Pattern Recognition 122 (2022) 108250

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

JSPNet: Learning joint semantic & instance segmentation of point clouds via feature self-similarity and cross-task probability

Feng Chen[a], Fei Wu[a,*], Guangwei Gao[b,*], Yimu Ji[c,d], Jing Xu[e], Guo-Ping Jiang[a], Xiao-Yuan Jing[a,f]

[a] *College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China*
[b] *Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo, Japan*
[c] *Nanjing Center of HPC China, Nanjing, China*
[d] *Jiangsu HPC and Intelligent Processing Engineer Research Center, Nanjing, China*
[e] *School of Law, Hohai University, Nanjing, China*
[f] *School of Computer Science, Wuhan University, Wuhan, China*

ARTICLE INFO

Article history:
Received 12 September 2020
Revised 9 August 2021
Accepted 11 August 2021
Available online 13 August 2021

Keywords:
Instance & semantic segmentation
Point could processing
Multi-task learning

ABSTRACT

In this paper, we propose a novel method named JSPNet, to segment 3D point cloud in semantic and instance simultaneously. First, we analyze the problem in addressing joint semantic and instance segmentation, including the common ground of cooperation of two tasks, conflict of two tasks, quadruplet relation between semantic and instance distributions, and ignorance of existing works. Then we introduce our method to reinforce mutual cooperation and alleviate the essential conflict. Our method has a shared encoder and two decoders to address two tasks. Specifically, to maintain discriminative features and characterize inconspicuous content, a similarity-based feature fusion module is designed to locate the inconspicuous area in the feature of current branch and then select related features from the other branch to compensate for the unclear content. Furthermore, given the salient semantic feature and the salient instance feature, a cross-task probability-based feature fusion module is developed to establish the probabilistic correlation between semantic and instance features. This module could transform features from one branch and further fuse them with the other branch by multiplying probabilistic matrix. Experimental results on a large-scale 3D indoor point cloud dataset S3DIS and a part-segmentation dataset ShapeNet have demonstrated the superiority of our method over existing state-of-the-arts in both semantic and instance segmentation. The proposed method outperforms PointNet with 12% and 26% improvements and outperforms ASIS with 2.7% and 4.3% improvements in terms of mIoU and mPre. Code of this work has been made available at https://github.com/Chenfeng1271/JSPNet.

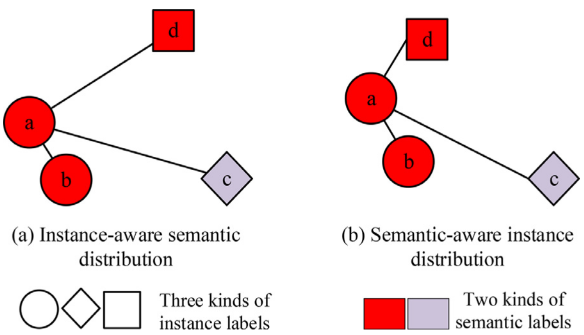
1. Introduction

Both semantic and instance segmentation [1,2] are fundamental and challenging tasks in computer vision, whose goals are to segment and classify a unit with a predefined category label and individual instance identity respectively. Recently, numerous deep learning based approaches [3,4] have achieved breakthrough progress in these two tasks. Then, with the growing availability of 3D scene data [5,6], point cloud semantic and instance segmentation also attract growing researcher's attention [7]. Inspired by the common ground of these two tasks, some recent works [8,9] focus on joint semantic and instance segmentation training. However, these works ignore to avoid the drawbacks brought by the conflict between semantic and instance segmentation.

In previous works, JSIS3D [10] proposed a multi-task network and multi-value Conditional Random Field (CRF) to handle joint segmentation tasks. However, this method is not an end-to-end framework and hard to constrain feature combinations of two kinds of segmentation. Besides, ASIS [8] designed an end-to-end two-branch framework to associate instance segmentation and semantic segmentation closely together. JSNet [9] then further introduced a novel fusion module to make instance segmentation and semantic segmentation mutually promote. However, these works only focus on the intuition of common ground of cooperation with each other by latent feature aggregation, while the conflict of them is not well analyzed. To our best knowledge, the conflict in addressing instance and semantic segmentation is not well studied. Therefore, in this work, we first make a problem statement of joint

[*] Corresponding authors.
E-mail addresses: wufei_8888@126.com (F. Wu), csggao@gmail.com (G. Gao).

https://doi.org/10.1016/j.patcog.2021.108250
0031-3203/© 2021 Elsevier Ltd. All rights reserved.



(a) Instance-aware semantic distribution

(b) Semantic-aware instance distribution

Fig. 1. Illustration of quadruplet relation in instance-aware semantic feature distribution and semantic-aware instance feature distribution.

training of semantic and instance segmentation, including common ground of mutual promotion, conflict between two tasks and ignorance of previous methods. Then, we propose a novel method, JSPNet, to address **J**oint semantic and instance segmentation simultaneously via exploiting feature **S**imilarity and cross-task **P**robability.

1.1. Problem statement of joint semantic and instance segmentation

Let $P = \{p_1, p_2, \ldots, p_N\}$ as the set of a point cloud scene where N is the total number of points, each points p_n is assigned a semantic label $s_n \in S$ and instance label $i_n \in I$. S and I are semantic category and instance identity respectively.

1.1.1. Common ground of mutual promotion

Semantic and instance segmentation are positively related in two cases.

- Points belonging to different semantic classes must belong to different instances. Given semantic labels s_m and s_n of two points, if $s_m \neq s_n$, i_m will be not equal to i_n.
- Points belonging to the same instance must belong to the same semantic class. Given instance labels i_m and i_n of two points, if $i_m = i_n$, s_m will be equal to s_n.

The benefits of mutual promotion include 1) correcting inaccurate feature prediction, 2) clarifying fuzzy feature prediction and 3) verifying accurate prediction.

1.1.2. Conflict of joint segmentation

Joint semantic and instance segmentation would also meet ambiguity in two cases.

- For points belonging to the same semantic class, it is unknown whether they belong to the same instance, i.e., the relationship between instance labels i_m and i_n of two points can't be determined by their same semantic labels s.
- For points belonging to different instances, it is also unknown whether they belong to the same semantic class, i.e., the relationship between semantic labels s_m and s_n of two points can't be inferred by their different instance labels i_m and i_n.

1.1.3. Quadruplet relation in joint semantic & instance segmentation

Notably, the condition that points sharing the same instance label belong to the different semantic labels would never exist. We analyze the quadruplet relation in practical learning, including instance-to-semantic (ins2sem) case and semantic-to-instance (sem2ins) case. As shown in Fig. 1, the quadruplet units have an anchor unit (point a with s_a and i_a), and three units (point b with s_a and i_a; point c with $s_{\bar{a}}$ and $i_{\bar{a}}$ where $\bar{a}$ denotes the label different from that of a; point d with s_a and $i_{\bar{a}}$). Namely, points a, b, d have the same semantic label, and d has another semantic label. Besides, a and b have the same instance label, while is different from c and d.

In instance-aware semantic distribution, points a and b would compact the intra-semantic distribution; points a and c would keep inter-semantic distribution away; However, the inter-instance distinctiveness between points a and d would disturb compactness of intra-semantic distribution. Moreover, in semantic-aware instance distribution, points a and b would compact the intra-instance distribution; a and c would keep inter-instance distribution away; Nevertheless, the intra-semantic compactness between points a and d would disturb the estrangement of inter-instance distribution.

Overall, the occurrence of points which have same semantic label and distinct instance labels is not rare. This condition should be specifically noticed for joint semantic & instance segmentation [11].

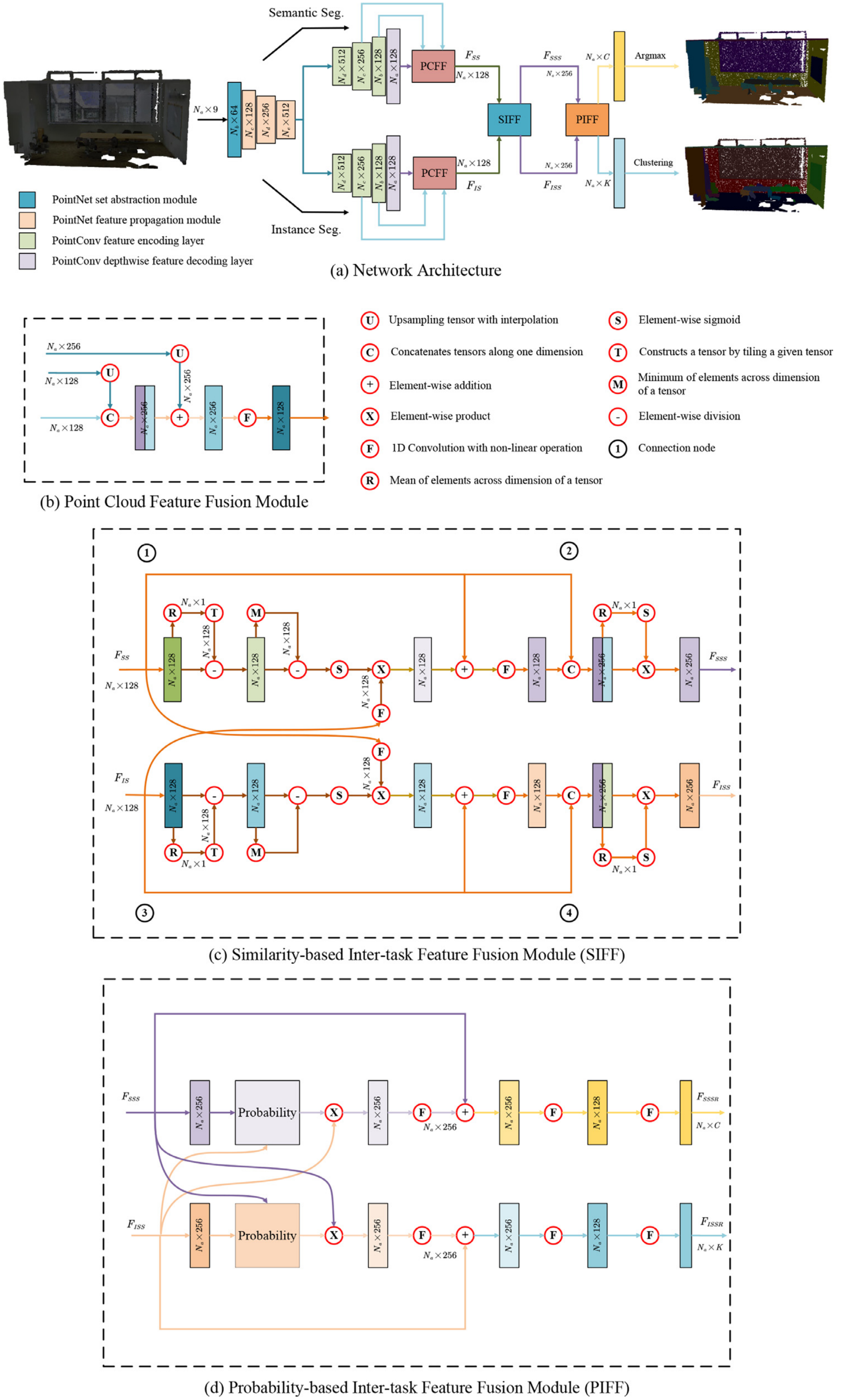
1.1.4. Ignorance of existing methods

Based on the common ground of mutual promotion, ASIS and JSNet introduced two-branch frameworks and joint semantic and instance feature fusion modules to address these two fundamental tasks simultaneously. However, the conflict of joint segmentation is ignored. For detail, they took advantage of feature aggregation (one 1D Convolution) and feature adaption (KNN and global pooling) to transfer features from another branch to current branch. These naive operations could not fully align the instance-wise and semantic-wise features by corresponding semantic-instance relationship. In practical learning, the relationship of mutual promotion is explicit, which is easy to formulate. However, the implicit relationship of conflict is challenging for extracting.

Besides, as a multi-task method, ASIS and JSNet lack powerful task-coherent constraint. Without modeling the quadruplet relation in structure or loss function, the control of consistency of joint training would stay at feature combination and feature flow.

1.2. Contributions of our work

In this paper, we propose a novel method, called JSPNet, for point cloud based joint semantic and instance segmentation. The proposed method contains four parts: a joint semantic and instance two-branch pipeline, a multi-scale feature fusion module for each branch, a similarity-based inter-task feature fusion module (SIFF) and a probability-based inter-task feature fusion module (PIFF). The two-branch pipeline and multi-scale feature fusion module are developed to extract effective high-level features of point cloud. Then SIFF and PIFF modules follow the complete-to-validate scheme to allow two tasks to promote each other. SIFF could figure out the inconspicuous features and alleviate the influence brought by the aforementioned conflict. PIFF would benefit from the discriminative feature to build a semantic-instance relation matrix. In particular, by measuring the similarity between global feature and point-wise feature, SIFF could notice the area of fuzzy content and then compensate it from another branch according to their demand, i.e., similarity matrix. Moreover, given the discriminative per-point information, PIFF is proposed to explore the cross-task probability between semantic and instance segmentation. This probability takes both the common ground of mutual promotion and conflict of two tasks into consideration, which could act as an explicit codebook and constraint for correcting and validating discriminative features. Thus, our method could be used to learn instance-aware semantic feature maps and semantic-aware instance feature embedding which are more discriminative and accurate for point prediction.


Semantic Seg.
Instance Seg.
$N_a \times 9$
$N_b \times 64$
$N_c \times 128$
$N_d \times 256$
$N_e \times 512$
PCFF
SIFF
PIFF
F_{SS}
F_{IS}
F_{SSS}
F_{ISS}
$N_a \times C$
$N_a \times K$
Argmax
Clustering
PointNet set abstraction module
PointNet feature propagation module
PointConv feature encoding layer
PointConv depthwise feature decoding layer
U Upsampling tensor with interpolation
C Concatenates tensors along one dimension
+ Element-wise addition
X Element-wise product
F 1D Convolution with non-linear operation
R Mean of elements across dimension of a tensor
S Element-wise sigmoid
T Constructs a tensor by tiling a given tensor
M Minimum of elements across dimension of a tensor
- Element-wise division
1 Connection node
$N_a \times 256$
$N_a \times 128$
$N_a \times 1$
Probability
F_{SSR}
F_{ISR}

(a) Network Architecture

(b) Point Cloud Feature Fusion Module

(c) Similarity-based Inter-task Feature Fusion Module (SIFF)

(d) Probability-based Inter-task Feature Fusion Module (PIFF)

Fig. 2. Overview of our JSPNet. (a) illustrates the network architecture of our method. (b) (c) and (d) elaborate on the components of PCFF, SIFF and PIFF respectively. Different colored blocks represent different modules in (a), while those colors represent different features in (b), (c) and (d).

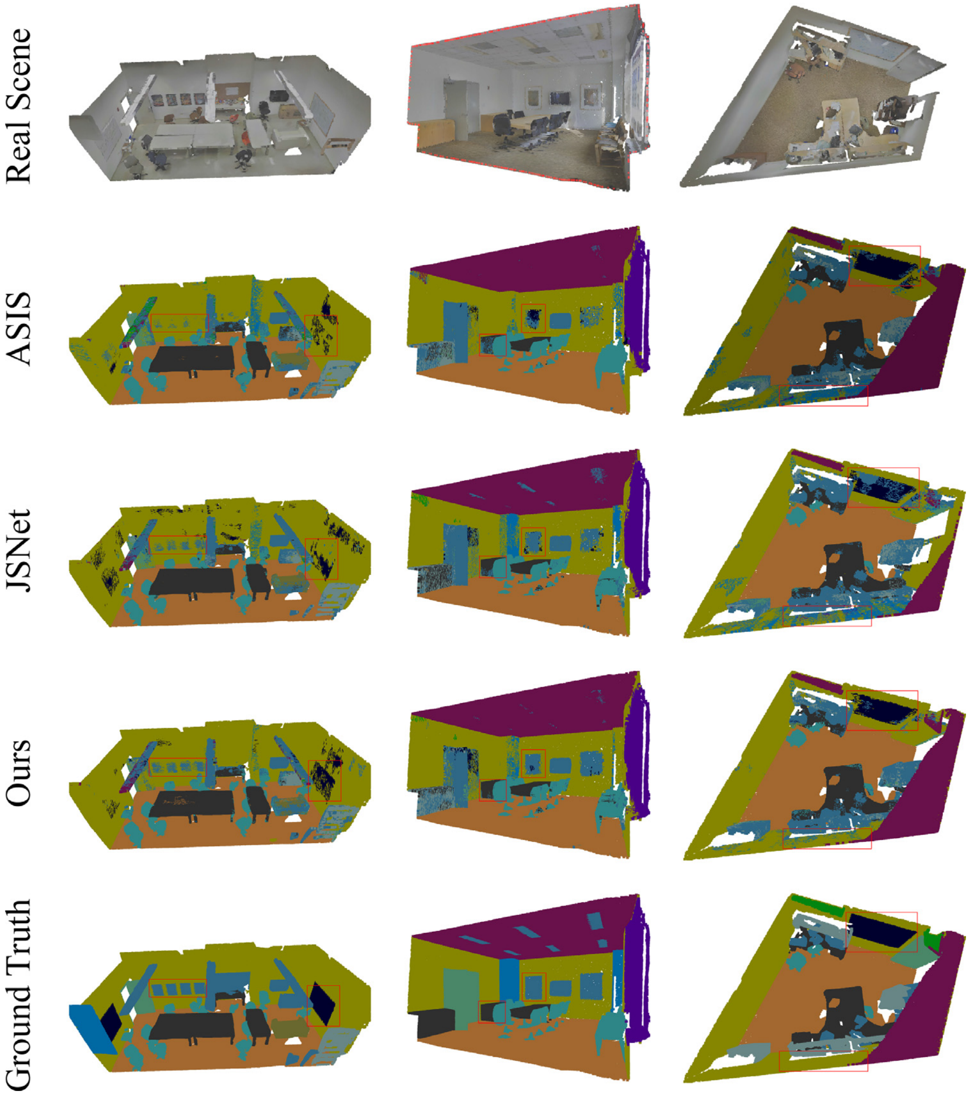


Fig. 3. Qualitative comparison of 3D semantic segmentation between ASIS, JSNet and our method on Area5 of S3DIS.

Overall, our contributions can be summarized as follows:

(1) We analyze the challenges in cooperation and conflict of joint semantic and instance segmentation in detail, which is beneficial to upcoming works to notice.

(2) We propose a novel framework, named JSPNet, to address joint semantic and instance segmentation accurately and feasibly. Specifically, two modules based on feature similarity and cross-task probability could make these two tasks cooperate with each other, as well as solving the challenges proposed in problem statement.

(3) With the end-to-end JSPNet, we achieve state-of-the-art performance on indoor S3DIS dataset [6]. Furthermore, the experimental results on ShapeNet [5] also validate the impressive part-segmentation capability of our method.

2. Related work

Multi-task Learning: Multi-task learning is a learning paradigm that aims to leverage useful information contained in multiple related tasks to improve the generalization performance of all the tasks [12]. Feature selection approach and task relation learning approach are two mainstream approaches of multi-task learning.

The purpose of feature selection approach is to select important information from other tasks to promote current task. Obozinski et al. [13] first proposed to study the multi-task feature selection problem by constrain $l_{2,1}$ norm in classical objective function. Then, the exclusive Lasso model is introduced by Zhou et al. [14] to select useful features in different tasks without overlapping. As to task relations and outlier tasks that can be identified, [15] proposed a probabilistic interpretation and probabilistic framework for multi-task feature selection based on the matrix-variate generalized normal prior.

The task relation learning approach aims to learn quantitative relations between tasks to promote each other. For available task-relation, [16,17] designed regularizers to guide the learning of multiple tasks by adopting task similarities for each pair of tasks. Görnitz et al. [18] built a tree structure to describe task relations. Therefore, parameters of a task corresponding to a node in the tree are enforced to be similar to those of its parent node. For unavailable task-relation, [19] proposed a multi-task Gaussian process (MTGP) by defining a prior vector to learn task-relations from data automatically.

Deep Learning on Point Cloud: Deep learning based approaches have been successfully applied in nature language processing, image processing [20,21], etc [22,23]. However, due to the unorder and spare nature of point cloud, existing deep learning expertise can't be directly applied in point cloud processing [24]. The most intuitive way is to use 3D convolution, however, it con-

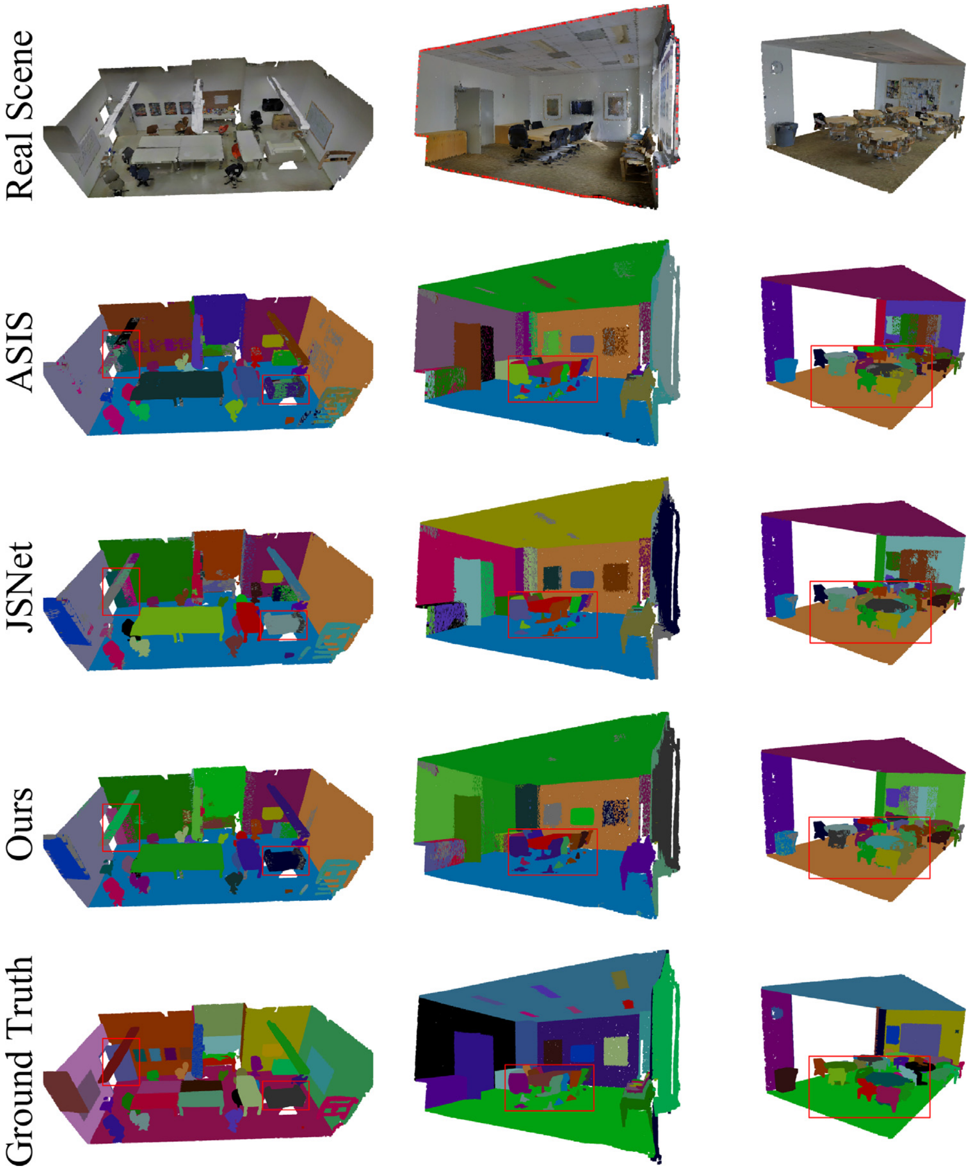


Fig. 4. Qualitative comparison of instance segmentation between ASIS, JSNet and our method on Area5 of S3DIS.

sumes heavy computation cost and is infeasible for large scale scene [25]. Then, to inherit the strong representation ability of 2D CNN, MV3D [26] projected point cloud to bird's eyes view (BEV) plane as pseudo-images and then combined it with RGB images to learn more discriminative representations of objects. However, the details of point cloud are lost during projecting because compressing 3D information to 2D plane would unavoidably leak dimensional information. PointNet [27] proposed Multi-Layer Perception (MLP) to directly encode the unstructured raw point cloud. Later, PointNet++ [28] introduced a hierarchical structure to solve the weakness in capturing local features. Besides, VoxelNet [29] and PointPillar [30] divided the whole point cloud space into evenly spaced grid and further tiled the encoded feature of each voxel to generate pseudo-images. This kind of approaches always transformed the continuous 3D space into a discrete grid representation, then, a shared convolutional operators could handle each volumetric grid, which is similar to operating convolutions in the 2D image domain. However, this volumetric operation is inconsistent with the nature of point cloud, making local details losing in qualification.

Semantic&Instance Segmentation: FCN [1] has achieved tremendous success in image semantic segmentation, as an end-to-end fully connected network. Apart from it, Faster RCNN [31] and YOLO [32] are proposed as anchor-based [33] and anchor-free approaches for object detection and instance segmentation. Then [2] combined the semantic and instance segmentation as panoptic segmentation. In point cloud semantic segmentation, 3D-FCNN [34] used 3D fully connected network to predict voxel-wise semantic labels. PointNet [27] and PointNet++ [28] designed multi-layer perception (MLP) to obtain fine-grained prediction. For point cloud instance segmentation, SGPN [4] modeled the point-wise similarity of point cloud to obtain proposals. Moreover, GSPN [3] proposed an analysis-by-synthesis strategy to tackle 3D object proposal.

For joint semantic and instance segmentation [35,36], JSIS3D [10] used multi-value conditional random field (MV-CRF) as a post processing to refine the semantic and instance prediction. However, separate MV-CRF module following MT-PNet can't be optimized with main framework simultaneously. Then ASIS [8] first proposed a novel end-to-end two-branch model to address these two tasks together. With a shared encoder and two separate de-

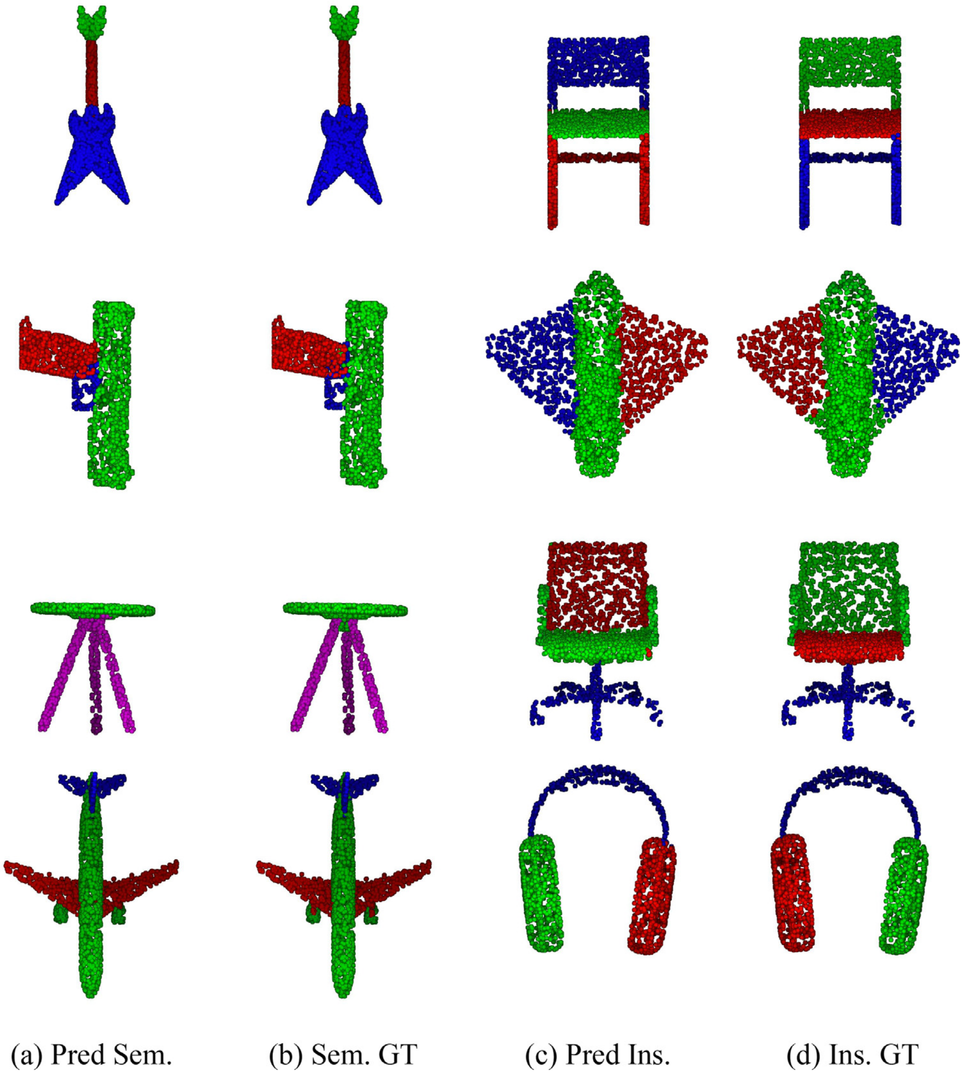


Fig. 5. Qualitative results of our method on ShapeNet. (a), (b), (c) and (b) are semantic prediction of our method, ground truth of semantic segmentation, instance prediction of our method, ground truth of instance segmentation respectively.

coders for two tasks, ASIS [8] associated semantic and instance features by mutual and sequential connection. JSNet revised the encoder and decoder of ASIS, and introduced JISS module, which takes advantage the category-wise information by global average pooling, to improve prediction accuracy. These two works adopted feature aggregation (1D convolution) and feature adaptation (KNN [37] or global average pooling) to generate instance-aware semantic features and semantic-aware instance features. Nevertheless, these naive transformation can't fully model the correlation between instance and semantic features [38,39].

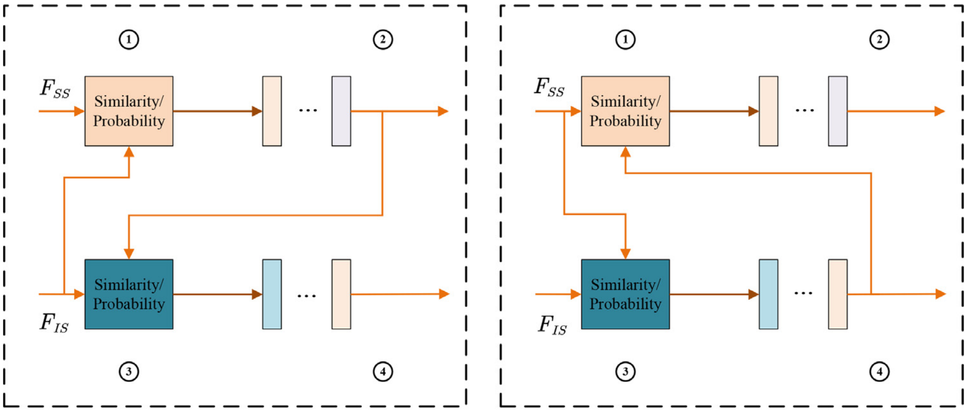
3. Our method

We propose JSPNet to address joint semantic and instance segmentation simultaneously. Section 3.1 elaborates on the network architecture of our method. Then, Sections 3.2 and 3.3 introduce two main components of our method, i.e., similarity-based inter-task feature fusion module (SIFF) and probability-based inter-task feature fusion module (PIFF).

3.1. Network architecture

As illustrated in Fig. 2 (a), our method is composed by four parts: a joint semantic and instance two-branch pipeline, a multi-scale feature fusion module (PCFF), a similarity-based inter-task feature fusion module (SIFF) and a probability-based inter-task feature fusion module (PIFF). The two branch pipeline has a shared encoder and two parallel decoders. These two decoders are developed for two tasks respectively: one aims to capture point-wise semantic feature, the other one is for extracting point-wise instance embedding. For the structure of encoder and decoder, we follow JSNet to combine PointNet++ [27,28] and PointConv to avoid losing details. Specifically, the shared encoder is constructed by a abstraction module of PointNet++ and three encoding layers of PointConv [40] sequentially. Two decoders share the same structure which is composed by three PointConv's depthwise feature decoding layers. Moreover, the following point cloud fusion module PCFF is applied in each branch to capture multi-scale high-level features. Additionally, in order to enhance point-wise saliency, SIFF generates a self-similarity matrix to aware the vague content, and then transforms related features from another task to compensate this blur. Given the salient features generated by SIFF, PIFF could further quantify the task-relatedness via semantic-to-instance/instance-to-semantic probability. Finally, the semantic and instance predictions are generated by argmax and clustering operations respectively.

For the flow of the whole framework, our network takes $N_a \times 9$ raw data which has N_a points, as input. The shared encoder encodes the input to $\frac{1}{128}N_a \times 512$ and two decoders then upsample



(a) Scheme: ins2sem then sem2ins

(b) Scheme: sem2ins then ins2sem

Fig. 6. Illustration of arrangement of connection flow. (a) and (b) depict ins2sem then sem2ins scheme and sem2ins then ins2sem scheme respectively. The connection nodes are same place as in Fig. 2 (c).

the output to $N_a \times 128$. PCFF combines features of $\frac{1}{16}$, $\frac{1}{4}$ and 1 scales from decoder to $N_a \times 128$. Next, SIFF generates a $N_a \times 128$ self-similarity matrix to figure out the vague content and then supplement the inconspicuous content by adding corresponding features from the other branch. Proceeding by few convolutions, the feature with $N_a \times 128$ is fed into PIFF to investigate task-related probability. Finally, the semantic features are shaped to $N_a \times C$ with softmax for semantic prediction, meanwhile, the instance embedding is shaped to $N_a \times K$ for instance prediction, where C and K are the number of semantic categories and the dimension of instance vector respectively.

For training optimization, semantic segmentation loss L_{sem} and instance embedding loss L_{ins} contribute to the total loss L_{total}:

$$L_{total} = L_{sem} + L_{ins}. \tag{1}$$

L_{sem} is the point-wise classical cross entropy loss, as shown in Eq. (2).

$$L_{sem} = -\sum_{i}^{C}\sum_{j}^{N_a} y_i^j \log \hat{y}_i^j, \tag{2}$$

where y and $\hat{y}$ are the semantic ground truth and perdition. Besides, we take advantages of discriminative function [41–43] as instance embedding loss. Specifically, the instance embedding loss L_{ins} could be formulated as:

$$L_{ins} = L_{push} + L_{pull}, \tag{3}$$

where L_{push} forces the mean embedding of different instances to be far away from each other. L_{pull} aims to pull the mean embedding of the same instance to be close with each other. For detail, L_{push} and L_{pull} could be denoted as follows:

$$L_{push} = \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{j=1}^{M}\left[2\delta_d - \left\|\mu_i - \mu_j\right\|_1\right]_+^2, \tag{4}$$

$$L_{pull} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N_m}\sum_{n=1}^{N_m}[\|\mu_m - e_n\|_1 - \delta_v]_+^2, \tag{5}$$

where M is the number of instances, N_m is the number of total elements in mth instance, e_n is the point-wise feature embedding, μ_m is the mean embedding of mth instance. $[x]_+$ is equal to $max(0, x)$, and $\|\cdot\|$ is the same as L_1 distance. Furthermore, δ_d and δ_v are marginal thresholds for L_{push} and L_{pull} respectively.

For test inference, predicted semantic label is generated by argmax. As for instance label, we use mean-shift [44] clustering on embedding to generate final instance prediction.

3.2. Similarity-based inter-task feature fusion module

In fact, it is hard for a single joint semantic and instance module to correct inaccuracy, verify accuracy and clarify vague synchronously. We observe that correcting inaccuracy and verifying accuracy could be included in task-relatedness. Therefore, we first design similarity-based inter-task feature fusion module (SIFF) to compensate the unclear content, which is beneficial to following relation extraction. The main idea of SIFF is similar to that of feature selection approach in multi-task learning. However, duo to discussion analyzed in Section 1.1.3, we make most of the positive benefit and alleviate the drawback brought by negative conditions, i.e., when points having same semantic label belong to different instances.

As shown in Fig. 2 (c), based on the intuition that global pooling extracts global context for categorization, we match the global feature and original point-wise feature, as a feature similarity, to obtain the possibility of inconspicuous point-wise features. Then we transform the corresponding discerning features from the other branch to adaptively fuse it to unclear content. For detail, the demand of requiring information of unclear content is measured by the magnitude of similarity matrix. Therefore, the negative condition would not heavily affect the discriminative feature maps. Besides, it does exist some errors happening in vague area that result from conflict of two tasks, however, it could be further fixed in following PIFF.

Given a main-branch input feature map $A \in \mathbb{R}^{N_a \times C}$ and a sub-branch input feature map $B \in \mathbb{R}^{N_a \times C}$, a global average pooling layer is applied to capture global feature $p \in \mathbb{R}^{1 \times C}$. Then the global feature p is tiled to P which shares the same resolution as A. In order to obtain point-aware demand for saliency, we measure the point-wise similarity by calculating the euclidean distance matric $D \in \mathbb{R}^{N_a \times C}$ between A and P. Then we transform B from the other branch to meet the demand of D. The final output F is generated by the sum of A and $D \times B$, activating by p. The whole procedure could be formulated as:

$$p = GlobalAveragePooling(A), \tag{6}$$

$$P = Tile(p), \tag{7}$$

$$D = Sigmoid(||P - A||_2 - Mean(||P - A||_2)), \tag{8}$$

$$F = p \cdot (D \times Conv1D(B) + A). \tag{9}$$

We apply SIFF module into two branches to mutually aid each other. Therefore, the instance-aware semantic feature F_{SSS} and semantic-fused instance feature F_{ISS} could be represented as:

$$F_{SSS} = SIFF(F_{IS}, F_{SS}), \tag{10}$$

$$F_{ISS} = SIFF(F_{SS}, F_{IS}), \tag{11}$$

3.3. Probability-based inter-task feature fusion module

Understanding and utilizing relation between point-wise semantic and instance features are important for joint training. However, the extracting implicit relation is infeasible for linear transformation. In this work, we combine the dictionary learning and residual connection into our probability-based inter-task feature fusion module (PIFF) to investigate the inherent co-occurrent relation between semantic and instance information. We tackle this co-occurrent relation between semantic and instance feature as a probabilistic problem. As shown in Fig. 2 (d), by measuring the similarity between target feature (main branch) and co-occurrent feature (sub-branch), the possibility distribution of target features conditioned on co-occurrent features could be estimated.

Given target feature $A \in \{x_1, x_2, \dots x_{N_a}\}$ and co-occurrent feature $B \in \{y_1, y_2, \dots y_{N_a}\}$, the measurement of probability in PIFF could be formulated as:

$$F = PIFF(A, B) = \sum_{i=1}^{N_a} \sum_{j=1}^{N_a} p(y_j|x_i) \cdot \phi_j, \tag{12}$$

where ϕ_j is generated by learnable feature transformation from y_j. The probability $p(y_j|x_i)$ of co-occurrent feature y_j based on target feature x_i is:

$$p(y_j|x_i) = \frac{e^{s(y_j,x_i)}}{\sum_{k=1}^{N_a} e^{s(y_k,x_i)}}, \tag{13}$$

where $s(y_k, x_i)$ measures the similarity between y_k and x_i. PIFF could work as modulator without customization in semantic-to-instance and instance-to-semantic feature transformation. Taking semantic-to-instance case as an example, the semantic feature F_{SSS} is first transformed to F_{S2I} by PIFF. Then, with a 1D Convolution, F_{S2I} is concatenated to F_{ISS} as F_{ISSC} to keep spatial correlation. In order to highlight important information, F_{ISSC} is activated by a mean of element across dimension (Mean) and a element-wise sigmoid to generate F_{ISSR}. Finally, with proceeding F_{ISSR}, two 1D convolutions are applied to produce final instance embedding feature $F_{ISSR} \in \mathbb{R}^{N_a \times K}$. The whole procedure can be represented as:

$$F_{S2I} = PIFF(F_{SSS}, F_{ISS}), \tag{14}$$

$$F_{ISSC} = Concat(F_{ISS}, Conv1D(F_{S2I})), \tag{15}$$

$$F_{ISSR} = F_{ISSC} \cdot Sigmoid(Mean(F_{ISSC})), \tag{16}$$

$$F_{ISSR} = Conv1D(Conv1D(F_{ISSR})) \tag{17}$$

4. Experiments

4.1. Dataset and evaluation metric

Two benchmark datasets are applied to evaluate our proposed method: Stanford Large-Scale 3D Indoor Space (S3DIS) [6] and part-segmentation ShapeNet [5]. S3DIS is an indoor 3D instance and semantic benchmark, which contains 6 areas and 272 rooms. Each point in the scene has a semantic and an instance annotation where the semantic annotation is involved in 13 categories. We follow the principle evaluation as ASIS and JSNet to assess our method on 5th fold and 6-fold validation. Apart from S3DIS, we also evaluate the part segmentation ability of our method on ShapeNet. ShapeNet has 16,881 3D CAD from 16 categories. The total dataset is annotated in 50 types of parts. Most categories are labeled in two to five parts. As the official setting, the total dataset is split into 795 scenes for training, 654 for testing. The instance labels are generated by DASCAN as the protocol of SGPN [4], acting as instance ground truth.

For 3D semantic segmentation, three evaluation metrics, i.e., overall accuracy (oAcc), mean accuracy (mAcc) and mean IoU (mIoU), are utilized to validate the performance of our method. Each metric is applied by calculating across over all categories. For 3D instance segmentation, mean precision (mPre), mean recall (mRec) with 0.5 IoU threshold, and (weight) converge (Conv, WConv) are adopted to evaluate our method. Conv denotes the average instance-wise IoU of prediction according to ground truth. Based on Conv, WConv takes advantage of the weight calculated by the size of ground truth instance for a balanced assessment. Let $\mathcal{G}$ denote ground truth regions and $\mathcal{P}$ is prediction regions, Conv and WConv can be formulated as:

$$Cov(\mathcal{G}, \mathcal{P}) = \sum_{m=1}^{|\mathcal{G}|} \frac{1}{|\mathcal{G}|} \max_n \text{IoU}\left(r_m^{\mathcal{G}}, r_n^{\mathcal{P}}\right), \tag{18}$$

$$WCov(\mathcal{G}, \mathcal{P}) = \sum_{m=1}^{|\mathcal{G}|} w_m \max_n \text{IoU}\left(r_m^{\mathcal{G}}, r_n^{\mathcal{P}}\right), \tag{19}$$

$$w_m = \frac{\left|r_m^{\mathcal{G}}\right|}{\sum_k \left|r_k^{\mathcal{G}}\right|}, \tag{20}$$

where $\left|r_m^{\mathcal{G}}\right|$ is the total number of points in ground truth region i.

4.2. Training and inference details

For the indoor S3DIS dataset, each point has 9-dim feature vector (XYZ, RGB and normalized coordinate as to the room). In training, the rooms are split to $1m \times 1m$ overlapped blocks along the ground plane, as the experimental setting of PintNet. Each overlapped block contains 4096 points in total. The inter-distribution and intra-distribution marginal thresholds δ_d and δ_v are set to $\delta_d = 1.5$, $\delta_v = 0.5$, and the dimension K of instance feature embedding is 5. Our network is trained with 100 epochs, batch size 12, learning rate 0.001 on a single NVIDIA GTX1080Ti. Adam optimizer is adopted for network optimization. Momentum is set to 0.9 and the decay step is set to 12,500 iterations for cutting the initial learning rate half. As for testing, instance objects are generated by mean-shift clustering with bandwidth 0.6. Then BlockMerging algorithm is utilized for merging blocks of instance. During training ShapeNet dataset, each shape is sampled to 2048 points which contain 6 dimensions (XYZ, normalized coordinate as to the shape).

4.3. Semantic segmentation results on S3DIS

As reported in Table 1, we provide the performance of semantic segmentation of our method on S3DIS. Compared with existing state-of-the-arts including PointNet [27], SEGCloud [45], RSNet [46], 3P-RNN [47], MT-PNet [10], MV-CRF [10], ASIS [8] and JSNet [9], our method outperforms most of them with significant margin in both 5th fold and 6 fold validations. Especially for ASIS and JSNet which share a similar framework as our method, our method achieves more than 1.5% improvement for mAcc and mIoU matrixes in 5th area. In 6 fold validation, our method almost obtains

Table 1
Comparison of semantic segmentation results on S3DIS.

	Method	mAcc	oAcc	mIoU
5th fold	PointNet	52.1	83.5	43.4
	SEGCloud	54.7	-	48.9
	RSNet	59.4	-	51.9
	3P-RNN	**71.3**	85.7	53.4
	ASIS	60.9	86.9	53.4
	JSNet	61.4	87.7	54.5
	Ours	63.8	**88.2**	**56.1**
6 fold	PointNet	60.3	80.3	48.9
	3P-RNN	**73.6**	86.9	56.3
	MT-PNet	-	86.7	-
	MV-CRF	-	87.4	-
	ASIS	70.1	86.2	59.3
	JSNet	71.7	88.7	61.7
	Ours	72.6	**89.7**	**62.5**

Table 2
Comparison of Instance segmentation results on S3DIS.

	Method	mConv	mWconv	mRec	mPre
5th fold	SGPN	32.7	35.5	28.7	36.0
	ASIS	44.6	47.8	42.4	55.3
	JSNet	48.7	51.5	46.9	**62.1**
	Ours	**50.7**	**53.5**	**48.0**	59.6
6 fold	SGPN	37.9	40.8	31.2	38.2
	MT-PNet	-	-	-	24.9
	MV-CRF	-	-	-	36.3
	ASIS	51.2	55.1	47.5	63.6
	2D-BoNet	-	-	47.6	65.6
	JSNet	54.1	58.0	53.9	**66.9**
	Ours	**54.9**	**58.8**	**55.0**	66.5

comparable, even better performance over latest approaches. For qualitative evaluation, Fig. 3 illustrates the well-segmented scenes predicted by our method. Our method could carefully segment objects with different scales in complex environments.

4.4. Instance segmentation results on S3DIS

To validate the performance of our method in instance segmentation, Table 2 depicts the experimental comparison with state-of-the-arts on S3DIS. In 5th fold validation, our method achieves significant improvement: 2.0% mConv, 2.0% mWconv, 1.1% mRecall over those of JSNet. For mean prediction, let alone other approaches, JSNet is slightly better than our method. For generalization, we also compare our method with state-of-arts in 6 fold validation. Compared with SGPN, the performance is improved with 17.0%, 18.0%, 23.8%, 28.2% for four matrixes respectively. As to the latest method 3D-BoNet [7], our method could also gain better results than it. Moreover, qualitative results are illustrated in Fig. 4, which indicates the well-segmented instance capability of our method.

4.5. Results on ShapeNet

Apart from evaluating our method on S3DIS, we also further make a thorough experiment on ShapeNet. Since the instance ground truth is generated by algorithm, instead of hand-labeling, we follow the setting as JSNet and ASIS that only provide qualitative results of instance segmentation in Fig. 5. We can see our method could favor the part instance segmentation. For semantic segmentation, Table 3 reports the performance comparison between PointNet, PointNet++, ASIS, JSNet with our method. Compared with PointNet++, our method makes a 1.3% mIoU improvement over it. As for ASIS and JSNet, our method outperforms ASIS with 1.2% improvement and achieves slightly better results than JSNet, indicating our SIFF and PIFF could further advance the performance of joint semantic & instance segmentation methods. Furthermore, Fig. 5 also illustrates the qualitative semantic visualization of our method on ShapeNet. The results represent that the proposed method could make a good application on part semantic segmentation.

Table 3
Semantic segmentation results on ShapeNet.

Method	mIoU
PointNet	83.7
PointNet+	84.9
ASIS	85.0
JSNet	85.8
Ours	**86.2**

Table 4
Ablation experiment of semantic segmentation for pipeline and backbone on Area 5 of S3DIS. 'S&I module' denotes the joint semantic & instance segmentation module.

Method	Pipeline	Backbone	S&I module	mAcc	oAcc	mIoU
-	ASIS	PointNet	w/o	53.8	82.6	44.3
-	ASIS	PointNet+	w/o	57.9	85.0	50.0
Ours	ASIS	PointNet	SIFF&PIFF	57.2	84.6	48.2
Ours	ASIS	PointNet+	SIFF&PIFF	58.2	86.5	52.0
-	JSNet	PointNet+	w/o	59.0	86.3	53.7
Ours	JSNet	PointNet+	SIFF&PIFF	60.5	87.9	55.0

5. Ablation study

In this section, we extensively analyze the effectiveness of each part of our method, including SIFF, PIFF, pipeline, backbone, training strategy, etc.

To validate the sensitiveness of SIFF and PIFF to specific pipeline, we insert them into ASIS's and JSNet's to replace corresponding joint semantic and instance segmentation module (S&I module). As shown in Tables 4 and 5, SIFF and PIFF could revise the fundamental results on semantic segmentation and on instance segmentation. The results report two proposed modules are robust to various pipelines and backbones.

To validate the effectiveness of SIFF, PIFF and training strategy, we make ablation experiments of them in the full framework, as shown in Table 6. Compared with groups (2), (3) and (5), we can notice that single SIFF only provide limited improvement and single PIFF also can't do its best. When combining these two modules together, additional improvement could be further obtained. We believe that the conspicuous feature capability brought by SIFF could pave the way for PIFF. Furthermore, in groups (4) and (5), the enriched multi-scale information is beneficial for joint semantic and instance segmentation training. We also analyze the training strategy of early stopping and random sample in groups (6), (7) and (8). The random sample could enhance the robustness and generalization of our model. The early stopping strategy may provide better results than model fully trained.

Different from ASIS and JSNet, our SIFF and PIFF are parallel transforation. Exactly, our framework has four interaction between semantic and instance branches: The instance-to-semantic (ins2sem) and semantic-to-instance (sem2ins) transformation in two modules are processed simultaneously, instead of specific arrangement. In Table 7, we also elaborate on necessity of depending on scheming: sem2ins then ins2sem or ins2sem then sem2ins. We make ablation experiments on each module. As depicted in Fig. 6, the sem2ins then ins2sem scheme is ①→③→④→①, and ins2sem then sem2ins scheme is ③→①→②→③, for both SIFF and PIFF, where connection nodes share the place as in

Table 5
Ablation experiment of instance segmentation for pipeline and backbone on Area 5 of S3DIS. 'S&I module' denotes the joint semantic & instance segmentation module.

Method	Pipeline	Backbone	S&I module	mConv	mWconv	mRec	mPre
-	ASIS	PointNet	w/o	38.1	40.6	35.1	42.5
-	ASIS	PointNet+	w/o	42.9	45.6	40.9	53.5
Ours	ASIS	PointNet	SIFF&PIFF	42.0	43.9	36.8	46.5
Ours	ASIS	PointNet+	SIFF&PIFF	45.1	48.8	43.6	56.2
-	JSNet	PointNet	w/o	44.0	48.2	43.0	53.0
Ours	JSNet	PointNet+	SIFF&PIFF	48.2	53.0	47.2	63.2

Table 6
Ablation experimental results of different components and training strategies on Area5 of S3DIS. 'ES' and 'RS' represent early stopping and random sample respectively.

Group	Component			Strategy		Metric	
	PCFF	SIFF	PIFF	ES	RS	mPrec	mIoU
(1)						53.2	53.5
(2)	√	√				53.6	54.0
(3)	√		√			54.9	54.2
(4)		√	√			54.3	53.6
(5)	√	√	√			56.8	54.3
(6)	√	√	√	√		59.0	55.3
(7)	√	√	√		√	60.0	56.1
(8)	√	√	√	√	√	60.5	56.5

Table 7
Results of schemed circuit of SIFF and PIFF on Area5 of S3DIS.

Module	Scheme	mPre	mIoU
SIFF	ins2sem→sem2ins	59.65	52.66
	sem2ins→ins2sem	56.32	50.93
	w/o	56.64	53.27
PIFF	ins2sem→sem2ins	58.33	53.96
	sem2ins→ins2sem	57.96	54.06
	w/o	60.23	55.24

Fig. 2 (c). Compared with the setting without scheme, the results in Table 7 indicates our modules don't rely on arrangement like ASIS and JSNet.

6. Conclusion

In this paper, we propose the end-to-end JSPNet for joint point cloud semantic and instance segmentation. We exhaustively analyze the common ground of cooperation and conflict between two tasks. To model the correlation between semantic and instance segmentation, we propose SIFF and PIFF to reinforce mutual cooperation and alleviate the essential conflict. These two fusion modules follow the complete-to-validate scheme. By measuring self-similarity with the global feature, SIFF could locate the inconspicuous feature content and revise it according to per-point demand. Based on discriminative features, PIFF could investigate the task-relatedness between semantic and instance segmentation. Finally, our method achieves state-of-the-art performance on S3DIS and ShapeNet, outperforming similar previous works, i.e., ASIS and JSNet.

Essentially, the joint semantic and instance segmentation is a multi-task problem with relatively explicit cooperation manner which could be modeled or learned through reasoning the relations between two sub-tasks. Therefore, for the general multi-task learning whose tasks may share same attributes as our task, developing simultaneous cross-task alignment or fusion by reasoning the relations of sub-tasks would be more beneficial, since usually task-wise proceeding has few specifically consequent arrangements. In our future work, we would focus on applying our method on autonomous driving scene, serving for accurate scene understanding in complicated environment.

Declaration of Competing Interest

The authors have no declaration of interest to report.

Acknowledgements

The work in this paper was supported by the National Natural Science Foundation of China (Nos. 62076139, 61972212, 61702280), Natural Science Foundation of Jiangsu Province (No. BK20190089), National Postdoctoral Program for Innovative Talents (No. BX20180146), National Key Re&D Program of China (Nos. 2017YFB1401302, 2017YFB0202200), NSFC-Key Project of General Technology Fundamental Research United Fund (No. U1736211), China Postdoctoral Science Foundation (No. 2019M661901), Jiangsu Planned Projects for Postdoctoral Research Funds (No. 2019K024), the Six Talent Peaks Project in Jiangsu Province (No. RJFW-011), Key R&D Program of Jiangsu (No. BE2017166), Zhejiang Lab (No. 2021KF0AB05), and 1311 Talent Program of Nanjing University of Posts and Telecommunications.

References

[1] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[2] A. Kirillov, K. He, R. Girshick, C. Rother, P. Dollár, Panoptic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9404–9413.
[3] L. Yi, W. Zhao, H. Wang, M. Sung, L.J. Guibas, GSPN: generative shape proposal network for 3D instance segmentation in point cloud, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3947–3956.
[4] W. Wang, R. Yu, Q. Huang, U. Neumann, SGPN: similarity group proposal network for 3D point cloud instance segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2569–2578.
[5] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, J. Xiao, 3D ShapeNets: a deep representation for volumetric shapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1912–1920.
[6] I. Armeni, O. Sener, A.R. Zamir, H. Jiang, I. Brilakis, M. Fischer, S. Savarese, 3D semantic parsing of large-scale indoor spaces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1534–1543.
[7] B. Yang, J. Wang, R. Clark, Q. Hu, S. Wang, A. Markham, N. Trigoni, Learning object bounding boxes for 3D instance segmentation on point clouds, in: Advances in Neural Information Processing Systems, 2019, pp. 6740–6749.
[8] X. Wang, S. Liu, X. Shen, C. Shen, J. Jia, Associatively segmenting instances and semantics in point clouds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4096–4105.
[9] L. Zhao, W. Tao, JSNet: joint instance and semantic segmentation of 3D point clouds, in: Advances in Neural Information Processing Systems, 2020, pp. 12951–12958.
[10] Q.-H. Pham, T. Nguyen, B.-S. Hua, G. Roig, S.-K. Yeung, 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, 2019, pp. 8827–8836.
[11] S. Serikawa, H. Lu, Underwater image dehazing using joint trilateral filter, Computers & Electrical Engineering 40 (1) (2014) 41–50.
[12] Y. Zhang, Q. Yang, A survey on multi-task learning, arXiv preprint arXiv:1707.08114(2017).
[13] G. Obozinski, B. Taskar, M. Jordan, Multi-task feature selection, Statistics Department, UC Berkeley, Tech. Rep 2 (2.2) (2006) 2.

[14] Y. Zhou, R. Jin, S.C.-H. Hoi, Exclusive lasso for multi-task feature selection, in: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2010, pp. 988–995.

[15] Y. Zhang, D.-Y. Yeung, Q. Xu, Probabilistic multi-task feature selection, in: Advances in Neural Information Processing Systems, 2010, pp. 2559–2567.

[16] T. Evgeniou, C.A. Micchelli, M. Pontil, Learning multiple tasks with kernel methods, J. Mach. Learn. Res. 6 (Apr) (2005) 615–637.

[17] T. Kato, H. Kashima, M. Sugiyama, K. Asai, Multi-task learning via conic programming, in: Advances in Neural Information Processing Systems, 2008, pp. 737–744.

[18] N. Görnitz, C. Widmer, G. Zeller, A. Kahles, G. Rätsch, S. Sonnenburg, Hierarchical multitask structured output learning for large-scale sequence segmentation, in: Advances in Neural Information Processing Systems, 2011, pp. 2690–2698.

[19] E.V. Bonilla, K.M. Chai, C. Williams, Multi-task gaussian process prediction, in: Advances in Neural Information Processing Systems, 2008, pp. 153–160.

[20] Z. Chen, H. Lu, S. Tian, J. Qiu, T. Kamiya, S. Serikawa, L. Xu, Construction of a hierarchical feature enhancement network and its application in fault recognition, IEEE Trans. Ind. Inf. 17 (7) (2020) 4827–4836.

[21] H. Lu, R. Yang, Z. Deng, Y. Zhang, G. Gao, R. Lan, Chinese image captioning via fuzzy attention-based DenseNet-BiLSTM, ACM Trans. Multimedia Comput.Commun. Appl. 17 (1s) (2021) 1–18.

[22] H. Lu, Y. Li, M. Chen, H. Kim, S. Serikawa, Brain intelligence: go beyond artificial intelligence, Mob. Netw. Appl. 23 (2) (2018) 368–375.

[23] P. Wang, D. Wang, X. Zhang, X. Li, T. Peng, H. Lu, X. Tian, Numerical and experimental study on the maneuverability of an active propeller control based wave glider, Appl. Ocean Res. 104 (2020) 102369.

[24] T. He, Y. liu, C. Shen, X. Wang, C. Sun, Instance-aware embedding for point cloud instance segmentation, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 567–575.

[25] M. Engelcke, D. Rao, D.Z. Wang, C.H. Tong, I. Posner, Vote3Deep: fast object detection in 3D point clouds using efficient convolutional neural networks, in: Proceedings of IEEE International Conference on Robotics and Automation, 2017, pp. 1355–1361.

[26] X. Chen, H. Ma, J. Wan, B. Li, T. Xia, Multi-view 3D object detection network for autonomous driving, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1907–1915.

[27] C.R. Qi, H. Su, K. Mo, L.J. Guibas, PointNet: deep learning on point sets for 3D classification and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 652–660.

[28] C.R. Qi, L. Yi, H. Su, L.J. Guibas, PointNet++: deep hierarchical feature learning on point sets in a metric space, in: Advances in Neural Information Processing Systems, 2017, pp. 5099–5108.

[29] Y. Zhou, O. Tuzel, VoxelNet: end-to-end learning for point cloud based 3D object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4490–4499.

[30] A.H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, O. Beijbom, PointPillars: fast encoders for object detection from point clouds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12697–12705.

[31] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[32] J. Redmon, A. Farhadi, YOLOv3: an incremental improvement, arXiv preprint arXiv:1804.02767(2018).

[33] Z. Huang, L. Huang, Y. Gong, C. Huang, X. Wang, Mask scoring R-CNN, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6409–6418.

[34] J. Huang, S. You, Point cloud labeling using 3D convolutional neural network, in: International Conference on Pattern Recognition, 2016, pp. 2670–2675.

[35] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.

[36] S. Minaee, Y.Y. Boykov, F. Porikli, A.J. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: a survey, IEEE Trans. Pattern Anal. Mach. Intell. (2021) inpress, doi:10.1109/TPAMI.2021.3059968.

[37] P. Fränti, S. Sieranoja, How much can k-means be improved by using better initialization and repeats? Pattern Recognit. 93 (2019) 95–112.

[38] Q. Wang, J. Xie, W. Zuo, L. Zhang, P. Li, Deep CNNs meet global covariance pooling: better representation and generalization, IEEE Trans. Pattern Anal. Mach. Intell. 43 (8) (2021) 2582–2597.

[39] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.

[40] W. Wu, Z. Qi, L. Fuxin, PointConv: deep convolutional networks on 3D point clouds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9621–9630.

[41] B. De Brabandere, D. Neven, L. Van Gool, Semantic instance segmentation with a discriminative loss function, arXiv preprint arXiv:1708.02551(2017).

[42] G. Gao, J. Yang, X.-Y. Jing, F. Shen, W. Yang, D. Yue, Learning robust and discriminative low-rank representations for face recognition with occlusion, Pattern Recognit. 66 (2017) 129–143.

[43] G. Gao, Y. Yu, J. Xie, J. Yang, M. Yang, J. Zhang, Constructing multilayer locality-constrained matrix regression framework for noise robust face super-resolution, Pattern Recognit. 110 (2021) 107539.

[44] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Trans. Pattern Anal. Mach. Intell. 24 (5) (2002) 603–619.

[45] L. Tchapmi, C. Choy, I. Armeni, J. Gwak, S. Savarese, SEGCloud: semantic segmentation of 3D point clouds, in: Proceedings of International Conference on 3D Vision, 2017, pp. 537–547.

[46] Q. Huang, W. Wang, U. Neumann, Recurrent slice networks for 3D segmentation of point clouds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2626–2635.

[47] X. Ye, J. Li, H. Huang, L. Du, X. Zhang, 3D recurrent neural networks with context fusion for point cloud semantic segmentation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 403–417.

Feng Chen is pursuing the Master degree in computer technology from Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include computer vision and image processing.

Fei Wu received the Ph.D. degree in computer science from Nanjing University of Posts and Telecommunications (NJUPT), China, in 2016. He is currently an associate professor with the College of Automation in NJUPT. He has authored over forty scientific papers, such as TPAMI, TIP, TCYB, PR, TSE, TR, CVPR, AAAI, IJCAI and WWW. His research interests include pattern recognition, artificial intelligence, and computer vision.

Guangwei Gao received the Ph.D. degree in pattern recognition and intelligence systems from Nanjing University of Science and Technology, Nanjing, China, in 2014. Now, he is an associate professor with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China. His research mainly focuses on pattern recognition and computer vision.

Yimu Ji is a professor in Nanjing University of Posts and Telecommunications, Nanjing, China. His research mainly focuses on intelligent driving and data processing.

Jing Xu is an undergraduate of Hohai University, Nanjing, China. Her major is criminal law and her re search interests include fairness of artificial intelligence and computer vision.

Guo-Ping Jiang received the Ph.D. degree in control theory and engineering from Southeast University, Nanjing, in 1997. Now he is a professor with the College of Automation, Nanjing University of Posts and Telecommunications, China. His research interests include complex dynamical networks and artificial intelligence.

Xiao-Yuan Jing received the doctoral degree of pattern recognition and intelligent system in the Nanjing University of Science and Technology, 1998. Now he is a professor with the School of Computer, Wuhan University, and with the College of Automation, Nanjing University of Posts and Telecommunications, China. His research interests include artificial intelligence and pattern recognition.