SPCR: SEMI-SUPERVISED POINT CLOUD INSTANCE SEGMENTATION WITH PERTURBATION CONSISTENCY REGULARIZATION

Yongbin Liao[1], *Hongyuan Zhu*[2], *Tao Chen*[1*], *Jiayuan Fan*[3]

[1] School of Information Science and Technology, Fudan University
[2] Agency for Science, Technology and Research
[3] Academy for Engineering and Technology, Fudan University

2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506359

ABSTRACT

Point cloud instance segmentation is steadily improving with the development of deep learning. However, current progress is hindered by the expensive cost of collecting dense point cloud labels. To this end, we propose the first semi-supervised point cloud instance segmentation architecture, which is called semi-supervised point cloud instance segmentation with perturbation consistency regularization (SPCR). It is capable to alleviate the data-hungry bottleneck of existing strongly supervised methods. Specifically, SPCR enforces an invariance of the predictions over different perturbations applied to the input point clouds. We firstly introduce various perturbation schemes on inputs to force the network to be robust and easily generalized to the unseen and unlabeled data. Further, perturbation consistency regularization is then conducted on predicted instance masks from various transformed inputs to provide self-supervision for network learning. Extensive experiments on the challenging ScanNet v2 dataset demonstrate our method can achieve competitive performance compared with the state-of-the-art of fully supervised methods.

Index Terms— Point cloud, Instance segmentation, Semi-supervised learning

1. INTRODUCTION

With the increasing availability and affordability of depth sensors, 3D scene understanding has attracted numerous attention and has been widely applied in many applications such as virtual reality and autonomous driving. Point cloud instance segmentation is one of the fundamental and challenging tasks in 3D computer vision that requires to simultaneously predict semantic and instance labels for each object in one scene.

In recent years, many deep learning based methods [1–12] for point cloud instance segmentation have emerged and boosted the performance in a large margin. These methods could be divided into two categories. One way is to directly group the input point cloud into object instances according to the learned point embeddings. SGPN [2] groups points according to the similarity matrix measured by the semantic predictions. ASIS [7] and BAN [8] leverage discriminative loss to separate different objects based on the learned feature embeddings. Occuseg [4] introduces an occupancy signal to complement point embeddings for instance clustering. The other category is first to generate instance candidates and further mine precise instance contours within candidate regions. GSPN [3] predicts instances according to the object proposals produced by its proposed generative shape proposal network. 3D-BoNet [5] directly generates bounding box proposals and foreground instance masks simultaneously. 3D-MPA [6] outputs the final predictions by aggregating multi-proposals produced through object center voting. PointGroup [1] conducts its proposed clustering algorithm on both original point set and offset-shifted point coordinate set for instance prediction.

However, most of the existing methods for point cloud instance segmentation are fully supervised and severely rely on dense point-level annotations, which are always costly for collection. Therefore, the applications of these methods are limited in the real scenarios. There are a lot of methods trying to tackle this problem. Semi-supervised learning is a promising choice which requires only few labeled data.

Many efforts have been made to adopt semi-supervised learning on 2D images. For example, The Π-model [13] encourages a consistency over two different perturbations applied to one input image. Mean-teacher [14] enforces similar predictions of student network and teacher network whose weights are transferred from the student. VAT [15] improves the prediction by approximating the perturbations which influence model's results the most. FixMatch [16] demonstrates the importance of strong and varied perturbations. Inspired by the recent success of semi-supervised learning on 2D images, there are also some attempts for 3D scene understanding. Tang *et al*. [17] propose a transferable semi-supervised 3D object detection from RGB-D input through cross-category learning which requires 2D labels for all object classes. SESS [18] takes pure point cloud as input and constructs a semi-supervised architecture leveraging mean-teacher network. However, these methods merely pay attention to 3D object detection. There is no consideration taken into point cloud in-

* Tao Chen is the corresponding author

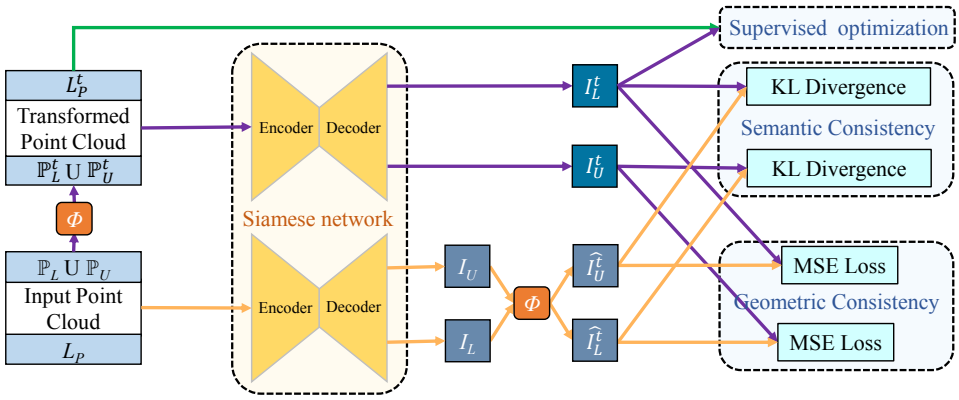

Fig. 1. **The network architecture of our proposed SPCR**. Given the original input point cloud (either labeled or unlabeled) and their transformed point cloud, we first generate instance candidates through the siamese network. The predictions of the original point cloud are then transformed with the same perturbation Φ. Finally, these instance predictions are optimized by the supervised loss and our proposed perturbation consistency regularization mechanism.

stance segmentation, a more challenging task which requires more detailed point-level annotations.

Considering the promising potential of semi-supervised learning on point cloud instance segmentation task, we propose SPCR, a self-supervised perturbation consistency regularization mechanism for semi-supervised point cloud instance segmentation. The perturbation consistency regularization is to enforce an invariance of the predictions over some perturbations applied to the input point cloud. As a result, we do not need a large amount of well-annotated training samples since unlabeled data could provide self-supervision through perturbation consistency themselves, which largely reducing the cost for data annotations. Specifically, we first propose multiple perturbation schemes for the input point cloud to learn the underlying knowledge of unlabeled data to full advantage. Furthermore, our perturbation consistency regularization will guide the model to be consistent with its predictions under different random perturbations. We propose two consistency terms consisting of both geometric and semantic information for better prediction-invariant constraint under different perturbations. With extensive experiments, we obtain competitive results compared with recent fully supervised methods and demonstrate the effectiveness of our proposed perturbation consistency regularization mechanism in a semi-supervised setting.

2. METHOD

2.1. SPCR Architecture

The overall architecture of our proposed SPCR is depicted in Fig. 1. We introduce the implementation of semi-supervised point cloud instance segmentation by a shared-weight siamese network which is composed of the state-of-the-art PointGroup [1]. To be specific, We take a mixture of labeled and unlabeled point clouds as our input denoted as $\{\mathbb{P}_L \cup \mathbb{P}_U\}$, where $\mathbb{P}_L$ and $\mathbb{P}_U$ represent the labeled and unlabeled point cloud respectively. To extract additional training signal for self-supervision, we conduct a perturbation Φ on the original point cloud to obtain transformed point cloud denoted as $\{\mathbb{P}_L^t \cup \mathbb{P}_U^t\}$. After that, the original point cloud and transformed point cloud are passed to the siamese network simultaneously for instance prediction. The output instance candidates are represented by $\{I_L, I_U\}$ and $\{I_L^t, I_U^t\}$ respectively. $\{I_L, I_U\}$ are further transformed to $\left\{\hat{I}_L, \hat{I}_U\right\}$ by the same perturbation conducted on the original point cloud. For instance candidates I_L^t, we optimize them with transformed ground truths L_P^t through supervised loss defined in PointGroup. For other predictions, we propose a perturbation consistency regularization to constrain the pair-wise instance candidates to be consistent both in semantic categories and geometric properties. Details of the perturbation scheme and perturbation consistency regularization mechanism will be described in the next two sub-sections.

2.2. Perturbation Scheme

An important factor for consistency regularization is the perturbations applied to the input point cloud. We propose three types of perturbation consisting of random jitter, flipping and rotation to prevent the labeled data from overfitting and leverage the unlabeled data for self-supervised learning. At first, we initial the perturbation matrix $\mathbf{M}$ of Φ with an identity matrix of shape 3×3. For jitter, the perturbation matrix will be added with a random matrix of the same shape as $\mathbf{M}$. For flipping, we update $\mathbf{M}$ by multiplying its first element with a random variable selected from $\{1, -1\}$, where -1 means flipping along y-axis and 1 means no flipping. For rotation, we firstly generate a rotation angel formulated as $\theta = 2\pi\delta$ where δ is a random variable and further define the corresponding rotation matrix as

$$R(\theta) = \begin{bmatrix} cos(\theta) & sin(\theta) & 0 \\ -sin(\theta) & cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

Method	Metric	10%	20%	30%	40%	50%	70%	100%
PointGroup [1]	mAP	13.4	20.8	25.8	27.6	29.6	31.6	33.7
	mAP@50%	27.3	40.3	47.1	48.8	50.3	51.7	54.8
	mAP@25%	42.1	55.6	63.5	65.2	67.0	68.8	70.6
Ours	mAP	19.6	27.1	29.3	31.2	31.7	32.8	35.4
	mAP@50%	35.6	47.8	50.7	52.9	53.9	54.9	57.5
	mAP@25%	51.4	62.3	66.5	68.1	69.1	70.1	71.6
Improvements	mAP	6.2	6.3	3.5	3.6	2.1	1.2	1.7
	mAP@50%	8.3	**7.5**	**3.6**	**4.1**	**3.6**	**3.2**	**2.7**
	mAP@25%	**9.3**	6.7	3.0	2.9	2.1	1.3	1.0

Table 1. Comparison of inductive learning with PointGroup on ScanNet v2 val set. Absolute improvements between Ours and PointGroup are reported in terms of mAP, mAP@50% and mAP@25% respectively.

After that, we will adjust **M** by multiplying it with rotation matrix $R(\theta)$ and thus get the final perturbation matrix.

With the established perturbation matrix **M** computed via the above random perturbation scheme, a training batch with a mixture of labeled and unlabeled samples will be transformed by multiplying it with **M**. In particular, the point-level labels L_P of labeled samples $\mathbb{P}_L$ are also transformed by the same perturbation before supervised optimization. To ensure the validity of consistency regularization, the instance predictions $\{I_L, I_U\}$ of original inputs $\{\mathbb{P}_L \cup \mathbb{P}_U\}$ are transformed as well.

2.3. Consistency Regularization

As stated above, in order to provide additional self-supervision for network training, we rely on enforcing a consistency regularization between the predictions under different perturbations. It is obvious that the consistency of predictions means being consistent both in semantic categories and geometric properties. To this end, we propose two terms of consistency regularization for semi-supervised point cloud instance segmentation and define the consistency regularization loss as

$$L_{CR} = \lambda_1 L_{semantic} + \lambda_2 L_{geometric} \tag{2}$$

where the semantic term $L_{semantic}$ enables consistency regularization by enforcing the similar predictions of semantic categories and the geometric term imposes the structural constraint on instance candidates for maintaining consistent geometric properties. λ_1 and λ_2 are the weights to control the importance for each term.

1) Semantic Consistency Term. For predictions $\{I_L^t, I_U^t\}$ and $\left\{\hat{I_L^t}, \hat{I_U^t}\right\}$ of the original and transformed input point clouds, let $\{P_L^t, P_U^t\}$ and $\left\{\hat{P_L^t}, \hat{P_U^t}\right\}$ represent the semantic probabilities of points on these instance candidates respectively. We define the semantic consistency regularization term as the KL-divergence:

$$L_{semantic} = \frac{\sum D_{KL}(p_L^t \parallel \hat{p_L^t}) + \sum D_{KL}(p_U^t \parallel \hat{p_U^t})}{\left|\hat{P_L^t}\right| + \left|\hat{P_U^t}\right|} \tag{3}$$

2) Geometric Consistency Term. Using the semantic term alone is to merely constrain the predicted semantic class of instance candidates, while ignore another import constraint of geometric information. Similarly, we denote the prediction of per-point offset to its instance center as $\{O_L^t, O_U^t\}$ and $\left\{\hat{O_L^t}, \hat{O_U^t}\right\}$ respectively. The geometric consistency regularization are then formulated as

$$L_{geometric} = \frac{\sum(o_L^t - \hat{o_L^t}) + \sum(o_U^t - \hat{o_U^t})}{\left|\hat{O_L^t}\right| + \left|\hat{O_U^t}\right|} \tag{4}$$

3. EXPERIMENTS

3.1. Dataset and Evaluation Criteria

Dataset. We evaluate our semi-supervised point cloud instance segmentation method on ScanNet v2 [19] dataset. The official dataset separation has 1201 scans for training, 312 scans for validation and 100 scans for testing. Point-level semantic-instance labels are well annotated for each scene of the training and validation sets.

Evaluation Criteria. Following the common experimental protocol for point cloud instance segmentation, we adopt the same 18 object classes for evaluation as reported in [1]. We use the mean average precision at overlap 25% (mAP@25%), overlap 50% (mAP@50%) and overlaps in the range [0.5 : 0.95 : 0.05] (mAP) as our evaluation criteria as defined in the ScanNet benchmark.

3.2. Implementation Details

Network. We adopt PointGroup [1] as the structure of our siamese network since it is the state-of-the-art method for point cloud instance segmentation. The inputs of our network are batches of point clouds with a mixture of labeled and unlabeled samples whose batch size is set to the same number of 2. Following the procedure of [1], the maximum point number of the input point clouds is limited to 250,000 for efficient network training. For weights of each consistency term, we set $\lambda_1 = 2$ and $\lambda_2 = 1$ empirically.

Training. With the available labeled point cloud, we pre-train PointGroup for 384 epochs with batch size 8. After that, we

Method	Metric	10%	20%	30%	40%	50%	70%
PointGroup [1]	mAP	16.0	24.6	30.1	31.9	35.5	36.9
	mAP@50%	32.9	44.9	52.6	55.2	59.2	61.2
	mAP@25%	47.2	61.9	69.4	70.0	73.9	74.6
Ours	mAP	22.3	29.8	32.2	34.3	36.4	37.4
	mAP@50%	39.9	52.5	55.1	57.9	60.3	61.8
	mAP@25%	53.4	65.8	70.3	71.2	74.5	75.7
Improvements	mAP	6.3	5.2	2.1	2.4	0.9	0.5
	mAP@50%	**7.0**	**7.6**	**2.5**	**2.7**	**1.1**	0.6
	mAP@25%	6.2	3.9	0.9	1.2	0.6	**1.1**

Table 2. Comparison of transductive learning with PointGroup on ScanNet v2 val set. Absolute improvements between Ours and PointGroup are reported in terms of mAP, mAP@50% and mAP@25% respectively.

Method	mAP	mAP@50%	mAP@25%
Baseline	13.4	27.3	42.1
SCR	15.4	33.2	47.3
GCR	16.3	32.8	48.7
Both SCR & GCR	**19.6**	**35.6**	**51.4**

Table 3. Ablation study of different terms of perturbation consistency regularization on ScanNet v2 validation set.

initial the siamese network with the pre-trained weights, and train our SPCR network on both the labeled and unlabeled data for another 128 epochs. We take Adam as the optimizer with an initial learning rate of 0.001 for the pre-train stage, which is decreased by 10 for fine-tuning of our SPCR.

Inference. During inference, we forward the point clouds of entire scenes as inputs to the siamese network to produce the instance candidates. we then post-process these predicted instance candidates by a 3D NMS module with an Intersection-over-Union (IoU) threshold of 0.3. Note that the IoU here is computed according to instances rather than bounding boxes.

3.3. Comparison with Fully-supervised Methods

To the best of our knowledge, our method is the first semi-supervised point cloud instance segmentation network. So we quantitatively compare our SPCR with the state-of-the-art fully-supervised PointGroup [1] to validate the effectiveness of our proposed framework. Specifically, we select various ratios of labeled data from the entire training set of ScanNet v2 dataset and train PointGroup in a fully-supervised way under these different ratio settings. For our proposed SPCR, we train it in a semi-supervised way with both the labeled and unlabeled samples for each label ratio setting.

Inductive Learning. For inductive learning, we conduct the comparison against PointGroup on the unseen validation set. As listed in Table 1, it is shown that our SPCR outperforms PointGroup under all the label ratio settings for three different evaluation metrics. Given 10% labeled data, our network obtains 6.2%, 8.3%, and 9.3% absolute improvements for mAP, mAP@50% and mAP@25% respectively, which demonstrates that our SPCR is able to learn from unlabeled data especially when the labeled data is rather limited.

Additionally, SPCR can achieve comparable performance when the label ratio is set as 50% compared with the fully-supervised PointGroup. Moreover, we are also able to further improve the performance for fully supervised setting with 100% labeled data. They indicate that our proposed self-supervised perturbation consistency regularization is indeed effective.

Transductive Learning. For transductive learning, we compare our network with PointGroup on the given unlabeled data and the results are reported in Table 2. we can see similar improvements as demonstrated in inductive learning, which further validate the superiority of our method.

3.4. Ablation Studies

In this section, we analyze the contribution of each term of proposed perturbation consistency regularization, including semantic consistency regularization (SCR) and geometric consistency regularization (GCR).

We conduct the ablation studies on ScanNet v2 with 10% labeled data and the comparison is reported in Table 3. It is obvious that the performance always improves either by adding semantic consistency regularization or geometric consistency regularization. Besides, the geometric consistency term has more contributions than the semantic consistency term. Finally, the combination of the two consistency terms produces better performance than their terms solely, which proves that all terms of our proposed perturbations consistency regularization indeed boost the process of semi-supervised point cloud instance segmentation.

4. CONCLUSION

In this paper, we propose the first semi-supervised point cloud instance segmentation architecture with self-supervised perturbation consistency regularization. We implement our semi-supervised framework by a shared-weighted siamese network and propose a perturbation consistency regularization mechanism to provide self-supervision for network learning from unlabeled data. Our SPCR network does not require a large number of well-annotated training data, but achieving competitive results compared with its fully-supervised counterpart. Extensive experiments validate the effectiveness and superiority of our SPCR and demonstrate that semi-supervised learning is a promising learning paradigm to solve the data-hungry problem of point cloud instance segmentation.

Acknowledgement. This work is supported by National Natural Science Foundation of China (No. 62071127 and U1909207), Shanghai Pujiang Program (No.19PJ1402000), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), Shanghai Engineering Research Center of AI Robotics and Engineering Research Center of AI Robotics, Ministry of Education in China and the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project A18A2b0046).