recall and precision of semantic prediction with varying score threshold τ . The dashed lines denote the recall and precision with hard semantic prediction.

Figure 4 shows the recall and precision (averaged over classes) with the varying score thresholds τ compared with those of hard semantic prediction. With hard semantic prediction, the recall is 79.1%, indicating more than 20% amount of points over classes are not covered by the predictions. When using the score threshold, the recall increases as the score threshold decreases. However, the small score threshold also leads to low precision. We propose a top-down refinement stage mitigate the low precision problems. The precision can be interpreted as the relation between foreground and background points of object instances. We set the threshold to 0.2 with precision near 50%, leading to the ratio between foreground and background points for ensuring stage is balanced.

3.3. Top-Down Refinement

The top-down refinement stage classifies and refines the instance proposals from the bottom-up grouping stage. A feature extractor layer processes each proposal to extract its corresponding backbone features. The extracted features are fed into a tiny U-Net network (a U-Net style network with a small number of layers) before predicting classification scores, instance masks, and mask scores at the ensuing branches.

Classification Branch. The classification branch starts with a global average pooling layer to aggregate the feature of all points in the instance, followed by a MLP to predict the classification scores $C = \{c_1, \dots, c_K\} \in \mathbb{R}^{K \times (N_{\text{class}}+1)}$, where K is the number of instances. We directly derive the object category and classification confidence score from the output of the classification branch.

We note that existing grouping-based methods typically derive the object category from semantic predictions. However, instances may come from objects with noisy semantic predictions. The proposed method directly uses the output of the classification branch as the instance class. The classification branch aggregates all point features of the instance and classifies the instance with a single label, leading to more reliable predictions.

Segmentation Branch. As shown in Section 3.2, the instance proposals contain both foreground and background points, we construct a segmentation branch to predict an instance mask within each proposal. The segmentation branch is a point-wise MLP of two layers that output an instance mask \mathbf{m}_k for each instance k .

Mask Scoring Branch. The mask scoring branch shares the same structure as the classification branch. This branch outputs the mask scores $\mathbf{E} = \{e_1, \dots, e_K\} \in \mathbb{R}^{K \times N_{\text{class}}}$, which estimate the IoU of a predicted mask with the ground truth. The mask score is combined with the classification score by multiplication to get the final confidence score.

Learning Targets. Training the top-down refinement branches requires the target labels for each branch. To this end, we follow the logic in existing 2D object detection and segmentation methods. We treat all instance proposals having IoU with a ground-truth instance higher than 50% as the positive samples and the rest as negatives. Every positive sample is assigned to a ground-truth instance with the highest IoU. The classification target of a positive sample is the category of the corresponding ground-truth instance. The total number of classes is $N_{\text{class}} + 1$ (N_{class} foreground classes and one background class). The segmentation and mask scoring branches are trained on positive samples only. The mask target of a positive sample is the mask of the assigned ground-truth instance. The mask score target is the IoU between the predicted mask and the ground truth. The training loss of these branches is the combination of cross-entropy, binary cross-entropy, and ℓ_2 regression losses, following [10, 14].

$$L_{\text{class}} = \frac{1}{K} \sum_{k=1}^K \text{CE}(\mathbf{c}_k, \mathbf{c}_k^*), \quad (4)$$

$$L_{\text{mask}} = \frac{1}{\sum_{k=1}^K \mathbb{1}_{\{\mathbf{m}_k\}}} \sum_{k=1}^K \mathbb{1}_{\{\mathbf{m}_k\}} \text{BCE}(\mathbf{m}_k, \mathbf{m}_k^*), \quad (5)$$

$$L_{\text{mask_score}} = \frac{1}{\sum_{k=1}^K \mathbb{1}_{\{\mathbf{e}_k\}}} \sum_{k=1}^K \mathbb{1}_{\{\mathbf{e}_k\}} \|\mathbf{e}_k - \mathbf{e}_k^*\|_2. \quad (6)$$

Here, \mathbf{c}^* , \mathbf{m}^* , \mathbf{e}^* are the classification, segmentation, and mask scoring targets, respectively. K is the total number of proposals and $\mathbb{1}_{\{\cdot\}}$ indicates whether the proposal is a positive sample.

3.4. Multi-task Learning

The whole network can be trained in an end-to-end manner using a multi-task loss.

$$L = L_{\text{semantic}} + L_{\text{offset}} + L_{\text{class}} + L_{\text{mask}} + L_{\text{mask_score}}, \quad (7)$$

where L_{semantic} and L_{offset} are the semantic and offset losses defined at subsection Section 3.1 while L_{class} , L_{mask} and $L_{\text{mask_score}}$ are the classification, segmentation and mask score losses defined at Section 3.3.

4. Experiments

4.1. Experimental Settings

Datasets. The experiments are conducted on standard benchmarked ScanNet v2 [6] and S3DIS [1] dataset. The ScanNet dataset contains 1613 scans which is divided into training, validation, and testing sets of 1201, 312, 100 scans, respectively. Instance segmentation is evaluated on 18 object classes. Following existing methods, the benchmarked results are reported on the hidden test split. The ablation study is conducted on the validation set.

The S3DIS dataset contains 3D scans of 6 areas with 271 scenes in total. The dataset consists of 13 classes for instance segmentation evaluation. Following existing methods, two settings are used to evaluate the instance segmentation results: testing on Area 5 and 6-fold cross-validation.

Evaluation Metrics. The evaluation metric is the standard average precision. Here, AP_{50} and AP_{25} denote the scores with IoU thresholds of 50% and 25%, respectively. Likewise, AP denotes the averaged scores with IoU threshold from 50% to 95% with a step size of 5%. Additionally, the S3DIS is also evaluated using mean coverage (mCov), mean weighed coverage (mWCov), mean precision (mPrec), and mean recall (mRec).

Implementation Details. The implementation details follow those of existing methods [4, 15]. The model is implemented using PyTorch deep learning framework [28] and trained on 120k iterations with Adam optimizer [16]. The batch size is set to 4. The learning rate is initialized to 0.001 and scheduled by a cosine annealing [24]. The voxel size and grouping bandwidth b are set to 0.02m and 0.04m, respectively. The score threshold for soft grouping τ is set to 0.2. At training time, the scenes are randomly cropped at a maximum number of points of 250k. At inference, the whole scene is fed into the network without cropping. For the S3DIS with high point density, scenes are randomly downsampled at a ratio of 1/4 before cropping. At inference, the scene is divided into four parts before feeding into the model, and then the features of these parts are merged right after the U-Net backbone. The ensuing components process the merged features as those on ScanNet dataset.

We note that the source code and trained models for existing high-performing methods are publicly available on ScanNet v2 only. In this work, the source code and trained models on both ScanNet v2 and S3DIS will be released to support result reproducibility.

| Method | AP ₅₀ | bathtub | bed | bookshe. | cabinet | chair | counter | curtain | desk | door | other | picture | fridge | s. curtain | sink | sofa | table | toilet | window |
|-------------------------|------------------|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|-------------|-------------|-------------|--------------|-------------|
| SGPN [41] | 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 |
| GSPN [47] | 30.6 | 50.0 | 40.5 | 31.1 | 34.8 | 58.9 | 5.4 | 6.8 | 12.6 | 28.3 | 29.0 | 2.8 | 21.9 | 21.4 | 33.1 | 39.6 | 27.5 | 82.1 | 24.5 |
| 3D-SIS [12] | 38.2 | 100.0 | 43.2 | 24.5 | 19.0 | 57.7 | 1.3 | 26.3 | 3.3 | 32.0 | 24.0 | 7.5 | 42.2 | 85.7 | 11.7 | 69.9 | 27.1 | 88.3 | 23.5 |
| MASC [21] | 44.7 | 52.8 | 55.5 | 38.1 | 38.2 | 63.3 | 0.2 | 50.9 | 26.0 | 36.1 | 43.2 | 32.7 | 45.1 | 57.1 | 36.7 | 63.9 | 38.6 | 98.0 | 27.6 |
| PanopticFusion [27] | 47.8 | 66.7 | 71.2 | 59.5 | 25.9 | 55.0 | 0.0 | 61.3 | 17.5 | 25.0 | 43.4 | 43.7 | 41.1 | 85.7 | 48.5 | 59.1 | 26.7 | 94.4 | 35.9 |
| 3D-Bonet [46] | 48.8 | 100.0 | 67.2 | 59.0 | 30.1 | 48.4 | 9.8 | 62.0 | 30.6 | 34.1 | 25.9 | 12.5 | 43.4 | 79.6 | 40.2 | 49.9 | 51.3 | 90.9 | 43.9 |
| MTML [17] | 54.9 | 100.0 | 80.7 | 58.8 | 32.7 | 64.7 | 0.4 | 81.5 | 18.0 | 41.8 | 36.4 | 18.2 | 44.5 | 100.0 | 44.2 | 68.8 | 57.1 | 100.0 | 39.6 |
| 3D-MPA [7] | 61.1 | 100.0 | 83.3 | 76.5 | 52.6 | 75.6 | 13.6 | 58.8 | 47.0 | 43.8 | 43.2 | 35.8 | 65.0 | 85.7 | 42.9 | 76.5 | 55.7 | 100.0 | 43.0 |
| Dyco3D [11] | 64.1 | 100.0 | 84.1 | 89.3 | 53.1 | 80.2 | 11.5 | 58.8 | 44.8 | 43.8 | 53.7 | 43.0 | 55.0 | 85.7 | 53.4 | 76.4 | 65.7 | 98.7 | 56.8 |
| PE [49] | 64.5 | 100.0 | 77.3 | 79.8 | 53.8 | 78.6 | 8.8 | 79.9 | 35.0 | 43.5 | 54.7 | 54.5 | 64.6 | 93.3 | 56.2 | 76.1 | 55.6 | 99.7 | 50.1 |
| PointGroup [15] | 63.6 | 100.0 | 76.5 | 62.4 | 50.5 | 79.7 | 11.6 | 69.6 | 38.4 | 44.1 | 55.9 | 47.6 | 59.6 | 100.0 | 66.6 | 75.6 | 55.6 | 99.7 | 51.3 |
| GICN [22] | 63.8 | 100.0 | 89.5 | 80.0 | 48.0 | 67.6 | 14.4 | 73.7 | 35.4 | 44.7 | 40.0 | 36.5 | 70.0 | 100.0 | 56.9 | 83.6 | 59.9 | 100.0 | 47.3 |
| OccuSeg [9] | 67.2 | 100.0 | 75.8 | 68.2 | 57.6 | 84.2 | 47.7 | 50.4 | 52.4 | 56.7 | 58.5 | 45.1 | 55.7 | 100.0 | 75.1 | 79.7 | 56.3 | 100.0 | 46.7 |
| SSTNet [20] | 69.8 | 100.0 | 69.7 | 88.8 | 55.6 | 80.3 | 38.7 | 62.6 | 41.7 | 55.6 | 58.5 | 70.2 | 60.0 | 100.0 | 82.4 | 72.0 | 69.2 | 100.0 | 50.9 |
| HAIS [4] | 69.9 | 100.0 | 84.9 | 82.0 | 67.5 | 80.8 | 27.9 | 75.7 | 46.5 | 51.7 | 59.6 | 55.9 | 60.0 | 100.0 | 65.4 | 76.7 | 67.6 | 99.4 | 56.0 |
| SoftGroup (Ours) | 76.1 | 100.0 | 80.8 | 84.5 | 71.6 | 86.2 | 24.3 | 82.4 | 65.5 | 62.0 | 73.4 | 69.9 | 79.1 | 98.1 | 71.6 | 84.4 | 76.9 | 100.0 | 59.4 |

Table 1. 3D instance segmentation results on ScanNet v2 hidden test set in terms of AP₅₀ scores. The proposed SoftGroup achieves the highest average AP₅₀, outperforming the previous strongest method by a significant margin. Reported results are from the ScanNet benchmark on 13/11/2021.

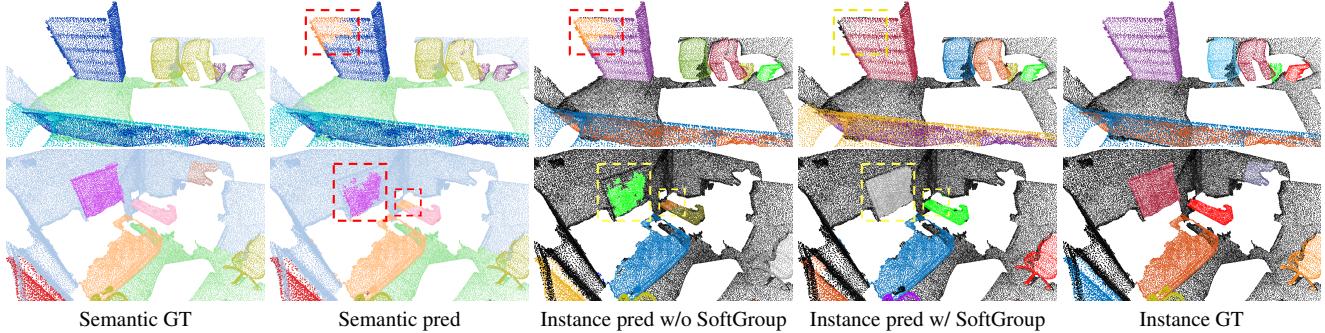


Figure 5. Qualitative results on ScanNet v2 validation set. Instance prediction without SoftGroup output low-quality instance mask at the region of wrong semantic prediction (highlighted by dashed boxes). SoftGroup produces more accurate instance masks at these regions.

4.2. Benchmarking Results

ScanNet v2. Table 1 shows the results of SoftGroup and recent state-of-the-art methods on the hidden test set of ScanNet v2 benchmark. We submit our model and report the results from the server. The proposed SoftGroup achieves the highest average AP₅₀ of 76.1%, surpassing the previous strongest methods a significant margin of 6.2%. Regarding class-wise scores, our method achieves the best performance in 12 out of 18 classes.

S3DIS. Table 2 summaries the results on Area 5 and 6-fold cross-validation of S3DIS dataset. On both Area 5 and cross-validation evaluations, the proposed SoftGroup achieves higher overall performance compared to existing method. Notably, on Area 5 evaluation, SoftGroup achieves

AP/AP₅₀ of 51.6/66.1(%), which is 8.9/6.8(%) improvement compared to the second-best. The state-of-the-art performance on both ScanNet v2 and S3DIS datasets shows the generalization advantage of our method.

Segmentation and Detection Results. We further report the instance segmentation and object detection results on ScanNet v2 validation set. To obtain object detection results, we follow the approach in [7] to extract a tight axis-aligned bounding box from the predicted point mask. Table 3 reports the instance segmentation and object detection results. Our method achieves significant improvement compared to the second-best by 3.2, 3.3, 6.3, and 7.3(%) of AP₅₀, AP₂₅, box AP₅₀, and box AP₂₅, respectively.

| Method | AP | AP ₅₀ | mCov | mWCov | mPrec ₅₀ | mRec ₅₀ |
|------------------------------|-------------|------------------|-------------|-------------|---------------------|--------------------|
| SGPN [†] [41] | - | - | 32.7 | 35.5 | 36.0 | 28.7 |
| ASIS [†] [42] | - | - | 44.6 | 47.8 | 55.3 | 42.4 |
| PointGroup [†] [15] | - | 57.8 | - | - | 61.9 | 62.1 |
| SSTNet [†] [20] | 42.7 | 59.3 | - | - | 65.5 | 64.2 |
| HAIS [†] [4] | - | - | 64.3 | 66.0 | 71.1 | 65.0 |
| SoftGroup[†] | 51.6 | 66.1 | 66.1 | 68.0 | 73.6 | 66.6 |
| SGPN [‡] [41] | - | - | 37.9 | 40.8 | 38.2 | 31.2 |
| PartNet [‡] [26] | - | - | - | - | 56.4 | 43.4 |
| ASIS [‡] [42] | - | - | 51.2 | 55.1 | 63.6 | 47.5 |
| 3D-BoNet [‡] [46] | - | - | - | - | 65.6 | 47.7 |
| OccuSeg [‡] [9] | - | - | - | - | 72.8 | 60.3 |
| GICN [‡] [22] | - | - | - | - | 68.5 | 50.8 |
| PointGroup [‡] [15] | - | 64.0 | - | - | 69.6 | 69.2 |
| SSTNet [‡] [20] | 54.1 | 67.8 | - | - | 73.5 | 73.4 |
| HAIS [‡] [4] | - | - | 67.0 | 70.4 | 73.2 | 69.4 |
| SoftGroup[‡] | 54.4 | 68.9 | 69.3 | 71.7 | 75.3 | 69.8 |

Table 2. 3D instance segmentation results on S3DIS dataset. Methods marked with [†] are evaluated on Area 5, and methods marked with [‡] are evaluated on 6-fold cross validation.

| Method | AP ₅₀ | AP ₂₅ | Box AP ₅₀ | Box AP ₂₅ |
|------------------|------------------|------------------|----------------------|----------------------|
| F-PointNet [31] | - | - | 10.8 | 19.8 |
| GSPN [47] | 37.8 | 53.4 | 17.7 | 30.6 |
| 3D-SIS [12] | 18.7 | 35.7 | 22.5 | 40.2 |
| VoteNet [30] | - | - | 33.5 | 58.6 |
| 3D-MPA [7] | 51.9 | 72.4 | 49.2 | 64.2 |
| PointGroup [15] | 51.7 | 71.3 | 48.9 | 61.5 |
| SSTNet [20] | 64.3 | 74.0 | 52.7 | 62.5 |
| HAIS [4] | 64.4 | 75.6 | 53.1 | 64.3 |
| SoftGroup | 67.6 | 78.9 | 59.4 | 71.6 |

Table 3. Instance segmentation and object detection results on ScanNet v2 validation set. Our method achieves better results on both mask and box AP.

Runtime Analysis. Table 4 report the runtime per scan of different methods on ScanNet v2 validation set. For a fair comparison, the reported runtime is measured on the same Titan X GPU model. The inference time of our method is 345ms per scan, which is extra 6ms over the fastest model. Regarding our component-time, the point-wise prediction network, soft grouping algorithm, and top-down refinement latencies are 152ms, 132ms, and 70ms, respectively. The results show that our method achieves high accuracy while remaining computationally efficient.

4.3. Qualitative Analysis

Figure 5 shows the visualization examples from ScanNet v2 dataset. Without SoftGroup, the semantic prediction errors are propagated to instance segmentation predictions

| Method | Component time (ms) | Total (ms) |
|-------------------------|---|------------|
| SGPN [41] | Backbone (GPU): 2080 Group merging (CPU): 149000 Block merging (CPU): 7119 | 158439 |
| ASIS [42] | Backbone (GPU): 2083 Mean shift (CPU): 172711 Block merging (CPU): 7119 | 181913 |
| GSPN [47] | Backbone (GPU): 1612 Point sampling (GPU): 9559 Neighbour search (CPU): 1500 | 12702 |
| 3D-BoNet [46] | Backbone (GPU): 2083 SCN (GPU): 667 Block merging (CPU): 7119 | 9202 |
| GICN [22] | Backbone (GPU): 1497 SCN (GPU): 667 Block merging (CPU): 7119 | 8615 |
| OccuSeg [9] | Backbone GPU): 189 Supervoxel (CPU): 1202 Clustering (GPU+CPU): 513 | 1904 |
| PointGroup [15] | Backbone (GPU): 128 Clustering (GPU+CPU): 221 ScoreNet (GPU): 103 | 452 |
| SSTNet [20] | Backbone (GPU) 125 Tree network (GPU+CPU): 229 ScoreNet (GPU): 74 | 428 |
| HAIS [4] | Pointwise prediction (GPU): 154 Hier. aggr. (GPU+CPU): 118 Intra-inst. prediction (GPU): 67 | 339 |
| SoftGroup (Ours) | Pointwise prediction (GPU): 152 Soft grouping (GPU+CPU): 123 Top-down refinement (GPU): 70 | 345 |

Table 4. Inference time per scan of different methods on ScanNet v2 validation set. For a fair comparison, the runtime is measured on the same Titan X GPU model.

(highlighted by dashed boxes). In contrast, SoftGroup effectively corrects the semantic prediction errors and thus generates more accurate instance masks.

4.4. Ablation Study

Component-wise Analysis. We provide experimental results of SoftGroup when different components are omitted. The considered baseline is a model with hard grouping and the confidence scores of output instances are ranked by a ScoreNet branch [15, 20]. Table 5 shows the ablation results. The baseline achieves 39.5/61.1/75.5(%) in terms of AP/AP₅₀/AP₂₅. Significant improvement is obtained by either applying soft grouping or top-down refinement. Combining these two components achieves the best overall performance AP/AP₅₀/AP₂₅ of 46.0/67.6/78.9(%), which is

| Baseline | Soft grouping | Top-down refinement | AP | AP ₅₀ | AP ₂₅ |
|----------------------------|---------------|---------------------|-------------|------------------|------------------|
| ✓ | | | 39.5 | 61.1 | 75.5 |
| | ✓ | | 41.6 | 63.8 | 79.2 |
| | | ✓ | 44.3 | 65.4 | 78.1 |
| ✓ | ✓ | ✓ | 46.0 | 67.6 | 78.9 |
| Overall improvement | | | +6.5 | +6.5 | +3.4 |

Table 5. Component-wise analysis on ScanNet v2 validation set. Our model achieves significant improvement over the baseline.

| τ | AP | AP ₅₀ | AP ₂₅ |
|--------|-------------|------------------|------------------|
| None | 44.3 | 65.4 | 78.1 |
| 0.01 | 40.1 | 58.5 | 69.2 |
| 0.1 | 45.3 | 66.5 | 78.5 |
| 0.2 | 46.0 | 67.6 | 78.9 |
| 0.3 | 45.2 | 66.8 | 78.5 |
| 0.4 | 44.7 | 46.1 | 78.3 |
| 0.5 | 43.9 | 64.8 | 77.7 |

Table 6. Ablation experiments on varying score threshold τ for soft grouping. “None” denotes the threshold is not used, and the hard semantic prediction is used for grouping instead.

significantly higher than the baseline by 6.5/6.5/3.4(%).

Score Threshold for Soft Grouping. Table 6 shows the experimental results with varying score thresholds for soft grouping. The baseline is with τ being “None”, indicating the threshold is deactivated and the hard predicted label is used for grouping. The baseline achieves AP/AP₅₀/AP₂₅ of 44.3/65.4/78.1(%). When τ is too high or too low the performance is even worse than the baseline. The best performance is obtained at τ of 0.2, which confirms our analysis at the Section 3.2, where the number of positive and negative samples are balanced.

Top-Down Refinement. We further provide the ablation results on the top-down refinement, on Table 7. With only the classification branch, our method achieves AP/AP₅₀/AP₂₅ of 41.1/64.6/79.7(%). When mask branch and mask scoring branch are in turn applied, the performance tends to improve on the higher IoU threshold regions. Combining all branches yields the performance AP/AP₅₀/AP₂₅ of 46.0/67.6/78.9(%).

Instance Category from Classification Branch. Table 8 reports the results of different schemes to obtain object categories. The results show that deriving the object category from semantic prediction yields the AP/AP₅₀/AP₂₅ of 45.0/65.6/76.2(%). The proposed method directly uses the

| Class | Mask | Mask score | AP | AP ₅₀ | AP ₂₅ |
|-------|------|------------|-------------|------------------|------------------|
| ✓ | | | 41.1 | 64.6 | 79.7 |
| ✓ | ✓ | | 45.7 | 68.4 | 79.5 |
| ✓ | ✓ | ✓ | 46.0 | 67.6 | 78.9 |

Table 7. The impact of each branch in top-down refinement on ScanNet v2 validation set.

| Category from class branch? | AP | AP ₅₀ | AP ₂₅ |
|-----------------------------|-------------|------------------|------------------|
| N | 45.0 | 65.6 | 76.2 |
| Y | 46.0 | 67.6 | 78.9 |

Table 8. Ablation study on instance category. “N” indicates that the instance category is taken from majority vote of semantic prediction. “Y” indicates that the instance category is taken from classification branch

output of the classification branch as the instance class. The classification branch aggregates all point features of the instance and classifies the instance with a single label, leading to more reliable prediction. The results show that directly using classification output as object category improves the AP/AP₅₀/AP₂₅ to 46.0/67.6/78.9(%).

5. Conclusion

We have presented SoftGroup, a simple yet effective method for instance segmentation on 3D point clouds. SoftGroup performs grouping on soft semantic scores to address the problem stemming from hard grouping on locally ambiguous objects. The instance proposals obtained from the grouping stage are assigned to either positive or negative samples. Then a top-down refinement stage is constructed to refine the positives and suppress the negatives. Extensive experiments on different datasets show that our method outperforms the existing state-of-the-art method by a significant margin of +6.2% on the hidden ScanNet v2 test set and +6.8% on S3DIS Area 5 in terms of AP₅₀.

Acknowledgements

This work was partly supported by Institute for Information communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT) (2021-0-01381, Development of Causal AI through Video Understanding, and partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01371, Development of brain-inspired AI with human-like intelligence).

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 *CVPR*, 2016. [2](#), [5](#)
- [2] Mathieu Aubry, Ulrich Schlickewei, and Daniel Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. In *ICCV workshops*, 2011. [2](#)
- [3] Michael M Bronstein and Iasonas Kokkinos. Scale-invariant heat kernel signatures for non-rigid shape recognition. In *CVPR*, 2010. [2](#)
- [4] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *ICCV*, 2021. [1](#), [3](#), [4](#), [5](#), [6](#), [7](#)
- [5] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. [2](#)
- [6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. [2](#), [5](#)
- [7] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *CVPR*, 2020. [6](#), [7](#)
- [8] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. [2](#), [3](#)
- [9] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *CVPR*, 2020. [2](#), [6](#), [7](#)
- [10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. [5](#)
- [11] Tong He, Chunhua Shen, and Anton van den Hengel. Dyco3d: Robust instance segmentation of 3d point clouds through dynamic convolution. In *CVPR*, 2021. [6](#)
- [12] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *CVPR*, 2019. [2](#), [4](#), [6](#), [7](#)
- [13] Binh-Son Hua, Minh-Khoi Tran, and Sai-Kit Yeung. Pointwise convolutional neural networks. In *CVPR*, 2018. [2](#)
- [14] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *CVPR*, 2019. [5](#)
- [15] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *CVPR*, 2020. [1](#), [3](#), [4](#), [5](#), [6](#), [7](#)
- [16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICVLR*, 2015. [5](#)
- [17] Jean Lahoud, Bernard Ghanem, Marc Pollefeys, and Martin R Oswald. 3d instance segmentation via multi-task metric learning. In *ICCV*, 2019. [2](#), [6](#)
- [18] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, 2019. [2](#)
- [19] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhuan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *NIPS*, 2018. [2](#)
- [20] Zhihao Liang, Zhihao Li, Songcen Xu, Mingkui Tan, and Kui Jia. Instance segmentation in 3d scenes using semantic superpoint tree networks. In *ICCV*, 2021. [1](#), [3](#), [4](#), [6](#), [7](#)
- [21] Chen Liu and Yasutaka Furukawa. Masc: Multi-scale affinity with sparse convolution for 3d instance segmentation. *arXiv:1902.04478*, 2019. [6](#)
- [22] Shih-Hung Liu, Shang-Yi Yu, Shao-Chi Wu, Hwann-Tzong Chen, and Tyng-Luh Liu. Learning gaussian instance segmentation in point clouds. *arXiv:2007.09860*, 2020. [2](#), [4](#), [6](#), [7](#)
- [23] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *CVPR*, 2019. [2](#)
- [24] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICVLR*, 2017. [5](#)
- [25] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IROS*, 2015. [2](#)
- [26] 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 *CVPR*, 2019. [7](#)
- [27] Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. *arXiv:1903.01177*, 2019. [6](#)
- [28] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. [5](#)
- [29] Quang-Hieu Pham, 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 *CVPR*, 2019. [2](#)
- [30] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *CVPR*, 2019. [7](#)
- [31] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *CVPR*, 2018. [7](#)
- [32] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. [2](#)
- [33] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv:1706.02413*, 2017. [2](#)
- [34] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *CVPR*, 2017. [2](#)
- [35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. [3](#)
- [36] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *ICVRA*, 2009. [2](#)

- [37] Radu Bogdan Rusu, Nico Blodow, Zoltan Csaba Marton, and Michael Beetz. Aligning point cloud views using persistent feature histograms. In *IROS*, 2008. [2](#)
- [38] Yiru Shen, Chen Feng, Yaoqing Yang, and Dong Tian. Mining point cloud local structures by kernel correlation and graph pooling. In *CVPR*, 2018. [2](#)
- [39] Martin Simonovsky and Nikos Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *CVPR*, 2017. [2](#)
- [40] 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 *ICCV*, 2019. [2](#)
- [41] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *CVPR*, 2018. [2, 6, 7](#)
- [42] Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, and Jiaya Jia. Associatively segmenting instances and semantics in point clouds. In *CVPR*, 2019. [7](#)
- [43] 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 TOG*, 2019. [2](#)
- [44] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *CVPR*, 2019. [2](#)
- [45] Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *ECCV*, 2018. [2](#)
- [46] 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 *NeurIPS*, 2019. [2, 4, 6, 7](#)
- [47] 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 *CVPR*, 2019. [2, 6, 7](#)
- [48] Biao Zhang and Peter Wonka. Point cloud instance segmentation using probabilistic embeddings. In *CVPR*, 2021. [3](#)
- [49] Biao Zhang and Peter Wonka. Point cloud instance segmentation using probabilistic embeddings. In *CVPR*, 2021. [6](#)
- [50] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021. [2](#)