CF-SIS: Semantic-Instance Segmentation of 3D Point Clouds by Context Fusion with Self-Attention

Xin Wen
School of Software, Tsinghua University
Beijing, China
x-wen16@mails.tsinghua.edu.cn

Zhizhong Han
Department of Computer Science, University of Maryland
College Park, USA
h312h@umd.edu

Geunhyuk Youk
School of Software, Tsinghua University
Beijing, China
rmsgurkjg@gmail.com

Yu-Shen Liu*
School of Software, BNRist, Tsinghua University
Beijing, China
liuyushen@tsinghua.edu.cn

ABSTRACT

3D Semantic-Instance Segmentation (SIS) is a newly emerging research direction that aims to understand visual information of 3D scene on both semantic and instance level. The main difficulty lies in how to coordinate the paradox between mutual aid and sub-optimal problem. Previous methods usually address the mutual aid between instances and semantics by direct feature fusion or hand-crafted constraints to share the common knowledge of the two tasks. However, they neglect the abundant common knowledge of feature context in the feature space. Moreover, the direct feature fusion can raise the sub-optimal problem, since the false prediction of instance object can interfere the prediction of the semantic segmentation and vice versa. To address the above two issues, we propose a novel network of feature context fusion for SIS task, named CF-SIS. The idea is to associatively learn semantic and instance segmentation of 3D point clouds by context fusion with attention in the feature space. Our main contributions are two context fusion modules. First, we propose a novel inter-task context fusion module to take full advantage of mutual aid and relive the sub-optimal problem. It extracts the context in feature space from one task with attention, and selectively fuses the context into the other task using a gate fusion mechanism. Then, in order to enhance the mutual aid effect, the intra-task context fusion module is designed to further integrate the fused context, by selectively merging the similar feature through the self-attention mechanism. We conduct experiments on the S3DIS and ShapeNet datasets and show that CF-SIS outperforms the state-of-the-art methods on semantic and instance segmentation task.

*Corresponding author. This work was supported by National Key R&D Program of China (2018YFB0505400), in part by the NSFC (61672307), National Key R&D Program of China (2019YFB1405703) and TC190A4DA/3, and in part by Tsinghua-Kuaishou Institute of Future Media Data.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413829

CCS CONCEPTS

• **Computing methodologies** → **Scene understanding**; **Object detection**.

KEYWORDS

3D shape recognition, 3D shape segmentation, point cloud

ACM Reference Format:
Xin Wen, Zhizhong Han, Geunhyuk Youk, and Yu-Shen Liu. 2020. CF-SIS: Semantic-Instance Segmentation of 3D Point Clouds by Context Fusion with Self-Attention. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413829

1 INTRODUCTION

Understanding 3D shapes and scenes has received a growing concern due to the fast development of 3D computer vision research in many real-world applications [11, 16, 22, 23, 28, 41]. It is a basic research topic in different multimedia applications, such as auto-navigation, robotics, augmented reality (AR) [33], and shape retrieval [34]. In this paper, we address the task of semantic-instance segmentation of 3D point clouds, which serves as the most fundamental problem in understanding 3D scene. Specifically, semantic segmentation aims to distinguish the class label or object category (e.g. chair, desk) for every point in 3D scenes, while the instance segmentation identifies the point set that represents an independent instance object. Therefore, the semantic-instance segmentation aims to simultaneously identify each instance object with its semantic label, which is a basic requirement for many real-world applications. For example, the auto-navigation often requires identifying the object (e.g. cars, passengers) on both semantic level and instance level.

The biggest challenge of associative segmentation lies in the paradox between sub-optimal problem and the mutual aid [31] of two segmentation tasks. Specifically, the incorrect predictions raise the sub-optimal problem, where false prediction of instance object can confuse the categorization of its semantic label and vice versa. On the other hand, the mutual aid, which lies in the endogenous relationships between semantics and instances, can serve as an instructive clue to distinguish the instance objects and the semantic categories behind the points, i.e. the points in the same instance must have the same semantic label and the points with different semantic label must belong to the different instances. Such

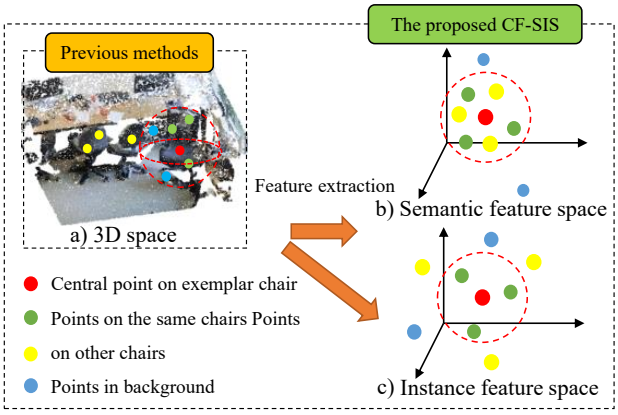


Figure 1: Illustration of the context considered by previous methods and proposed CF-SIS. For example, in CF-SIS, the instance segmentation can distinguish the points (red & green points) belong to the same instance object (chair), by excluding the spatial outliers (yellow) in 3D space (a) and the semantic outliers (blue) in semantic feature space (b).

insight has prompted several studies [9, 18, 31] to utilize the inner connection of semantic and instance segmentation and relieve the negative effect of sub-optimal problem.

To efficiently exploit the endogenous relationships between the instances and semantics, a typical practice is to regard the associative training of the two tasks as a multi-task learning problem, where the label relationships can be implicitly learned and shared by the direct point-wise feature fusion [31] or hand-crafted constraints [18]. It is usually achieved through a specially designed pipeline network or optimization objective that connects different task branches. However, such practice merely learns the spatial context in neighborhood region of a central point in 3D space, but neglects the feature context among the features embedded in feature space, which is more informative to capture the label relationships compared with the single point feature. Specifically, the *spatial context* that we refer to is the representation vector aggregated according to distance in 3D space, which contains the information of the spatial relationships between central point and the other point in point clouds. And the *feature context* is aggregated according to feature similarity in feature space, which contains the semantic relationships between central point and the other point. The visual illustration about the advantages of utilizing context in both 3D and feature space is shown in Figure 1. Let's take the instance prediction of the central red point as the example: the red point can distinguish its partners (green points) that belong to the same instance object (chair) by excluding the spatial outliers in 3D space (yellow) and the context outliers in semantic feature space (blue). Therefore, it can produce a more precise instance segmentation by considering context in 3D and semantic feature space. Except for the absence of feature context in previous methods, the practice of direct feature fusion may also raise the sub-optimal problem, since it can bring in false prior prediction from one task and confuse the downstream prediction of the other task.

Therefore, in this paper, we propose a novel network of feature context fusion for SIS, named CF-SIS, to address the above problems by learning and selectively fusing feature context in the feature space, which is neglected by many previous methods [18, 31]. The proposed CF-SIS consists of two modules, named *intra-task context fusion* and *inter-task context fusion*, both of which aim to efficiently leverage the contextual information flow in and between different tasks. The inter-task context fusion first aggregates the feature context in feature space, by modeling the relationships between neighbor and central points using the attention mechanism. Then, it selectively merges the aggregated context into the other task through the *fusion gate*. The intra-task context fusion further integrates the fused features to enhance the mutual aid effect using a self-attention mechanism, which can be regarded as a contextual-aware point feature extraction in the feature space.

The reasons that CF-SIS can address the mutual aid and sub-optimal problem are two-folds. First, compared with the previous methods which only learn context in 3D space, CF-SIS learns more informative context in both feature and 3D space, which can capture much more geometric, semantic and instance characteristics of the point clouds. Second, by introducing a learnable fusion gate, the CF-SIS can learn to filter the context information extracted from one task, and selectively fuse them into the other task. This enables the network to select the context information of real help and dropout the information about false prediction, which can effectively relieve the sub-optimal problem. The main contributions of our work can be summarized as follows:

- We propose a novel feature context fusion network for SIS task on 3D point clouds, named CF-SIS. The network can efficiently learn and fuse the feature context between the semantic and instance segmentations in the feature space, which enables the contextual-aware feature extraction for discriminative semantic-instance segmentation.
- We propose the inter-task context fusion module to exploit the mutual aid between semantics and instances. Compared with the previous methods, our module can learn to share much more informative common knowledge between semantics and instances, by means of feature context fusion with attention.
- We propose the intra-task context fusion module to further enhance the mutual aid effect. Such module can learn to decide the most related regions to the central points through the self-attention mechanism, and aggregates the informative context in the related regions to discriminatively predict the semantic categories or instance objects behind the points.

2 RELATED WORK

In the field of 3D computer vision, there are many studies concerning various representation form of 3D shapes (e.g. voxels[2, 17, 29], view[3–6, 8] and point cloud[7, 15, 32]), and in this paper we concern the segmentation task on the specific form of point cloud.

Instance segmentation. The studies concerning 3D instance segmentation can be roughly divided into two directions. The first direction is proposal-based methods [25, 26, 42], which predicts the instances by progressively proposing and refining the region proposals. As a typical work, SlidingShapes [25] is proposed to exploit handcrafted feature for predicting 3D object bounding boxes, and Frustum PointNet [19] considered the detection on 2D frame and

then back-projected into 3D, from which the final bounding box predictions are refined. In order to directly segment on point clouds, a bottom-up 3D proposal generation network and 3D proposal refining network is proposed in PointRCNN [24]. Similarly, SGPN [30] utilizes the point-level feature similarity for discovering the region proposal in point clouds. More recently, GSPN [40] is proposed as a generative framework for object proposal by reconstruction. The second direction is proposal-free methods, which directly predicts the instance features for all points and use clustering algorithm to predict instances as clusters in feature space [18, 31]. Such direction is newly proposed as a basic framework for incorporating semantic segmentation and instance segmentation associatively into one network, which will be detailed in the last part of this section.

Semantic segmentation. In the field of 3D semantic segmentation, great progress has been made in recent year because of the fast development of deep learning framework. Following the convolutional structure of PointNet [20] and PointNet++ [21], many successors [7, 12, 14, 37] investigate the convolution operations which aggregate the neighbors of a given point by edge attributes in the local region graph. In order to extract the rich representation of contextual relationships between object parts, SPG [10] is proposed to adopt the super-point graph for capture the spatial organization of 3D point clouds, in which a partition of the scanned scene is transformed into geometrically homogeneous elements, and can be further exploited by a graph convolutional network.

Associative segmentation of semantic and instance. Associatively segmenting semantics and instances is a newly merging area in the research of 3D point cloud. The associative learning of the two tasks provide a new solution for instance segmentation in point cloud, instead of the region proposals. Meanwhile, the semantic segmentation can benefit from the features learned by the instance segmentation network. To explore the potential of simultaneously learning both instance and semantic segmentation, JSIS3D [18] combined a multi-task framework with a multi-value conditional random field (MRF) model to establish the associations between semantic and instance labels, which is for separately learning the per-point semantic and instance embeddings. In the case of JSIS3D, the problem for learning both semantic and instance segmentation is formulated as the joint optimization of the energy function in the MRF model. ASIS [31], on the other hand, integrates the two tasks into an end-to-end parallel training framework, where two pipelines between the semantic and instance segmentation branches are designed to share the common knowledge in a soft and learnable fashion. Except for performing segmentations on raw point cloud input, 3D-SIS [9] considers the associative segmentation task on 3D scene represented by RGB-D scans, where two paralleled pipelines are designed to recognize the geometry and color feature, respectively. However, the above-mentioned methods usually consider the direct merging of per-point raw features from one task into the other one. In contrast, our CF-SIS takes one step further to consider more delicate extraction of common knowledge as the feature context, and carefully fuse the feature context from one task into the other using the fusion gate to purify the information.

3 THE CF-SIS NETWORK ARCHITECTURE

3.1 Overview

The overall architecture of the proposed CF-SIS is illustrated in Figure 2. The network consists of one shared feature extractor (i.e. the stacked PointNet++ [20] layers) and two paralleled branches for instance and semantic segmentation, respectively. The inter-task context fusion bridges the two branches for learning and fusing feature context, and the following intra-task context fusion aggregates the fused feature for predicting task-oriented embeddings.

Specifically, given the input point cloud of size N, the network first prepares the semantic and instance raw features, denoted by F_{sem} and F_{ins}, respectively, using the shared feature extractor with two separate multi-layer perceptrons (MLPs).

The instance branch first merges the instance raw features F_{ins} with the semantic context $F_{sem\text{-}c}$ that extracted from the semantic branch with self-attention mechanism, through the semantic-to-instance (*Sem2Ins*) sub-branch of inter-task context fusion. A fusion gate is applied to control the information that flows from the semantic branch into the instance branch for knowledge sharing. As a learnable layer, the fusion gate is trained to select how much of the semantic context should be allowed to flow into the instance branch. What's more, it also controls how much information of the instance features should be merged with the semantic context. The fused instance feature $F_{ins\text{-}f}$ is then fed into the instance-to-instance (*Ins2Ins*) sub-branch of intra-task context fusion. The intra-task context fusion module integrates the fused feature with self-attention mechanism, in order to enhance the mutual aid effect from semantic branch and predict a more discriminative instance embedding E_{ins}.

The semantic branch is nearly the same as instance branch, except that the instance context to be merged with semantic raw features F_{sem} is extracted from instance embeddings E_{ins}, through the instance-to-semantic (*Ins2Sem*) sub-branch, and without the fusion gate. The fused semantic feature $F_{sem\text{-}f}$ is fed to semantic-to-semantic (*Sem2Sem*) sub-branch of intra-task context fusion and produce the final semantic embeddings E_{sem}. The detailed architecture of each part is described as follows.

3.2 Inter-task Context Fusion

The inter-task context fusion consists of two same sub-branches, i.e. the Sem2Ins sub-branch and the Ins2Sem sub-branch, except that Ins2Sem has no fusion gate. Without losing generality, the Sem2Ins sub-branch is taken as the example for further explanation. Overall speaking, in this sub-branch, we first model the semantic correlation between points by calculating the semantic attention according to semantic raw features $F_{sem} = \{\boldsymbol{s}_i | i = 1, 2, .., N\}$. Then, the semantic context is aggregated by the weighted average of semantic raw features, and fused with the instance raw features $F_{ins} = \{\boldsymbol{u}_i | i = 1, 2, .., N\}$ using a fusion gate, in order to control the information flow from the semantic branch to the instance branch.

Specifically, given one point represented by semantic raw feature $\boldsymbol{s}_i$, the $N-1$ candidate points to form its corresponding semantic context are $\boldsymbol{S}_i = \{\boldsymbol{s}_{i_j} | j \in \{1, 2, ..., N-1\}\}$. In order to model the semantic relationships between the central point $\boldsymbol{s}_i$ and the other points, we propose the *spatial attention* to calculate the cosine similarity between the features of central and candidate

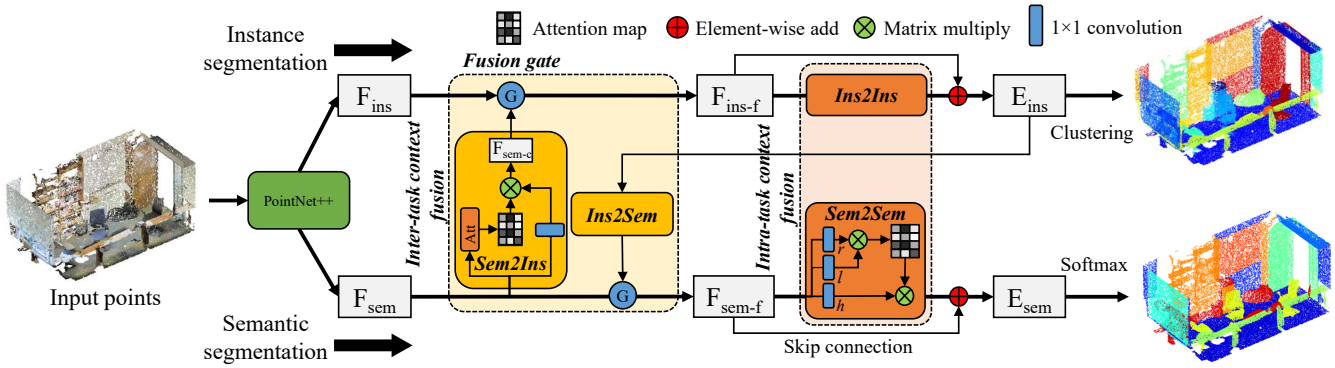


Figure 2: Illustration of the overall architecture of the proposed CF-SIS. The network consists of two task-oriented branches, where the inter-task context fusion bridges the two branches and controls the information flow through the fusion gate. The intra-task context fusion further integrates the features in each task and produces the final task-oriented embeddings for segmentation.

points as the scores to measure semantic relationships, given as $q_{i,j} = \frac{s_i^T s_{i_j}}{\|s_i\|_2 \|s_{i_j}\|_2}$, where $\| * \|_2$ denotes the L2 norm. The cosine distance can be regarded as a hand-crafted feature to characterize the semantic relationships between points. The reasons that we propose to use the cosine distance in the high dimensional feature space instead of a learnable attention mechanism are follows. First, the cosine distance can force the network to incorporate the spatial relationships between points into consideration, since two neighbor points will naturally generate similar feature and yield high cosine similarity. Second, the cosine distance can introduce more information from the other branches, since the cosine distance is an unsmoothed activation score, which has a significantly higher value compared with smoothed (by softmax activation) attention score used in learnable attention. The semantic feature context $F_{sem\text{-}c} = \{\boldsymbol{c}_i | i \in \{1, 2, .., n\}\}$ is computed as the weighted average over all candidate points as $\boldsymbol{c}_i = \sum_{j=1}^{N-1} q_{i,j} \boldsymbol{s}_{i_j}$.

Note that, the inter-task context fusion considers the context extraction over the entire input point set for the following two reasons. First, compared to the 3D space, the kNN searching for the neighbor points in the high dimensional feature space suffers from the heavy computational cost, making it hard to selectively compute attentions between point and its semantically related neighbors. Second, collecting the semantic context across the entire feature space can extend the perception range of the central point, thus force the network to learn a more discriminative choice of related points, and extract more informative semantic context.

Fusion gate in inter-task context fusion module. Since the Sem2Ins sub-branch is at the very bottom of the network compared with the Ins2Sem sub-branch, where the raw semantic features are used for context extraction. As a result, the extracted semantic context is noisy and may contain erroneous information, which can harm the performance of instance segmentation if directly fused into instance feature without preprocessing, especially at the early stage of training. Therefore, in order to filter out the noise and the erroneous information in the semantic context, and make the training less sensitive to the initialization and hyper-parameters of the model, we propose to use a fusion gate to control the information flow between the two tasks. The fusion gate enables the network to automatically learn to decide when and how much information is allowed to flow from one task into the other. Specifically, the fusion gate determines the quality of the semantic feature context $F_{sem\text{-}c}$ using a linear transformation, and predict a probability between 0 and 1 using sigmoid activation for each position of the input features. The predicted probability indicates the proportion of semantic context that is allowed to pass through. The process can be formulated as

$$\beta_{g,i} = \sigma(W_g \boldsymbol{c}_i + \boldsymbol{b}_g), \tag{1}$$

where σ represents the sigmoid activation, and $\{W_g, \boldsymbol{b}_g\}$ are learnable matrix and bias of linear transformation. The fused instance feature matrix $F_{ins\text{-}f} = \{\boldsymbol{u}_{f,i} | i \in \{1, 2, .., n\}\}$ is the weighted element-wise sum of the gated semantic feature context matrix $F_{sem\text{-}c}$ and the instance feature matrix F_{ins}, given as

$$\boldsymbol{u}_{f,i} = \beta_{g,i} \boldsymbol{c}_i + (1 - \beta_{g,i}) \boldsymbol{u}_i. \tag{2}$$

We note that the fusion gate is also deployed in the Sem2Ins branch, which has the same structure described above.

3.3 Intra-task Context Fusion

Except for aiding the context fusion between different tasks, the presence of attention inside the task can also help the model to capture the semantic and instance characteristics, by implicitly pointing out the related region to the central points. This is important for perceiving and extracting 3D visual information from point clouds, because the related regions contain more information about the semantic and instance characteristics compared to the single point, which can be further used to distinguish the semantic labels and instance objects behind the points, and yield a more discriminative segmentation result. In this subsection, we find inspiration from the previous self-attention based work [35] to learn the per point context in feature space, through an effective intra-task context fusion.

The module consists of two same sub-branches, i.e. Ins2Ins sub-branch and Sem2Sem sub-branch. Without losing generality, we only describe the detailed structure of Ins2Ins sub-branch. The input to this module is the fused instance feature $F_{ins\text{-}f} = \{\boldsymbol{u}_{f,i} | i \in$

$\{1, 2, .., N\}\}$. These features are first transformed into three different feature subspace $\{r, l, h\}$, through the multilayer perceptrons f with parameter $\boldsymbol{\theta}$, given as

$$\boldsymbol{u}_{\tau,i} = f(\boldsymbol{u}_{f,i}|\boldsymbol{\theta}_\tau), \tau \in \{r, l, h\}. \tag{3}$$

The attention is computed as the dot product between the feature vectors in subspace r and l, and is smoothed using softmax activation, which can be written as

$$a_{i,j} = \frac{\exp(\boldsymbol{u}_{l,i}^{\mathrm{T}} \boldsymbol{u}_{r,j})}{\sum_{j=1}^{N} \exp(\boldsymbol{u}_{l,i}^{\mathrm{T}} \boldsymbol{u}_{r,j})}. \tag{4}$$

The final instance point embedding $E_{ins} = \{\boldsymbol{e}_{ins,i}\}$ is the weighted average over all features in subspace h with a skip connections linking to $F_{ins\text{-}f}$, given as:

$$\boldsymbol{e}_{ins,i} = f(\lambda_o \sum_{j=1}^{N} a_{i,j} \boldsymbol{u}_{h,j} + \boldsymbol{u}_{f,i}|\boldsymbol{\theta}_o), \tag{5}$$

where λ_o is a learnable balance factor that randomly initialized and updated along with other parameters during training. Note that, the intra-task context fusion aims at refining the instance features, which will be used for predicting the per-point instance label. In order to preserve the information of the original instance feature, the smoothed attention with a relatively small score for context features is more desirable than the unsmoothed cosine distance.

3.4 Training and Details

3.4.1 Training Losses. We use the softmax cross entropy loss for semantic branch. And for instance branch, we follow the class-agnostic loss of ASIS. Specifically, given a point cloud with K instances, we first gather all the predicted instance embeddings for the k-th ground truth instance as $I_k = \{\boldsymbol{e}_{ins,k_j} | j = 1, 2, .., n_k\}$, where n_k denotes the number of ground truth points belong to instance I_k. Then, we optimize the instance segmentation branch by merging the embeddings belonging to the same instance object together, and repelling the embeddings of different instance objects away from each other. Let M_k denote the number of points in k-th instance. The optimization function can be written as follows:

$$L_m = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M_k} \sum_{j=1}^{M_k} \max(\|\mu_k - e_{k_j}\|_2 - \delta_m, 0), \tag{6}$$

$$L_r = \frac{1}{K(K-1)} \sum_{i \neq j} \max(\|\mu_i - \mu_j\|_2 - \delta_r, 0), \tag{7}$$

where the μ_i is the center of instance I_k and $\{\delta_m, \delta_r\}$ are predefined thresholds. Besides, we also regularize the instance center μ_i by $L_\mu = \frac{1}{K} \sum_{i=1}^{K} \|\mu_i\|_2$. The total instance loss is the weighted sum of L_m, L_r and L_μ, given as

$$L_{ins} = L_m + L_r + \lambda_\mu L_\mu, \tag{8}$$

where λ_μ is the weight factor fixed to 0.001.

3.4.2 Training Settings. For all experiments, we use the Adam optimizer with a initial learning rate of 0.001 for training. We set the hyper-parameter β_1 of Adam optimizer as 0.9 and use the default β_2. The learning rate is decayed for every 10 epoches with a decay rate of 0.7, and clipped at the minimum learning rate of 1×10^{-5}. We apply one dropout layer with 0.5 dropout rate before the output layer of both semantic and instance branches. We train the CF-SIS for 100 epoches with a batch size of 12 on S3DIS dataset and 200 epoches with a batch size of 16 on ShapeNet dataset.

Table 1: Instance segmentation results (%) on S3DIS.

Method	Area 5			6-fold CV		
	mAP	mRec	mCov	mAP	mRec	mCov
SGPN [30]	36.0	28.7	32.7	38.2	31.2	37.9
JSIS3D [18]	-	-	-	36.3	-	-
ASIS [31]	55.3	42.4	44.6	63.6	47.5	51.2
3D-BoNet [38]	57.5	40.2	-	**65.6**	47.6	-
Ours(CF-SIS)	**59.1**	**46.8**	**49.9**	65.1	**52.4**	**52.6**

Table 2: Semantic segmentation results (%) on S3DIS.

Method	Area 5			6-fold CV		
	oAcc	mAcc	mIoU	oAcc	mAcc	mIoU
PointNet [20]	83.5	52.1	43.4	78.6	-	47.7
JSIS3D [18]	-	-	-	87.4	-	-
ASIS [31]	86.9	60.9	53.4	86.2	70.1	59.3
Ours(CF-SIS)	**88.7**	**67.3**	**58.9**	**88.0**	**74.0**	**63.2**

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

For fair and comprehensive comparison, we follow the same settings of ASIS [31] to evaluate the CF-SIS under the S3DIS and ShapeNet dataset. S3DIS [1] is a large scale scan dataset that contains 3D scans from Matterport Scanners in 6 areas, which have 272 rooms in total and each point is assigned with an instance object and one semantic label of 13 categories. We report the experimental results on the official 6-fold cross validation (6-fold CV) in accord with [1] and also on Area 5. The ShapeNet [39] dataset contains 16,881 3D shapes from 16 categories, each point is assigned with one semantic label and without instance annotation. Therefore, for training, we used the generated instance annotations from [30] as the ground-truth labels. For semantic segmentations, we use the overall accuracy (oAcc) and mean IoU (mIoU) as the evaluation metrics. For instance segmentation, the mean recall (mRec), mean average precision (mAP) with IoU threshold 0.5 and the weighted coverage (wCov) [31] are adopted as the evaluation metric.

4.2 S3DIS Results

S3DIS quantitative comparison. The experiment on S3DIS dataset follows the basic settings in [30], where points are uniformly sampled into overlapped blocks of size $1m \times 1m$ with a stride of $0.5m$. In each block, we randomly sample 4, 096 points. The embeddings of all blocks are merged using the `BlockMerging` procedure in [31].

In Table 1, we compare the overall instance segmentation performance on both 6-fold CV and Area 5 on S3DIS dataset, where the CF-SIS achieves the best results on 5 out of 6 metrics on both splittings. In our opinion, the better performance of CF-SIS on these categories can be dedicated to the following reason. That is, in the 3D scene of a single room, the objects in these categories usually appear multiple times at different position, and the semantic context fusion can help the model find the common characteristics among these spatially distant objects. Taking the chair as an example, learning the semantic context can aggregate these spatially unrelated

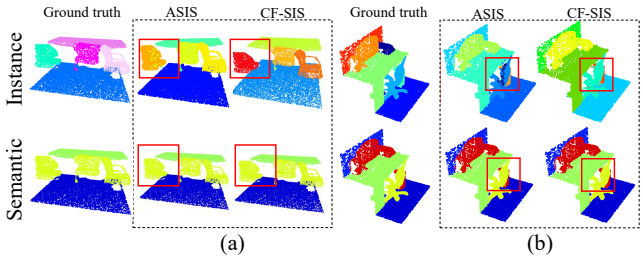


Figure 3: Illustrative comparison of CF-SIS with ASIS on S3DIS. Red rectangles denote the region that CF-SIS outperforms ASIS. For example, in (a), CF-SIS yields a cleaner instance segmentation on chair back, and makes less mistakes in semantic segmentation, compared to the ASIS.

chairs together and filter out the background objects such as wall, floor and tables, which makes the objects more distinguishable from the point cloud and easy to be segmented. A visual analysis about this fact will be given later in subsection of *Visualization of attentions*.

The overall segmentation results on 6-fold CV and Area 5 are shown in Table 2, in which we compare our CF-SIS with the counterpart methods ASIS [31] and PointNet [20]. Since the JSIS3D [18] did not report the commonly used IoU in semantic segmentation, we compare with the JSIS3D on overall accuracy in Table 2. SGPN [30] and 3D-BoNet [38] are not semantic segmentation methods which will also not going to be compared. The results show that CF-SIS outperforms the counterpart methods on both splittings. Note that although JSIS3D yields an oAcc comparable to our method, the CF-SIS significantly outperforms the JSIS3D in instance segmentation on mAP (27.8%), which proves that CF-SIS can achieve a more balanced performance between two tasks.

In Table 4, we show the detailed instance segmentation results in terms of per class mAP on 6-fold CV of S3DIS dataset. Note that JSIS3D didn't report the results on beam and column, while SGPN missed the result on clutter, of which the missing values are all marked by "-" in Table 4. Overall speaking, CF-SIS outperforms the counterpart methods in terms of mAP in 7 out of 13 classes. Especially, the segmentation performance on the categories of table, chair and sofa is improved by over 10%. And in Table 3, we show the detailed semantic segmentation results in terms of per class IoU on 6-fold CV of S3DIS dataset. We can find that CF-SIS achieves the best on 11 out of 13 classes. In Figure 4, we show more segmentation results on S3DIS dataset.

S3DIS visualization comparison. To further demonstrate the effectiveness of our method, we visualize the segmentation results in Figure 3 and illustratively compare CF-SIS with ASIS. We take Figure 3(a) as an example. Since ASIS did not adopt the fusion gate for feature fusion, the false predictions of semantics (top row of ASIS, on the chair back denoted by red rectangle) mislead the segmentation of instances (bottom row of ASIS). In contrast, although the semantic branch of CF-SIS makes the similar false semantic predictions on the same region, the instance branch can still make the correct prediction. This proves the effectiveness of the fusion gate in CF-SIS. In Figure 4, we visualize more results on rooms of S3DIS dataset, of which the predicted results show significant consistency with the ground truth.

S3DIS conclusions. From the comparison, we draw the following two conclusions. First, the experimental results show that the proposed CF-SIS outperforms the baseline SGPN [30] (also using PointNet++ backbone). Since SGPN cannot take the advantages of semantic label for instance segmentation, the better performance of CF-SIS proves the effectiveness of the mutual aid between the semantic and instance segmentation tasks. Second, CF-SIS yields better results compared with the point-wise fusion based ASIS. This proves the effectiveness of the context fusion module designed in CF-SIS, which aims to learning the instance-aware semantic feature and semantic-aware instance feature.

4.3 ShapeNet Results

For the ShapeNet dataset, we follow the practice of [30] to obtain the instance ground truths by clustering points in each semantic categories, using the DBSCAN [36] algorithm. The experiment on ShapeNet dataset is conducted for further analyzing the effectiveness of mutual aid and context fusion, in terms of the semantic segmentation performance. The reason is that, the clustered instance parts of object are subsets that are geometrically decomposed from the semantic categories (e.g. the semantic label of chair legs can be divided into several independent leg instances). Such geometric decompositions can be regarded as hand-crafted features for semantic segmentation. And CF-SIS can extract these features from the instance branch and fuse them into the semantic branch, which can provide more information about the detailed geometric structure of object, and improve the performance of semantic segmentation.

We compared with the backbone method PointNet++ and counterpart method ASIS in Table 5, in which we report part-averaged IoU (pIoU, %) and mean per-class pIoU (mpIoU, %) [12]. For fair comparison, in the ShapeNet experiment, the backbone of CF-SIS is exactly the same as PointNet++, including both the feature extraction and interpolation layers. Therefore, the improvement (by 1.1% of pIou and 1.8% of mpIoU) of CF-SIS compared to the PointNet++ proves the effectiveness of fusing the instance features into semantic segmentation task. And the better performance of CF-SIS compared with ASIS demonstrates that context fusion can learn and share the instance information more efficiently than point-wise feature fusion.

In Figure 5, we visualize the semantic and instance segmentation results of CF-SIS. Figure 5(a) and 5(b) are semantic segmentation ground truths and predictions. Figure 5(c) shows instance ground truth, in which spatially disconnected parts of same semantic category are successfully separated into independent objects, and Figure 5(d) shows that CF-SIS effectively learns such instance segmentation.

4.4 Model Analysis

In this subsection, we analyze the effectiveness of each part to the proposed CF-SIS. By default, all the ablation experiments are conducted on the splitting of Area 5 for efficient evaluation and use the spatial attention version of CF-SIS.

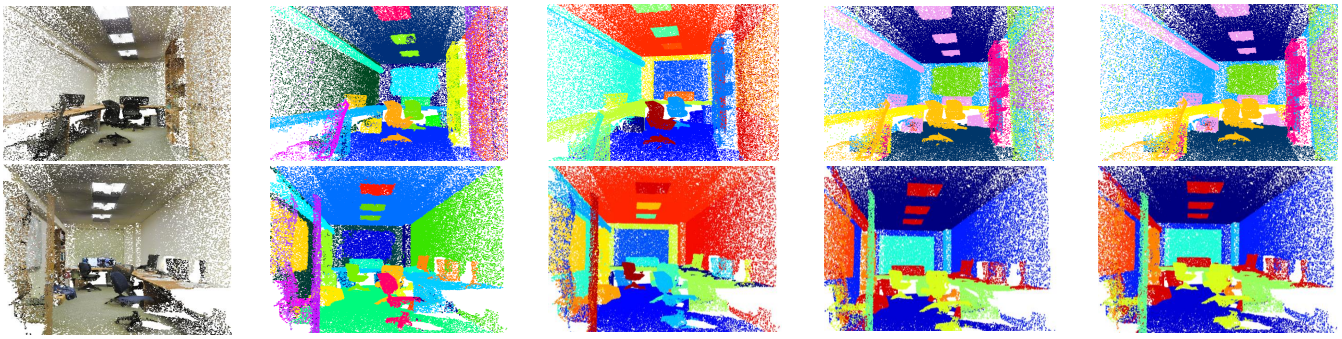
Effect of each part. We develop and evaluate five different variations of our model: (1) The *No-Gate* variation is the model

Table 3: Per class results (%) of semantic segmentation on S3DIS dataset with 6-fold cross validation in terms of IoU.

Method	cei.	floor	wall	beam	col.	win.	door	tab.	chair	sofa	book.	board	clut.
PointNet	88.8	**97.3**	69.8	42.4	23.1	46.3	10.8	52.6	58.9	40.3	5.9	26.4	33.2
ASIS	91.3	89.7	69.8	**45.8**	27.0	51.9	55.1	61	49.3	9.1	40.2	33.5	40.7
CF-SIS(Ours)	**94.2**	95.6	**77.6**	36.4	**37.4**	**53.3**	**66.9**	**65.8**	**66.6**	**38.6**	**52.6**	**53.1**	**59.0**

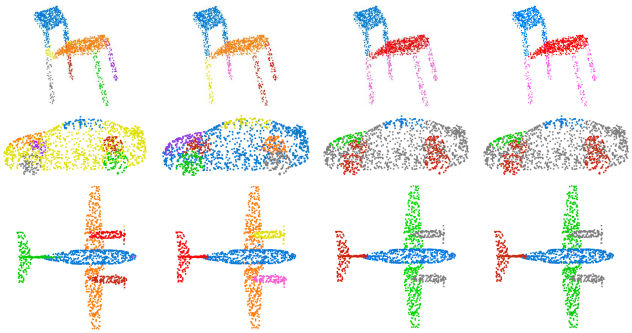
Table 4: Per class results (%) of instance segmentation on S3DIS dataset with 6-fold cross validation in terms of mAP.

Method	cei.	floor	wall	beam	col.	win.	door	tab.	chair	sofa	book.	board	clut.
JSIS3D	76.9	83.6	32.2	-	-	51.4	7.2	16.3	23.6	16.7	21.8	52.1	13.4
SGPN	79.4	66.3	**87.7**	**78.0**	**60.4**	66.6	56.8	46.9	40.8	6.38	47.6	11.1	-
ASIS	90.6	**88.0**	62.2	40.0	48.8	71.2	56.9	52.6	48.0	26.5	**50.8**	**86.3**	40.9
CF-SIS(Ours)	**91.8**	87.7	67.4	40.6	43.1	**74.9**	**62.8**	**60.2**	**70.8**	**42.2**	50.5	74.8	**52.3**



(a) Input (b) Pred Ins. (c) Gt Ins. (d) Pred Sem. (e) Gt Sem.

Figure 4: Illustrative evaluation of the segmentation results on S3DIS dataset. From left to right: (a) input point with RGB color; (b) predicted instance; (c) ground truth instance; (d) predicted semantics; (e) ground truth semantics.



(a) Gt Ins. (b) Pred Ins. (c) Gt Sem. (d) Pred Sem.

Figure 5: Illustrative evaluation of the segmentation results on ShapeNet.

that removes the gate fusion structure from the CF-SIS. As a result, the context from one task is allowed to completely flow into the other task without filtering. (2) The *Inter-only* variation is the model without the intra-task context fusion module. (3) The *Intra-only* variation is the reverse version of Inter-only model that removes the inter-task context fusion from the CF-SIS. (4) The *Sem2Ins-only* and (5) *Ins2Sem-only* are variations that only preserve one sub-branch of inter-task context fusion, in which the single direction of

Table 5: Semantic segmentation results (%) on ShapeNet.

Methods	pIoU	mpIoU
PointNet [20]	83.7	80.4
PointNet++ [21]	85.1	81.9
ASIS [31]	85.0	-
Point2Sequence [14]	85.2	-
LRC-Net [13]	85.3	-
SGPN [30]	85.8	-
SPLATNet [27]	85.4	83.7
Ours(CF-SIS)	**86.2**	**83.7**

context flow from semantic to instance is allowed or vice versa. The experimental results are shown in Table 6. Except for the above five variations, we also include the results of *Full* CF-SIS model and the *baseline* model (i.e. remove both inter-task and intra-task context fusion along with the fusion gate) for comparison.

From Table 6 we can find that each part contributes to the final performance of the full CF-SIS. Compared with the Full model, the semantic and instance performances of all variations drop at the same time, but with different scale. For Sem2Ins-only variation, the performance of semantic segmentation drops more drastically than the instance segmentation, and the opposite results can be observed in the Ins2Sem-only variation. Such phenomenon proves

Table 6: The effect (%) of each part to the CF-SIS.

Methods	Instance		Semantic	
	mAP	mRec	mAcc	mIoU
No-Gate	51.0	41.7	59.9	52.2
Inter-only	56.4	44.6	63.9	55.5
Intra-only	53.8	42.3	61.8	54.2
Sem2Ins-only	55.3	43.4	60.2	50.5
Ins2Sem-only	52.7	41.4	62.1	54.5
baseline	50.1	40.8	59.2	51.3
Full	**59.1**	**46.8**	**67.3**	**58.9**

the close connection between the two task branches, where both tasks can improve itself by taking contextual information from the other one and in return benefit the other task with its own context. The significance of the inter-task context fusion in respect to the two segmentation branches is more obvious, when comparing the Sem2Ins-only, Ins2Sem-only and the Inter-only variations to the baseline model. Especially, the performance of Inter-only model improves by 6.3% in terms of mAP for instance segmentation and 4.2% in terms of mIoU for semantic segmentation. The intra-task context fusion shows its effectiveness when comparing the Intra-only variation to the baseline model, in which the module increases the performance by 3.7% in terms of mAP and 2.9% in terms of mIoU.

Effect of each attention in context fusion module. In order to analyze the different attention mechanism used in inter-task context fusion and intra-task context fusion, w develop two variations: (1) *Intra-C* is the variation that replace the learnable attention with the cosine distance in intra-task context fusion. (2) *Inter-L* is the variation that replace the cosine distance in inter-task context fusion with the learnable attention. We compare the performance of these two variations with original *Full* model in Table 7, from which we can find that the original Full model achieves the best performance. The results are in accordance with the discussion in Sec 3.2 and Sec 3.3, in which we use the unsmoothed cosine distance in inter-task context fusion to introduce more context information from the other branch, and we use the smoothed learnable attention in intra-task context fusion to preserve the information of original feature vector when fusing the context.

Table 7: The effect (%) of attention in context fusion module.

Methods	Instance		Semantic	
	mAP	mRec	mAcc	mIoU
Intra-C	55.0	45.7	60.8	52.2
Inter-L	51.4	44.6	63.1	54.5
Full	**59.1**	**46.8**	**67.3**	**58.9**

Visualization of attentions. In Figure 6, we visualize the attentions learned in the inter-task context fusion in an $1m \times 1m$ block sampled from S3DIS, which is the standard input to CF-SIS on this dataset. The visual effects of both inter-task and intra-task context fusion modules are similar, so we exhibit the inter-task context fusion module for simplicity. Specifically, Figure 6(b) and 6(c) shows the attention assigned by red points (on the chairs in

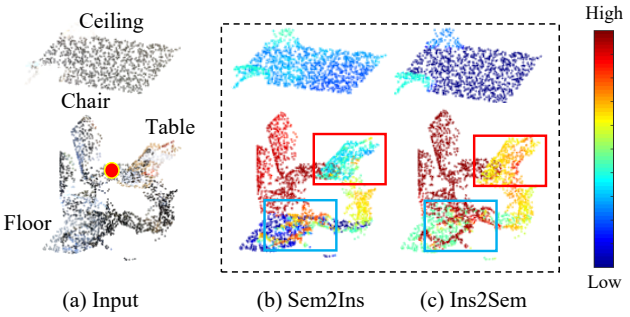


Figure 6: Visualization of the attention learned in our inter-task context fusion. The left (a) is the input $1m \times 1m$ block with RGB color. The right (b) and (c) is the visualized attention of all other points to the red point on the chair.

Figure 6(a)) to the other points in the block. In Figure 6(b) and 6(c), the region with high attention scores are colored by red, and the regions with low attention scores are colored by blue. The visualization shows that in both sub-branches, points belong to chairs are all assigned with significantly higher attentions than the other points, which proves that inter-task context fusion successfully learns to discriminate the related points in both instance and semantic feature space. What's more, the rectangles in Figure 6(b) and 6(c) further demonstrate the interaction of mutual aid between two tasks. In Figure 6(b), the semantic segmentation assigns a low attention scores to the table (red rectangles), while in Figure 6(c) the instance segmentation assigns a relatively high attention scores to the table (also red). Therefore, the mutual aid can share the correct segmentation predictions through Sem2Ins into instance branch. On the other hand, when semantic segmentation generates a poor prediction in the conjunctive region between chair and floor (blue rectangle in Figure 6(b)), it can collect correct prediction from the instance segmentation (also blue in Figure 6(c)) through Ins2Sem sub-branch.

5 CONCLUSION

We introduce a novel network, named CF-SIS, for semantic-instance segmentation. Through the proposed two novel modules, called inter-task context fusion and the intra-task context fusion, the CF-SIS can effectively learn and fuse the context from one task into the other, which enables to share the common information between the semantic and instance segmentation tasks. The shared information can be used as an instructive clue to more accurately distinguish the instance object and its category behind the points. Comprehensive experiments on S3DIS and ShapeNet dataset prove the effectiveness of the propose model.

REFERENCES

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. 2017. Joint 2D-3D-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105* (2017).
[2] Zhizhong Han, Chao Chen, Yu-Shen Liu, and Matthias Zwicker. 2020. DRWR: A Differentiable Renderer without Rendering for Unsupervised 3D Structure Learning from Silhouette Images. In *International Conference on Machine Learning (ICML)*.
[3] Zhizhong Han, Xinhai Liu, Yu-Shen Liu, and Matthias Zwicker. 2019. Parts4Feature: Learning 3D global features from generally semantic parts in

multiple views. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 766–773.
[4] Zhizhong Han, Honglei Lu, Zhenbao Liu, Chi-Man Vong, Yu-Shen Liu, Matthias Zwicker, Junwei Han, and CL Philip Chen. 2019. 3D2Seqviews: Aggregating sequential views for 3D global feature learning by CNN with hierarchical attention aggregation. *IEEE Transactions on Image Processing* 28, 8 (2019), 3986–3999.
[5] Zhizhong Han, Mingyang Shang, Yu-Shen Liu, and Matthias Zwicker. 2019. View Inter-Prediction GAN: Unsupervised representation learning for 3D shapes by learning global shape memories to support local view predictions. In *AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 33. 8376–8384.
[6] Zhizhong Han, Mingyang Shang, Zhenbao Liu, Chi-Man Vong, Yu-Shen Liu, Matthias Zwicker, Junwei Han, and CL Philip Chen. 2019. Seqviews2Seqlabels: Learning 3D global features via aggregating sequential views by RNN with attention. *IEEE Transactions on Image Processing* 28, 2 (2019), 658–672.
[7] Zhizhong Han, Xiyang Wang, Yu-Shen Liu, and Matthias Zwicker. 2019. Multi-Angle Point Cloud-VAE: Unsupervised Feature Learning for 3D Point Clouds From Multiple Angles by Joint Self-Reconstruction and Half-to-Half Prediction. In *IEEE International Conference on Computer Vision (ICCV)*. 10441–10450.
[8] Zhizhong Han, Xiyang Wang, Chi-Man Vong, Yu-Shen Liu, Matthias Zwicker, and CL Chen. 2019. 3DViewGraph: Learning global features for 3D shapes from a graph of unordered views with attention. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 758–765.
[9] Ji Hou, Angela Dai, and Matthias Nießner. 2019. 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4421–4430.
[10] Loic Landrieu and Martin Simonovsky. 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4558–4567.
[11] Xiang Li, Xiaojing Yao, and Yi Fang. 2018. Building-A-Nets: Robust building extraction from high-resolution remote sensing images with adversarial networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11, 10 (2018), 3680–3687.
[12] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. 2018. PointCNN: Convolution on x-transformed points. In *Advances in Neural Information Processing Systems*. 820–830.
[13] Xinhai Liu, Zhizhong Han, Fangzhou Hong, Yu-Shen Liu, and Matthias Zwicker. 2020. LRC-Net: Learning discriminative features on point clouds by encoding local region contexts. *Computer Aided Geometric Design* (2020), 101859.
[14] Xinhai Liu, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker. 2019. Point2Sequence: Learning the shape representation of 3D point clouds with an attention-based sequence to sequence network. In *AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 33. 8778–8785.
[15] Xinhai Liu, Zhizhong Han, Xin Wen, Yu-Shen Liu, and Matthias Zwicker. 2019. L2G Auto-Encoder: Understanding point clouds by local-to-global reconstruction with hierarchical self-attention. In *ACM International Conference on Multimedia (ACM MM)*. 989–997.
[16] Yu-Shen Liu, Yi Fang, and Karthik Ramani. 2009. Using least median of squares for structural superposition of flexible proteins. *BMC bioinformatics* 10, 1 (2009), 29.
[17] Yu-Shen Liu, K. Ramani, and M. Liu. 2011. Computing the inner distances of volumetric models for articulated shape description with a visibility graph. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 12 (2011), 2538–2544.
[18] Quang-Hieu Pham, Thanh Nguyen, Binh-Son Hua, Gemma Roig, and Sai-Kit Yeung. 2019. JSIS3D: Joint Semantic-Instance Segmentation of 3D Point Clouds with Multi-Task Pointwise Networks and Multi-Value Conditional Random Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8827–8836.
[19] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. 2018. Frustum Pointnets for 3D Object Detection from RGB-D Data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 918–927.
[20] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[21] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *In Advances in Neural Information Processing Systems*. 5099–5108.
[22] Mengwei Ren, Liang Niu, and Yi Fang. 2017. 3D-A-Nets: 3D deep dense descriptor for volumetric shapes with adversarial networks. *arXiv preprint arXiv:1711.10108* (2017).
[23] Yiting Shao, Qi Zhang, Ge Li, Zhu Li, and Li Li. 2018. Hybrid Point Cloud Attribute Compression Using Slice-based Layered Structure and Block-based Intra Prediction. In *Proceedings of the 26th ACM international conference on Multimedia*. 1199–1207.
[24] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 2019. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–779.
[25] Shuran Song and Jianxiong Xiao. 2014. Sliding Shapes for 3D Object Detection in Depth Images. In *European conference on computer vision*. Springer, 634–651.
[26] Shuran Song and Jianxiong Xiao. 2016. Deep sliding shapes for amodal 3D object detection in RGB-D images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 808–816.
[27] Hang Su, Varun Jampani, Deqing Sun, and Subhransu Maji. 2018. SPLATNet: Sparse Lattice Networks for Point Cloud Processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2530–2539.
[28] Gusi Te, Wei Hu, Amin Zheng, and Zongming Guo. 2018. RGCNN: Regularized Graph CNN for Point Cloud Segmentation. In *Proceedings of the 26th ACM international conference on Multimedia*. 746–754.
[29] Chao Wang, Yu-Shen Liu, Min Liu, Jun-Hai Yong, and Jean-Claude Paul. 2012. Robust shape normalization of 3D articulated volumetric models. *Computer-Aided Design* 44, 12 (2012), 1253–1268.
[30] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. 2018. SGPN: Similarity Group Proposal Ntwork for 3D Point Cloud Iinstance Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2569–2578.
[31] Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, and Jiaya Jia. 2019. Associatively Segmenting Instances and Semantics in Point Clouds.. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[32] Xin Wen, Tianyang Li, Zhizhong Han, and Yu-Shen Liu. 2020. Point cloud completion by skip-attention network with hierarchical folding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[33] Wanmin Wu and Cha Zhang. 2014. Immersive 3D Communication. In *Proceedings of the 26th ACM international conference on Multimedia*. 1229–1230.
[34] Jin Xie, Guoxian Dai, and Yi Fang. 2017. Deep multimetric learning for shape-based 3D model retrieval. *IEEE Transactions on Multimedia* 19, 11 (2017), 2463–2474.
[35] Saining Xie, Sainan Liu, Zeyu Chen, and Zhuowen Tu. 2018. Attentional ShapeContextNet for point cloud recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4606–4615.
[36] Xiaowei Xu, Martin Ester, Hanspeter Kriegel, and Jorg Sander. 1998. A distribution-based clustering algorithm for mining in large spatial databases. In *International Conference on Data Engineering*. 324–331.
[37] Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. 2018. SpiderCNN: Deep learning on point sets with parameterized convolutional filters. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 87–102.
[38] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. 2019. Learning Object Bounding Boxes for 3D Instance Segmentation on Point Clouds. In *Proceedings of the International Conference on Computer Vision*.
[39] Li Yi, Vladimir G Kim, Duygu Ceylan, Ichao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas J Guibas. 2016. A scalable active framework for region annotation in 3D shape collections. In *International Conference on Computer Graphics and Interactive Techniques*, Vol. 35. 210.
[40] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. 2019. GSPN: Generative Shape Proposal Network for 3D Instance Segmentation in Point Cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3947–3956.
[41] Na Zhao. 2018. End2End Semantic Segmentation for 3D Indoor Scenes. In *Proceedings of the 26th ACM international conference on Multimedia*. 810–814.
[42] Yin Zhou and Oncel Tuzel. 2018. VoxelNet: End-to-end Learning for Point Cloud Based 3D Oobject Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4490–4499.