Structural Relational Reasoning of Point Clouds

Yueqi Duan[1,2,3], Yu Zheng[1,2,3], Jiwen Lu[1,2,3,*], Jie Zhou[1,2,3], Qi Tian[4]
[1]Department of Automation, Tsinghua University, China
[2]State Key Lab of Intelligent Technologies and Systems, China
[3]Beijing National Research Center for Information Science and Technology, China
[4]Huawei Noah's Ark Lab, China
duanyq14@mails.tsinghua.edu.cn; yu-zheng15@mails.tsinghua.edu.cn; lujiwen@tsinghua.edu.cn;
jzhou@tsinghua.edu.cn; tian.qi1@huawei.com

Abstract

The symmetry for the corners of a box, the continuity for the surfaces of a monitor, the linkage between the torso and other body parts — it suggests that 3D objects may have common and underlying inner relations between local structures, and it is a fundamental ability for intelligent species to reason for them. In this paper, we propose an effective plug-and-play module called the structural relation network (SRN) to reason about the structural dependencies of local regions in 3D point clouds. Existing network architectures on point sets such as PointNet++ capture local structures individually, without considering their inner interactions. Instead, our SRN simultaneously exploits local information by modeling their geometrical and locational relations, which play critical roles for our humans to understand 3D objects. The proposed SRN module is simple, interpretable, and does not require any additional supervision signals, which can be easily equipped with the existing networks. Experimental results on benchmark datasets indicate promising boosts on the tasks of 3D point cloud classification and segmentation by capturing structural relations with the SRN module.

1. Introduction

Recent years have witnessed rapid development on 3D point cloud data due to the popularity of varying scanning devices. Typically, point cloud data is represented by sparse and unordered 3D points. Compared with 2D images which usually have regularly arranged pixels, it is more challenging to analyze 3D point clouds with the irregular structures [39]. Early works mainly focus on extracting hand-crafted 3D features, which aim to exploit statistical properties of point sets and are especially designed to present robustness to transformations [26]. In general, these methods can be divided into two categories: intrinsic features [3, 5, 33] and extrinsic features [30, 31, 22, 7, 15], where representative hand-crafed 3D features include spin image [15], fast point feature histograms (FPFH) [30] and heat kernel signature (HKS) [33].

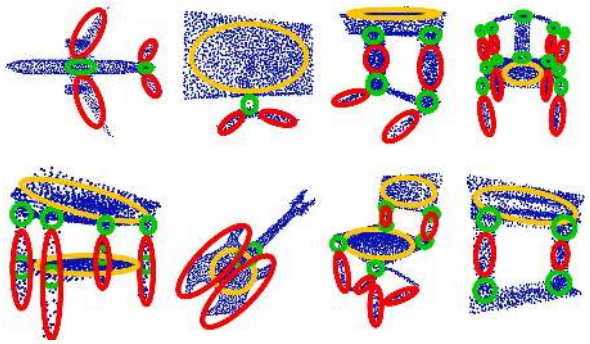

Figure 1. Examples of highly related local structures in common 3D objects. In the figure, red ellipses represent repetitive symmetric local regions, orange ellipses are continuous surfaces (on which the local structures share the same plane), and green ellipses show the junctions of connected parts. We observe that most real-world 3D objects contain highly related local structures and it is a fundamental ability for our humans to reason about them.

With the great success of deep learning methods on 2D image processing and understanding, pioneer works have been proposed to design network architectures for 3D point cloud data [26, 28, 25, 17, 43]. The initial PointNet directly learns global features from the input point clouds by point-wise spatial encoding and aggregation, which obtains encouraging performance [26]. While PointNet does not capture local information of point clouds, the extension work PointNet++ demonstrates the importance of local structure exploitation [28]. In order to capture local structures, PointNet++ designs sampling and grouping layers to obtain local sub-clouds, and individually encodes the local regions

* Corresponding author

by performing point feature embedding on each sub-cloud. Then, PointNet++ aggregates the local features for holistic representation with a pooling operation to acquire invariance to point-wise permutations. However, it fails to consider the underlying structural interactions between local regions due to the simple pooling operation, which is one of the key components in understanding 3D objects.

As examples shown in Figure 1, most real-world 3D objects have highly related local structures such as repetitive symmetric regions, continuous surfaces and connected parts, and it is a fundamental ability for our humans to reason about such inner structural relations when analyzing a 3D object. Structural relational reasoning plays a crucial role in 3D object understanding, especially for the point cloud data where only the coordinates of points are provided without further information. For example, if we are learning to recognize the species of "human" from their 3D shapes, it is far from satisfactory to individually remember the structures of all the body parts (head, torso, arms and legs). More importantly, we need to reason about their structural relations, such as the symmetry of two arms and of two legs, and the linkage between the torso and other body parts. We may also notice the symmetric two large wings at the first glance of an airplane. These relations are key components when understanding the holistic 3D structures of humans, airplanes as well as other objects, and we are able to better grasp the their semantics with the exploitation of structural relations.

In this paper, we propose a simple module named the structural relational network (SRN) to reason for the interactions between local regions, which can be plugged into existing networks without additional labels. Our SRN is inspired by the recent relational networks [32, 13, 45]. While most existing methods model the spatial or temporal relations for images and videos, SRN aims to capture the structural interactions between local regions in 3D point clouds. More specifically, we compute the geometrical and locational interactions between each local structure and others to reason for their relations, so that the learned local features encode not only the 3D structures but also the dependencies with other local regions. In our experiments, we equip our SRN with the widely-used PointNet++ architecture [28] to show the effectiveness, which is also widely applicable to networks with local structure exploitation modules on point clouds. Experimental results on the ModelNet [38], ScanNet [9] and ShapeNet [6] datasets show substantial improvements with the 3D point cloud classification and segmentation tasks compared to the PointNet++ pipeline.

2. Related Work

Deep Learning on Point Clouds: In recent years, deep learning methods have been employed to various 2D visual analysis tasks and have achieved outstanding performance [18, 10, 12]. However, these methods cannot be directly applied to 3D point clouds. While pixels regularly lie on the image plane for 2D images, the structure of 3D point cloud data is irregular so that some basic operations in CNN are not applicable. An intuitive idea to address the issue is to partition the 3D space into voxelized shapes [23, 38, 27, 9]. However, as 3D point cloud data is usually sparse, these methods suffer from low resolution and heavy computational cost. More recently, some deep learning structures have been especially designed to consume 3D point clouds as the input [26, 28, 25, 29, 39, 19, 20, 14, 2, 44, 40]. For example, Qi *et al.* [26] proposed a network architecture named PointNet by fusing point features into global representations with max pooling, which presented invariance to point-wise permutations. As PointNet fails to capture local structures which play critical roles for the success of convolutions, they further extended PointNet to PointNet++ by hierarchically grouping points to different levels for local feature extraction [28]. The following works such as self-organizing network (SO-Net) [19], similarity group proposal network (SGPN) [36] and PointCNN [20] also emphasized the importance of local structure exploitation for 3D point clouds. However, they fail to fully exploit the structural relations between local sub-clouds, which our humans largely rely on in 3D object understanding. Instead, the proposed SRN module aims to reason for such local interactions without any extra supervision signals.

Relational Reasoning: Relational reasoning aims to reason about the interactions between entities, which is a fundamental ability of humans. However, such relations are difficult for conventional network architectures to learn. More recently, relational modules have been carefully designed to solve the problems [32, 13, 37, 45, 34, 8, 16, 24]. For example, Santoro *et al.* [32] proposed a relational network (RN) for the task of visual question answering (VQA) and achieved outstanding performance. Hu *et al.* [13] presented an object relation module based on attention modules for object detection. Zhou *et al.* [45] designed a temporal relational network (TRN) to reason about the interactions between frames of videos in varying scales. While most of these methods aim to exploit the spatial or temporal relations in images and videos, very few works focus on relational reasoning for 3D data. Suwajanakorn *et al.* [34] proposed a KeypointNet for category-specific 3D keypoint extraction, which is the most relevant work in 3D point cloud reasoning. Although both methods are designed for 3D point cloud data, the goal of KeypointNet is to detect latent keypoints by reasoning for the relations between points and the categories. Our main contribution is the first attempt to reason about the structural relations of 3D objects, which are commonly-existed, important, but ignored by most existing methods. With the exploitation of structural relations, the models are able to understand 3D objects more comprehensively.

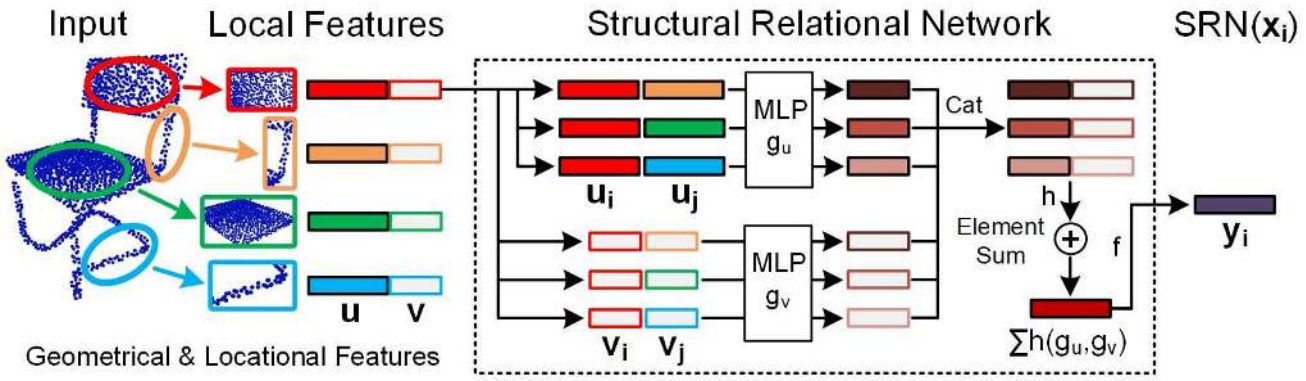

Figure 2. An illustration of the proposed structural relational network (SRN). For each input point cloud, we first obtain local sub-clouds to extract their geometrical features $\mathbf{u}$ and average coordinates $\mathbf{v}$, which are concatenated into local features $\mathbf{x}$. Then, we employ our SRN to capture the both geometrical and locational relations between the local sub-cloud P_i and other P_j. Lastly, we fuse the two kinds of relations to obtain the final structural relation $\mathbf{y}_i$ for P_i. In this figure, we utilize four different colors to represent varying local regions, and set the red sub-cloud as P_i for an easy illustration. (Best viewed in color.)

3. Proposed Approach

In this section, we first present the proposed structural relational network, which can be generally plugged into deep neural networks on 3D point cloud with local feature extraction modules. Then, we detail how to equip our SRN module to the widely-used PointNet++ architecture. Lastly, we highlight the differences with existing methods and introduce implementation details.

3.1. Structural Relational Network

Let P be a point set in which each point is represented by a 3D coordinate feature, where there are local sub-clouds P_i extracted from the holistic point cloud P. As there are only coordinates provided for point sets without further information, we extract the local geometrical feature $\mathbf{u}_i \in \mathbb{R}^d$ and average location $\mathbf{v}_i \in \mathbb{R}^3$ to describe each sub-cloud P_i, so that the features contain both geometrical and locational information of each local region.

Inspired by the relational reasoning modules for images and videos [32, 13, 45], we aim to reason about the structural relations between each P_i and other P_j. Both geometrical feature $\mathbf{u}$ and locational feature $\mathbf{v}$ of each local sub-cloud play crucial roles in structural interactions. For example, the repetitive local patterns are exploited through reasoning in geometry, and the linkage relations are captured through reasoning in location. Therefore, we define the structural interactions $\mathrm{SRN}(\mathbf{x}_i)$ between the ith local sub-cloud and others by jointly learning geometrical and locational relations:

$$\mathbf{y}_i = f\left(\sum_{\forall j} h\left(g_u\left(\mathbf{u}_i, \mathbf{u}_j\right), g_v\left(\mathbf{v}_i, \mathbf{v}_j\right)\right)\right), \qquad (1)$$

where both i and j are the indexes of local regions, f, g_u, g_v and h are functions, and $\mathbf{y}_i$ is the learned structural relational feature of P_i. In (1), the pairwise functions g_u and g_v aim to exploit the geometrical and locational relations between P_i and P_j, respectively, and then h fuses the two kinds of relations followed by an elementwise sum for all P_j. Lastly, we utilize the function f to obtain the final structural relations of P_i.

Figure 1 shows an illustration of the proposed SRN module, which aims to learn structural relations between P_i and other P_j. We learn the geometrical and locational interactions for structural relation exploitation, which are both important for 3D point cloud understanding. We follow [32] by firstly concatenating $\mathbf{u}_i$ and each $\mathbf{u}_j$, and $\mathbf{v}_i$ and each $\mathbf{v}_j$ to construct the input of g_u and g_v, respectively. The pairwise functions g_u and g_v capture the geometrical and locational relations between P_i and other P_j. Then, we utilize another pairwise function h to fuse the two kinds of relations for complete description, and sum up the results of varying P_j up to make (1) invariant to permutations. Lastly, we employ a function f to obtain the final relational representation $\mathbf{y}_i$. In SRN, we utilize multi-layer perceptrons (MLPs) to realize the functions g_u and g_v, with the parameters shared by all local sub-cloud pairs, respectively, and 1×1 convolutions for h and f, respectively. The learned $\mathbf{y}_i$ provides essential complementary information for local sub-cloud description. We utilize a residual block to sum up $\mathbf{y}_i$ and $\mathbf{u}_i$ (which share the same dimension), and then concatenate $\mathbf{v}_i$ as the final representation of the local sub-cloud P_i.

There are two key advantages of the proposed SRN module:

1) For each local sub-cloud P_i, the SRN module learns its structural interactions with all the local sub-clouds P_j. As point sets only contain point-wise coordinates

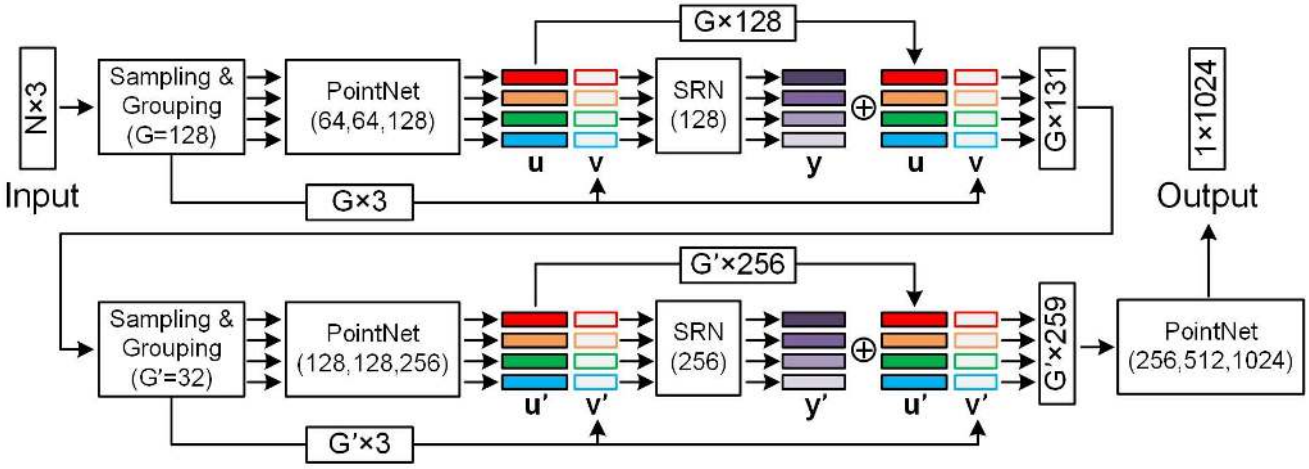

Figure 3. The network architecture of the SRN-equipped PointNet++. We first sample and group each input point cloud into G sub-clouds, following a PointNet module to extrate the local feature $\mathbf{u}$ for each sub-cloud. The PointNet module contains a multi-layer perceptron and a max-pooling layer for feature fusion. Then, we employ the proposed SRN module to obtain the relational features $\mathbf{y}$, which is summed up by $\mathbf{u}$ with a residual connection and concatenated by $\mathbf{v}$. We perform the procedure twice to extract multi-scale local features. Lastly, we employ a PointNet module to obtain the final output feature for the holistic point cloud. (Best viewed in color.)

without further information, SRN simultaneously considers the geometrical and locational relations between local regions, which both play crucial roles in 3D point cloud data. Compared with the existing local features for point clouds, SRN exploits additional information of structural relations as important complements.

2) SRN supports local sub-clouds in variable numbers and permutations. Even for the same point cloud, the number and permutation of local regions may be different due to varying sub-cloud grouping algorithms. With the shared parameters of the pairwise function g and the operation of element-wise sum, SRN is flexible with the input sizes by maintaining the same output size, and is invariant to varying permutations of local features.

3.2. SRN-PointNet++

The proposed SRN consumes the local geometrical features and average coordinates as the input, and learns the structural relational features as the output. As we simply employ the concatenation results of local features to replace the original ones, it can be generally applied to deep neural network architectures with local feature extraction modules for point clouds. PointNet++ [28] is a recent network architecture on 3D point clouds, which has achieved very promising performance. In this paper, we employ PointNet++ as a representative model to detail the equipment of SRN and analyze the experimental improvements.

Based on PointNet [26], PointNet++ utilizes an additional hierarchical architecture by sampling and grouping points for local structure exploitation. Rather than directly learning a global feature from the holistic point cloud, PointNet++ performs iterative farthest point sampling and ball query based grouping to obtain local sub-clouds at first, and then learns local features with PointNet individually. PointNet++ hierarchically learns local features through the three key layers including sampling layer, grouping layer and PointNet layer, and combine the features from multiple scales. Figure 3 shows the network structure of SRN-equipped PointNet++. Rather than simply utilizing local features $\mathbf{u}$ and $\mathbf{v}$ for representation, we learn the relational features $\mathbf{y}$ through SRN to provide essential complements, which is integrated with the original network with the residual block. For the point cloud classification task, we utilize the learned holistic feature for representation, and employ a 1024-512-256 fully connected layer followed by a softmax classifier. For the point cloud segmentation task, we follow [28] by employing a point feature propagation method to obtain point-wise scores for each point in the original point cloud.

3.3. Discussion

In this subsection, we compare our SRN with existing relational networks and local feature based methods to highlight the differences.

Difference with Existing Relational Networks: In recent years, several relational networks (RNs) have been proposed to reason about the interactions between entities [32, 13, 45]. They exploit spatial and semantic relations between objects for images [32, 13], or model temporal relations between frames in videos [45]. In our work, we focus on the structural relations between local sub-clouds for 3D point sets, which widely exist in real-world 3D object-

s and play important roles for our humans to understand them. In order to better capture structural relations, we simultaneously exploit geometrical and locational interactions of sub-clouds, which are both critical in 3D structure understanding. To our best knowledge, the proposed SRN is the first time to reason about structural relations for point cloud feature learning.

Difference with Local Feature Based Methods: Recent deep learning methods on 3D point cloud data demonstrate the importance of local regions [28, 39, 20]. In general, these methods employ hierarchical architectures to capture local structures in varying scales, and then integrate the local features for holistic representation. However, they independently process each local region and ignore the important structural relations thereby cannot completely understand the holistic 3D objects. Instead, our SRN captures both geometrical and positional relations to provide essential complement information for 3D local structure description, which is able to boost the performance of the existing network architecture for point clouds as a simple plug-and-play module.

3.4. Implementation Details

We used the Tensorflow [1] packages to construct our module throughout the experiments. For each 3D point cloud input, we randomly sampled 1,024 points from the 3D mesh to construct the point sets. The dimensions and other details of each layer are shown in Figure 3. In the first SRN module, we reasoned about the structural relations between each local region and the randomly-selected 32 local regions. In the second SRN module, we reason for all the sub-clouds for each sub-cloud. We did not perform data augmentation of point clouds, and fixed the dimension of the final representation as 1,024. We set the maximum training epoch number as 250 and the batchsize as 16 in the experiments. At the beginning of training, we empirically set the learning rate as 0.001 and the decay rate as 0.7 for every 200K steps.

4. Experiments

We conducted experiments on the widely-used 3D point cloud datasets to evaluate the proposed SRN module. More specifically, we first tested the performance of the SRN-equipped PointNet++ on point cloud classification and segmentation tasks, respectively, and compared with the state-of-the-art methods. We also designed cross-dataset evaluation to test the generalization ability. Then, we illustrated the effectiveness of SRN with ablation studies. Lastly, we visualized the t-SNE results and the learned structural relations between local sub-clouds for intuitive illustrations, and analyzed the key observations from the experiments. For fair comparisons, we utilized the same network structure of PointNet++ with our SRN-PointNet++ on all the datasets, with the only difference of the equipment of SRN modules.

Table 1. The comparison of the classification accuracy (%) with the state-of-the-art methods on the ModelNet40 and ScanNet datasets.

Method	ModelNet40	ScanNet
FPNN [21]	87.5	-
Vol. CNN [27]	89.9	74.9
O-CNN [35]	90.6	-
PointNet [26]	89.2	-
PointCNN [20]	**91.7**	77.9
PointNet++ [28]	90.6	77.1
SRN-PointNet++	**91.5**	**79.7**

Table 2. Cross-dataset evaluation of SRN-PointNet++ on ModelNet40 and ScanNet with classification accuracy (%).

Train / Test	ModelNet40	ScanNet
ModelNet40	91.5	75.6
ScanNet	86.5	79.7

4.1. Datasets

We employed three benchmark point cloud datasets for experimental evaluation, which included ModelNet40 [38], ScanNet [9] and ShapeNet [6]. We followed the standard evaluation protocols to test the performance.

1) The ModelNet40 dataset [38] contains 40 categories with 12,331 3D mesh models, which are split into 9,843 training models and 2,468 test models.

2) The ScanNet dataset [9] includes 1,513 scanned and reconstructed indoor scenes, and we employ 1,201 scenes as the training set and the rest 312 scenes as the test set.

3) The ShapeNet dataset [6] covers 55 common object categories with about 51,300 models, where we follow [41] by employing the ShapeNet Part dataset of 16 categories with 16,880 models. The models are divided into 14,006 training split and 2,874 test split, in which each point is associated with a point-wise label for the point cloud segmentation task.

4.2. Quantitative Results

In this subsection, we first evaluated the proposed SRN-PointNet++ on point cloud classification and segmentation tasks, respectively, and designed cross-dataset experiments. Then, we conducted ablation studies for more in-depth analysis of SRN.

Point Cloud Classification: Classification is usually considered as a touchstone task to test the effectiveness of

Table 3. Experimental comparison of the segmentation part IoU (%) with the recent methods on the ShapeNet Part dataset. In the table, the compared baseline methods are field probing neural networks [21], volumetric CNN [27], anisotropic CNN [4], SyncSpecCNN [42], PointNet [26], fully-convolutional point network [29] and PointNet++ [28], respectively.

Class	FPNN	Vol.	ACNN	SSCNN	PN	FCPN	PN++	SRN-PN++
Airplane	81.0	75.1	76.4	81.6	83.4	84.0	82.3	**82.4**
Bag	78.4	72.8	72.9	81.7	78.7	82.8	79.7	**79.8**
Cap	77.7	73.3	70.8	81.9	82.5	86.4	86.1	**88.1**
Car	75.7	70.0	72.7	75.2	74.9	88.3	78.2	77.9
Chair	87.6	87.2	86.1	90.2	89.6	83.3	90.5	**90.7**
EarPhone	61.9	63.5	71.1	74.9	73.0	73.6	73.7	69.6
Guitar	92.0	88.4	87.8	93.0	91.5	93.4	91.5	90.9
Knife	85.4	79.6	82.0	86.1	85.9	87.4	86.2	**86.3**
Lamp	82.5	74.4	77.4	84.7	80.8	77.4	83.6	**84.0**
Laptop	95.7	93.9	95.5	95.6	95.3	97.7	95.2	**95.4**
MotorBike	70.6	58.7	45.7	66.7	65.2	81.4	71.0	**72.2**
Mug	91.9	91.8	89.5	92.7	93.0	95.8	94.5	**94.9**
Pistol	85.9	76.4	77.4	81.6	81.2	87.7	80.8	**81.3**
Rocket	53.1	51.2	49.2	60.6	57.9	68.4	57.7	**62.1**
Skateboard	69.8	65.3	82.1	82.9	72.8	83.6	74.8	**75.9**
Table	75.3	77.1	76.7	82.1	80.6	73.4	82.8	**83.2**
Mean	81.4	79.4	79.6	84.7	83.7	84.0	85.1	**85.3**

the deep model, and the methods often perform well on other tasks if they achieve promising classification accuracies. We compared our SRN-PointNet++ with the state-of-the-art methods on the ModelNet40 [38] and ScanNet [9] datasets following the standard evaluation protocols.

Table 1 shows the classification results compared with the existing methods, where we share the same network architecture and hyperparameters of PointNet++ [28] and our SRN-PointNet++ for direct comparisons. We utilize the bold numbers to show that the performance of PointNet++ is improved equipped with SRN. In general, the point cloud data in ScanNet is more complicated than ModelNet40, as ScanNet contains indoor scenes and ModelNet40 includes 3D objects. While PointNet++ captures multi-scale local structures with hierarchical layers, it integrates local features with a simple max-pooling operation and fails to exploit their structural relations. Instead, the proposed SRN module explicitly reasons about the geometrical and locational relations of local structures, so that the learned deep model better understands the holistic structure of point clouds. The simple plug-and-play SRN module successfully boosts the performance of PointNet++ on both ModelNet40 and ScanNet datasets. More spcifically, we observe that the improvement for ScanNet is larger than ModelNet equipped with our SRN module. As the point clouds in ScanNet are more complicated with adaquate local structures, reasoning about the inner structural relations plays a more important role for more complete understanding. FPNN [21], Vol. CNN [27] and O-CNN [35] are volume-based deep learning methods which suffer from low resolution of 3D point cloud data. Our SRN-PointNet++ directly consumes point sets as inputs and outperforms these methods. PointCNN [20] designs a $\mathcal{X}$-convolution operation on 3D point cloud and achieves the state-of-the-art performance. Instead, the proposed SRN-PointNet++ model is able to obtain a comparable result on ModelNet40 and better performance on more complicated ScanNet by reasoning about the structural relations between local regions, where only multi-layer perceptrons and 1×1 convolutions are utilized for mapping rather than carefully-designed $\mathcal{X}$-convolution operation on 3D point cloud data.

As there are large differences between the point cloud data in ModelNet40 (3D objects) and ScanNet (indoor scenes), we also conducted cross-dataset experiments to test the generalization ability of the proposed SRN-PointNet++. For cross-dataset evaluation, we trained the network on one dataset and tested on the other dataset, extracting features with the learned model and employing linear SVM as classifier. Table 2 shows the experimental results of cross-dataset experiments on ModelNet40 and ScanNet. We observe that the performance drops if we chooses different training and test datasets due to the data discrepancy. However, the gaps are small and the cross-dataset experimental results are still comparable with existing recent methods. The cross-dataset experiments demonstrate the generalization ability of our SRN-PointNet++. Moreover, it shows that the learned structural relations are common for varying types of 3D point cloud data.

Table 4. The comparison of the classification accuracy (%) of SRN-PointNet++ under varying relations and integration methods on ModelNet40 (MN40).

Method	Relation	Integration	MN40
PointNet++	-	-	90.6
SRN-PointNet++	Geo	Concat	91.0
SRN-PointNet++	Geo	Res	90.7
SRN-PointNet++	Loc	Concat	91.1
SRN-PointNet++	Loc	Res	91.2
SRN-PointNet++	Geo + Loc	Concat	91.3
SRN-PointNet++	Geo + Loc	Res	**91.5**

Point Cloud Segmentation: The segmentation task is more challenging than classification as it requires further understanding of the point cloud data. We used the ShapeNet Part dataset [6] for comparison with existing methods on the point cloud segmentation task. We followed [42, 29] by employing part averaged IoU to evaluate the segmentation results, which calculated the weighted average of IoU for each category. Through the detailed comparisons on each 3D object, we are able to make complete observations of the proposed SRN-PointNet++.

Table 3 illustrates the experimental comparisons of point cloud segmentation on the ShapeNet Part dataset, where we also employ bold numbers to represent the improvement of SRN. Among the compared baseline methods, SyncSpecCNN [42] [11] is especially designed for the task of 3D semantic segmentation, while other methods are general 3D feature learning methods. We observe that the proposed SRN-PointNet++ obtains the state-of-the-art performance in the final result compared with existing methods on the task of point cloud segmentation. Equipped with our SRN module, SRN-PointNet++ outperforms PointNet++ in 13 out of 16 identities, which demonstrates the effectiveness of structural relational reasoning for 3D point cloud understanding. While SyncSpecCNN [42] learns synchronized spectral CNN for 3D data augmentation, SRN-PointNet++ achieves better performance on the point cloud segmentation task as a general point cloud analysis method. FCPN [29] is the most recent deep learning method for point clouds, which designs a fully-convolutional point network to process large-scale 3D data. Instead, our SRN-PointNet++ outperforms FCPN in the final result with the exploitation of structural relations by simple operations. Experimental results show the effectiveness of SRN-PointNet++ on the relatively hard point cloud segmentation task.

Ablation Study: Besides direct comparisons with the recent methods on benchmark datasets, we also conducted ablation experiments on ModelNet40 to further analyze the properties of SRN. In order to completely exploit structural relations between local regions, we simultaneously considered geometrical and locational interactions and we tested the influence of the relation types by only using each one of them. Moreover, we employed a residual block to aggregate local structural features and relations features in SRN-PointNet++, which we compared with another widely-used fusion method of concatenation.

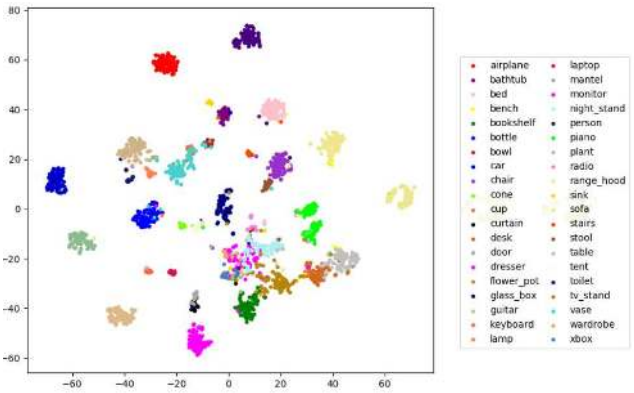

Figure 4. Visualization results of t-SNE on the ModelNet40 dataset for SRN-PointNet++.

Table 4 shows the classification accuracies on ModelNet40 of the ablation study. In the table, Geo and Loc are geometrical and locational relations, respectively. We compared the performance of SRN-PointNet++ by exploiting each or both of them. Concat and Res represent to fuse the structural features with concatenation ($[\mathbf{y}, \mathbf{u}, \mathbf{v}]$) and residual blocks ($[\mathbf{y} + \mathbf{u}, \mathbf{v}]$). We observe that capturing either geometrical or locational relation is able to boost the performance of PointNet++, which is close to our humans when understanding 3D objects as we may notice both repetitive structures and linkages of parts. The best performance is obtained by simultaneously considering both relations. Concatenation and residual blocks are popular methods for feature combination. Through the detailed comparisons, we observe that the performance gap is small for the two aggregation approaches, which shows the adaptivity of the learned relational features. As the residual blocks do not require to increase the feature length, we employ residual blocks throughout the experiments for computational cost reduction.

4.3. Qualitative Results

In this subsection, we first visualized the t-SNE results of the proposed SRN-PointNet++. Figure 4 shows the visualization results. We observe that most classes are separated with small intra-class variations, which demonstrates the discriminative power of SRN-PointNet++. For more detailed analysis, we look into the most mixed areas to find the categories which have larger difficulties to be correctly classified. We discover that the most confusing cate-

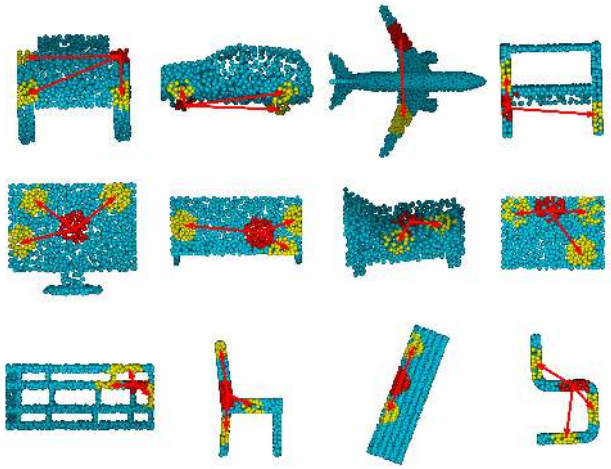

Figure 5. The visualization results of local regions with high responses. In the figure, different colors represent varying local regions. (Best viewed in color.)

gories for SRN-PointNet++ are: 1) night stand, 2) dresser, 3) wardrobe, and 4) x-box. These objects share similar 3D structures, which are also relatively simple compared with other classes such as airplane and car. On one hand, their inter-class distances are small which leads to large difficulty in feature learning and classification. On the other hand, our SRN module captures similar structural relations for these objects and results in similar encoding due to their simple and regular structures. Therefore, these classes are relatively easier to be misclassified by SRN-PointNet++ compared with other classes.

We also visualized the learned structural relations with the high responses of the second SRN in our SRN-PointNet++. Figure 5 shows the visualization results on the ModelNet40 dataset. We observe that highly related local structures are captured with both geometrical and locational interactions, such as symmetric repetitive parts (the first row), local regions sharing the same plane (the second row) and essential linkages of varying parts (the third row). Although the point clouds from different classes vary largely, our SRN module is still able to exploit the common and underlying relations between local structures. It should be noted that we do not utilize additional labels or carefully-designed loss functions to train such structural relations. Instead, these relations are captured with the goal of better understanding the semantics. In other words, SRN discovers similar local interactions with our humans when learning to understand the objects, which also demonstrates the importance of exploiting structural relations in 3D object analysis.

4.4. Analysis

The above experiments on benchmark datasets suggest the following four key observations:

1) The proposed SRN module successfully boosts the performance of PointNet++ on both classification and segmentation tasks, which shows the effectiveness of structural relational reasoning in 3D point cloud data. Moreover, the improvement is more significant for complicated point cloud data with adequate local structures.

2) Cross-dataset evaluations on ModelNet40 and ScanNet show that our SRN-PointNet++ presents strong generalization ability and captures common structural relations despite of data discrepancy.

3) Ablation studies show that both geometrical and locational interactions are important to describe structural relations between local regions. The best result is achieved when simultaneously exploiting both relations. Also, we show that different aggregation methods of the local structural features and relational features do not largely affect the performance of SRN-PointNet++.

4) While supervision signals only provide the categories of the point clouds, visualization results illustrate that our SRN module is able to capture highly relevant local structures without specific labels.

5. Conclusion

In this paper, we have proposed a simple and plug-and-play module named SRN to reason about structural relations between local regions for 3D point clouds, which play an important role for our humans to analyze 3D objects. While most existing methods aggregate local features by a simple pooling operation thereby ignoring the important structural interactions, our SRN explicitly captures their geometrical and locational relations to better understand the holistic structures. The proposed SRN module can be equipped with the existing models, where we detail the SRN-PointNet++ architecture as a representative method. Experimental results on benchmark datasets demonstrate that our SRN successfully boosts the performance of the original network on point cloud classification and segmentation tasks. Ablation studies and visualization results also show that our SRN module captures essential structural relations in geometry and location.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant U1813218, Grant 61822603, Grant U1713214, Grant 61672306, and Grant 61572271. The authors would like to thank Mr. Chaojian Li and Mr. Haidong Zhu for valuable discussions.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A. Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B. Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *arXiv*, abs/1603.04467, 2016. 5

[2] Mikaela Angelina Uy and Gim Hee Lee. PointNetVLAD: Deep point cloud based retrieval for large-scale place recognition. In *CVPR*, pages 4470–4479, 2018. 2

[3] Mathieu Aubry, Ulrich Schlickewei, and Daniel Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. In *ICCVW*, pages 1626–1633, 2011. 1

[4] Davide Boscaini, Jonathan Masci, Emanuele Rodolà, and Michael Bronstein. Learning shape correspondence with anisotropic convolutional neural networks. In *NIPS*, pages 3189–3197, 2016. 6

[5] Michael M Bronstein and Iasonas Kokkinos. Scale-invariant heat kernel signatures for non-rigid shape recognition. In *CVPR*, pages 1704–1711, 2010. 1

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, and Hao Su. Shapenet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2, 5, 7

[7] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3D model retrieval. In *Computer graphics forum*, pages 223–232, 2003. 1

[8] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *CVPR*, pages 7239–7248, 2018. 2

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 2, 5, 6

[10] Ross Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015. 2

[11] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3D semantic segmentation with submanifold sparse convolutional networks. *CVPR*, pages 9224–9232, 2018. 7

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[13] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, pages 3588–3597, 2018. 2, 3, 4

[14] Zi Jian Yew and Gim Hee Lee. 3DFeat-Net: Weakly supervised local 3D features for point cloud registration. In *ECCV*, pages 607–623, 2018. 2

[15] Andrew E Johnson and Martial Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *TPAMI*, pages 433–449, 1999. 1

[16] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross B Girshick. Inferring and executing programs for visual reasoning. In *ICCV*, pages 3008–3017, 2017. 2

[17] Roman Klokov and Victor Lempitsky. Escape from cells: Deep kd-networks for the recognition of 3D point cloud models. In *ICCV*, pages 863–872, 2017. 1

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 2

[19] Jiaxin Li, Ben M Chen, and Gim Hee Lee. SO-Net: Self-organizing network for point cloud analysis. In *CVPR*, pages 9397–9406, 2018. 2

[20] Yangyan Li, Rui Bu, Mingchao Sun, and Baoquan Chen. PointCNN. *arXiv preprint arXiv:1801.07791*, 2018. 2, 5, 6

[21] Yangyan Li, Soeren Pirk, Hao Su, Charles R Qi, and Leonidas J Guibas. FPNN: Field probing neural networks for 3D data. In *NIPS*, pages 307–315, 2016. 5, 6

[22] Haibin Ling and David W Jacobs. Shape classification using the inner-distance. *TPAMI*, 29(2):286–299, 2007. 1

[23] Daniel Maturana and Sebastian Scherer. VoxNet: A 3D convolutional neural network for real-time object recognition. In *IROS*, pages 922–928, 2015. 2

[24] Rasmus Berg Palm, Ulrich Paquet, and Ole Winther. Recurrent relational networks for complex relational reasoning. *arXiv preprint arXiv:1711.08028*, 2017. 2

[25] 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*, pages 918–927, 2018. 1, 2

[26] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, pages 652–660, 2017. 1, 2, 4, 5, 6

[27] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view CNNs for object classification on 3D data. In *CVPR*, pages 5648–5656, 2016. 2, 5, 6

[28] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, pages 5099–5108, 2017. 1, 2, 4, 5, 6

[29] Dario Rethage, Johanna Wald, Jurgen Sturm, Nassir Navab, and Federico Tombari. Fully-convolutional point networks for large-scale point clouds. In *ECCV*, pages 569–611, 2018. 2, 6, 7

[30] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *ICRA*, pages 3212–3217, 2009. 1

[31] Radu Bogdan Rusu, Nico Blodow, Zoltan Csaba Marton, and Michael Beetz. Aligning point cloud views using persistent feature histograms. In *IROS*, pages 3384–3391, 2008. 1

[32] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lilli-

crap. A simple neural network module for relational reasoning. In *NIPS*, pages 4967–4976, 2017. 2, 3, 4

[33] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, pages 1383–1392, 2009. 1

[34] Supasorn Suwajanakorn, Noah Snavely, Jonathan Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. In *NIPS*, 2018. 2

[35] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-CNN: Octree-based convolutional neural networks for 3D shape analysis. *TOG*, 36(4):72, 2017. 5, 6

[36] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. SGPN: Similarity group proposal network for 3D point cloud instance segmentation. In *CVPR*, pages 2569–2578, 2018. 2

[37] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 2

[38] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920, 2015. 2, 5, 6

[39] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. FoldingNet: Point cloud auto-encoder via deep grid deformation. In *CVPR*, pages 206–215, 2018. 1, 2, 5

[40] Xiaoqing Ye, Jiamao Li, Hexiao Huang, Liang Du, and Xiaolin Zhang. 3D recurrent neural networks with context fusion for point cloud semantic segmentation. In *ECCV*, pages 415–430, 2018. 2

[41] Li Yi, Vladimir G Kim, Duygu Ceylan, I Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3D shape collections. *TOG*, 35(6):210, 2016. 5

[42] Li Yi, Hao Su, Xingwen Guo, and Leonidas J Guibas. SyncSpecCNN: Synchronized spectral CNN for 3D shape segmentation. In *CVPR*, pages 6584–6592, 2017. 6, 7

[43] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. PU-Net: Point cloud upsampling network. In *CVPR*, pages 2790–2799, 2018. 1

[44] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. PCN: Point completion network. *arXiv preprint arXiv:1808.00671*, 2018. 2

[45] Bolei Zhou, Alex Andonian, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, pages 803–818, 2018. 2, 3, 4