

Towards Semi-supervised Dual-modal Semantic Segmentation

Qiulei Dong, *Member, IEEE*, Jianan Li, Shuang Deng

Abstract—With the development of 3D and 2D data acquisition techniques, it has become easy to obtain point clouds and images of scenes simultaneously, which further facilitates dual-modal semantic segmentation. Most existing methods for simultaneously segmenting point clouds and images rely heavily on the quantity and quality of the labeled training data. However, massive point-wise and pixel-wise labeling procedures are time-consuming and labor-intensive. To address this issue, we propose a parallel dual-stream network to handle the semi-supervised dual-modal semantic segmentation task, called PD-Net, by jointly utilizing a small number of labeled point clouds, a large number of unlabeled point clouds, and unlabeled images. The proposed PD-Net consists of two parallel streams (called original stream and pseudo-label prediction stream). The pseudo-label prediction stream predicts the pseudo labels of unlabeled point clouds and their corresponding images. Then, the unlabeled data is sent to the original stream for self-training. Each stream contains two encoder-decoder branches for 3D and 2D data respectively. In each stream, multiple dual-modal fusion modules are explored for fusing the dual-modal features. In addition, a pseudo-label optimization module is explored to optimize the pseudo labels output by the pseudo-label prediction stream. Experimental results on two public datasets demonstrate that the proposed PD-Net not only outperforms the comparative semi-supervised methods but also achieves competitive performances with some fully-supervised methods in most cases.

Index Terms—Point Clouds, Dual Modality, Semi-supervised Semantic Segmentation.

I. INTRODUCTION

WITH the rapid development of both 3D and 2D data acquisition techniques, the 3D point clouds and images of scenes could be easily acquired together by jointly utilizing 3D and 2D sensors. And the correspondences between 3D points and image pixels could be easily calculated with the intrinsic and extrinsic parameters of the sensors. Accordingly, unlike the existing works [1]–[8] that only segment uni-modal data, many segmentation methods [9]–[15] are proposed to combine the complementary information of point clouds and images to boost performances, and they are trained in a fully-supervised manner under a general pipeline shown in Figure 1(a). This category of methods generally trains a 3D-DNN-based extractor to learn features for segmenting point

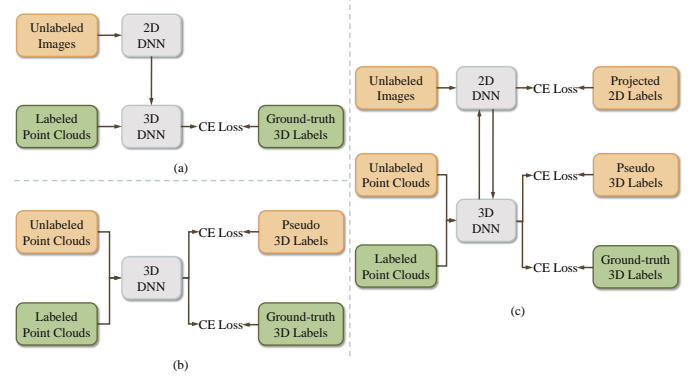


Fig. 1. Pipeline comparison of three segmentation categories: (a) fully-supervised dual-modal 3D segmentation, (b) semi-supervised uni-modal 3D segmentation, and (c) semi-supervised dual-modal segmentation. The orange boxes represent either the unlabeled data or its corresponding labels, while the green boxes represent either the labeled data or its corresponding labels. ‘CE Loss’ represents the cross-entropy loss.

clouds by minimizing the cross-entropy loss with regard to ground-truth 3D labels, with the assistance of the corresponding features extracted by a 2D-DNN-based extractor. However, it has to spend an expensive cost on data annotation.

In order to alleviate the above data annotation problem, some works [16]–[25] focus on semi-supervised semantic segmentation. Figure 1(b) shows the pipeline of semi-supervised uni-modal 3D segmentation, whose training data contains a small percentage of labeled point clouds and a large percentage of unlabeled point clouds. This category of methods generally trains a 3D-DNN-based extractor iteratively to learn features for segmenting point clouds, by minimizing the cross-entropy loss with regard to both the ground-truth 3D labels of the labeled point clouds and the predicted pseudo labels of the originally unlabeled point clouds. However, these semi-supervised methods only use uni-modal data, which could not make the most of the collected dual-modal data. Thus, how to utilize the complementary information in point clouds and images to solve the semi-supervised dual-modal segmentation problem remains to be investigated.

Figure 1(c) shows the general pipeline of semi-supervised dual-modal segmentation, whose training data contains a small percentage of labeled point clouds, a large percentage of unlabeled point clouds, and a set of unlabeled images. Since the camera parameters are either given by some datasets or easily calculated by some commonly-used calibration techniques, the labels of the projected 2D pixels by the 3D points could be identified according to the camera parameters. This category

Corresponding Author: Qiulei Dong.

Qiulei Dong and Jianan Li are with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, and also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: qldong@nlpr.ia.ac.cn, lijianan211@mails.ucas.ac.cn).

Shuang Deng is with the Autonomous Driving Division of X Research Department, JD logistics, Beijing 102600, China (e-mail: deng-shuang10@jd.com).

of methods employs an iterative optimization strategy: At each iteration, it trains and updates a 2D-DNN-based extractor to learn 2D features for segmenting images by minimizing a cross-entropy loss with regard to the labels of the projected 2D pixels, with the assistance of the corresponding features extracted from point clouds by a 3D-DNN-based extractor. Then with the assistance of the learned 2D features, a 3D-DNN-based extractor is trained and updated to learn features for segmenting point clouds by minimizing a cross-entropy loss with regard to both the ground-truth 3D labels of the labeled point clouds and the predicted pseudo labels of the unlabeled point clouds.

To handle the semi-supervised dual-modal semantic segmentation task, we propose a parallel dual-stream network, called PD-Net. It contains two parallel streams with the same architecture: an original stream, and a pseudo-label prediction stream whose parameters are updated by the Exponential Moving Average (EMA) strategy [24]. Each stream in PD-Net contains a 3D encoder-decoder branch, a 2D encoder-decoder branch, and multiple dual-modal fusion modules. The 3D and 2D branches are utilized to extract 3D and 2D features respectively. Intuitively, jointly leveraging dual-modal features could improve the segmentation performance, considering the complementarity of 3D features and 2D features (i.e., the 3D features contain rich geometric information but lack textural information, while the 2D features are enriched with color and textural information but are short of depth information). However, direct fusion may dilute the inter-modal attentive weights, which could undermine the performance instead. To fully exploit the complementary information in dual-modal data, we propose the dual-modal fusion module, which fuses the 3D and 2D latent features via a multi-head attention-based mechanism. Besides, a consistency loss term is designed to constrain the semantic consistency between the 3D and 2D features. The labeled point clouds and their corresponding images are only trained in the original stream, and the labels of the images are projected from the point clouds according to the sensor parameters. The unlabeled point clouds and their corresponding images are trained in both two streams. Specifically, the output of the original stream is supervised by the pseudo labels output by the pseudo-label prediction stream. To improve the quality of the pseudo labels generated by the pseudo-label prediction stream so that the effectiveness of the self-training strategy for the unlabeled point clouds and their corresponding images is guaranteed, we propose the pseudo-label optimization module to leverages pseudo labels of one modality to improve the quality of pseudo labels of another modality based on a voting mechanism. The pseudo-label optimization module is non-parametric, thus it is free from inductive bias and performance degeneration due to the domain gap between different modalities.

In sum, the main contributions of this paper include:

- We propose the dual-modal fusion module and the consistency loss term, which could effectively fuse the features of point clouds and images.
- We propose the pseudo-label optimization module, which is helpful for improving the quality of the predicted pseudo labels.

- We propose the PD-Net, which consists of the aforementioned dual-modal fusion module, consistency loss term, and pseudo-label optimization module. To our best knowledge, this work is the first attempt to investigate how to utilize dual-modal data to handle the semi-supervised segmentation task for both point clouds and images.

The remainder of this paper is organized as follows. Some existing methods on 3D semi-supervised semantic segmentation, 2D semi-supervised semantic segmentation, and fully-supervised dual-modal semantic segmentation are reviewed in Section II. The proposed method is introduced in detail in Section III. The experimental results are reported in Section IV. Finally, we conclude this paper in Section V.

II. RELATED WORKS

In this section, we first introduce the related semi-supervised segmentation methods of point clouds and images respectively. Then, we introduce the related fully-supervised segmentation methods that combine the point clouds and images.

A. 3D Semi-supervised Semantic Segmentation

To alleviate the annotation burden, label-efficient 3D semantic segmentation (including weakly-supervised [26], [27], semi-supervised [16]–[22], and unsupervised 3D semantic segmentation) has drawn growing attention among researchers. In this work, we delve into semi-supervised 3D semantic segmentation, which aims at utilizing a small number of densely labeled data and a large number of unlabeled data for model training.

Some early semi-supervised 3D segmentation works [16], [17] rely on additional information (i.e., expert annotation) to constrain the features of unlabeled point clouds. However, the application is limited because the introduced expert knowledge is not applicable to all circumstances. To overcome this defect, Li *et al.* [18] proposed to design an adversarial architecture to calculate the confidence discrimination of pseudo labels for the unlabeled point clouds, and select the pseudo labels with high reliability. Jiang *et al.* [19] proposed to utilize the contrastive loss based on the pseudo-label guidance to enhance the feature representation and model generalization ability in the semi-supervised setting. Deng *et al.* [21] proposed to combine the geometry and color-based super-points to optimize the pseudo labels to guarantee the reliability of the self-training of the unlabeled points. Taking the prior knowledge of LiDAR point clouds into consideration, Kong *et al.* [20] proposed to mix laser beams from different LiDAR scans and then encourage the model to make consistent and confident predictions before and after mixing. Li *et al.* [22] designed a soft pseudo-label method informed by LiDAR reflectivity to make full use of the limited labeled points and abundant unlabeled points.

B. 2D Semi-supervised Semantic Segmentation

The great advances of semi-supervised learning in image classification [23], [24] inspire the investigation of semi-supervised semantic segmentation [28]–[39] for images. Early

works on 2D semi-supervised semantic segmentation leveraged the Generative Adversarial Networks (GAN) to synthesize high-quality pseudo labels. Hung *et al.* [40] designed a fully-convolutional discriminator which enables semi-supervised learning by searching the reliable regions in predicted results of unlabeled images, thereby providing additional supervisory signals for training. Mittal *et al.* [29] proposed a GAN-based branch to improve the low-level details in segmentation predictions, which is helpful for alleviating low-level artifacts in the low-data regime.

Recently, researchers have paid more and more attention to consistency regularization and contrastive learning. Chen *et al.* [31] imposed the consistency between networks with different initialization and encouraged the high similarity between the predictions of the two networks, which expands the training data by regarding the pseudo labels as the supervision for unlabeled images. Mai *et al.* [39] proposed a coherent RankMatch network that leverages the inter-pixel correlations to construct more safe and effective supervision signals. Inspired by the great success of contrastive learning in unsupervised learning, Zhao *et al.* [32] designed a simple yet effective contrastive learning-based training strategy for semi-supervised segmentation, which increases intra-class compactness and inter-class separability. Liu *et al.* [41] proposed a contrastive learning framework designed at a regional level that performs semi-supervised pixel-level contrastive learning on a sparse set of hard negative pixels. Zhou *et al.* [42] designed the cross-set pixel-wise contrastive learning, which facilitates feature representation ability and feature alignment under perturbations. Alonso *et al.* [33] maintained a memory bank that is updated across the whole dataset, and then enforced the network to yield similar pixel-level feature representations for same-class samples. Wang *et al.* [34] proposed to apply regularization on the structure of the feature cluster, which is expected to increase the intra-class compactness in feature space. Zhong *et al.* [35] combined consistency regularization and contrastive learning, which simultaneously constrains the label-space consistency property between images under different perturbations and the feature space contrastive property among different pixels. Liang *et al.* [36] proposed to reason with logic-induced diagnoses and enable the recovery of erroneous pseudo labels, which could ultimately alleviate the notorious error accumulation problem. Tian *et al.* [43] proposed to conduct contrastive learning in both low-level image space and high-level feature space to maximize the utilization of available unlabeled data.

C. Fully-supervised Dual-modal Semantic Segmentation

In recent years, many methods [9]–[13], [44]–[47] have been proposed to jointly use the two modalities (i.e., 3D point clouds and images) to improve the semantic segmentation performances. Dai *et al.* [9] proposed to project the multi-view image features to the voxels and merge the multi-view features with the voxel features for better performance. Considering the computational complexity of the voxel representation, Jaritz *et al.* [10] designed a feature aggregation module to aggregate the 3D features projected from images to the original point clouds. Jaritz *et al.* [46] proposed to mutually project the sampled

image and point cloud features, and minimize the distribution discrepancies between the dual-modal features. Hu *et al.* [47] designed a bidirectional projection module where the point cloud and image features could interact with each other so that the advantages of these two modalities could be combined for better performance. Based on [47], Wang *et al.* [48] leveraged the semantic information to further enhance the mid-level features, which is proved to be helpful for improving both point cloud and image segmentation performances. Zhuang *et al.* [13] proposed a collaborative fusion scheme to exploit perceptual information from two modalities. Yan *et al.* [12] proposed a general training scheme to acquire semantic and structural information from the dual-modal data by distilling the information of 2D images to the 3D network. Li *et al.* [49] proposed a method named MSeg3D. It utilizes joint intra-modal feature extraction and inter-modal feature fusion to mitigate the modality heterogeneity and explores the asymmetric multi-modal diversified augmentation transformations for effective training.

The above-mentioned fully-supervised methods require expensive cost for labeling, while the proposed PD-Net could simultaneously segment the point clouds and images with a small number of 3D labels and no 2D label needed.

III. METHODOLOGY

In this section, we introduce our proposed PD-Net in detail. Firstly, we describe the overall architecture of the proposed network. Then, we introduce the designed dual-modal fusion module, consistency loss function, and pseudo-label optimization module respectively. Finally, we present the total loss function of the proposed network.

A. Architecture

The architecture of the proposed PD-Net is shown in Figure 2a. As seen from this figure, PD-Net employs a parallel two-stream structure: an original stream and a pseudo-label prediction stream. The original stream is utilized to simultaneously segment the point clouds and images. The pseudo-label prediction stream is utilized to predict pseudo labels for the unlabeled point clouds and their corresponding images for self-training in the original stream. The parameters of the pseudo-label prediction stream \mathbf{W}_{pl} are updated according to the original stream based on the EMA method [24]. The EMA method could retain the historical information via a progressive-update strategy, which could mitigate the negative influence brought by the false pseudo labels. Specifically, the updated parameters of the original stream are denoted as \mathbf{W}'_{ori} . In the s -th training step, the updated parameters of the pseudo-label prediction stream \mathbf{W}'_{pl} are formulated as:

$$\mathbf{W}'_{pl} = \alpha \times \mathbf{W}_{pl} + (1 - \alpha) \times \mathbf{W}'_{ori}, \quad (1)$$

where $\alpha = \min(1 - \frac{1}{s+1}, t_{ema})$, and t_{ema} is a predetermined threshold. The labeled point clouds and their corresponding images are trained in the original stream, and the labels of images are projected from the labels of point clouds. The unlabeled point clouds and their corresponding images are trained in both two streams, their corresponding outputs of

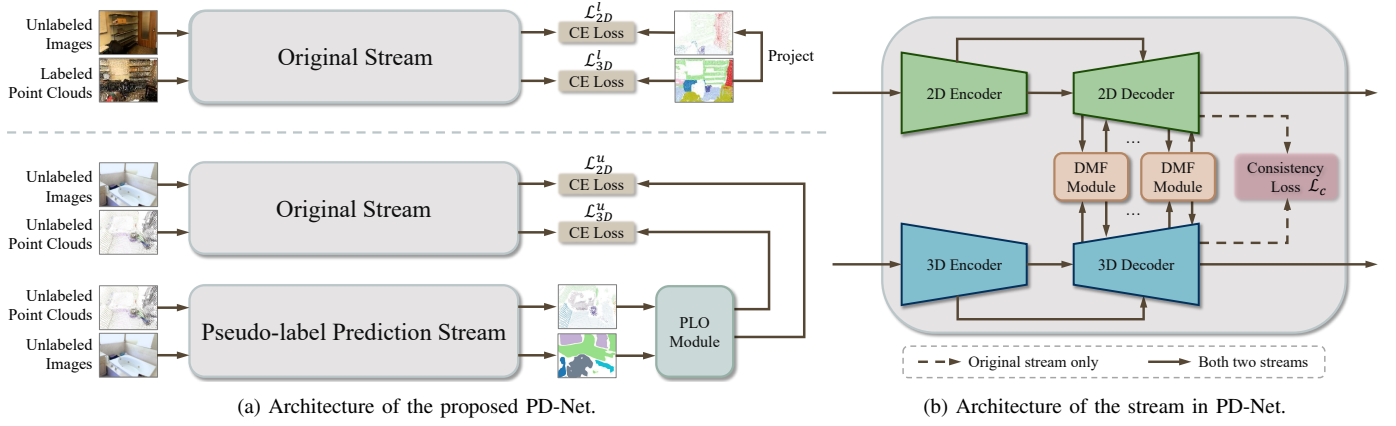


Fig. 2. Architecture of the proposed PD-Net and the original stream / the pseudo-label prediction stream in PD-Net. The proposed PD-Net contains an original stream and a pseudo-label prediction stream. The Pseudo-label Optimization (PLO) module is utilized to optimize the pseudo labels output by the pseudo-label prediction stream. CE Loss represents the cross entropy loss. The labeled point clouds and their corresponding images are only trained in the original stream, while the unlabeled point clouds and their corresponding images are trained in both two streams. The stream in PD-Net contains 3D and 2D encoder-decoder branches for dual-modal data, and multiple Dual-modal Fusion (DMF) modules to fuse the dual-modal latent features. The consistency loss function is utilized to constrain the dual-modal output features in the original stream.

the original stream are supervised by the pseudo labels output from the Pseudo-label Optimization (PLO) module.

In PD-Net, the original stream and the pseudo-label prediction stream have the same architecture, as shown in Figure 2b. Both two streams contain a 3D encoder-decoder branch, a 2D encoder-decoder branch, and multiple Dual-modal Fusion (DMF) modules. The 3D and 2D encoder-decoder branches are utilized to extract features from point clouds and images respectively.

B. Dual-modal Fusion Module

The DMF module is used to fuse the latent features of point clouds and images at each layer of the 3D and 2D decoders. The consistency loss term is explored to constrain the consistency of the 3D features and 2D features in the original stream. The PLO module is utilized to optimize the coarse pseudo labels output by the pseudo-label prediction stream.

The Dual-modal Fusion (DMF) module is designed to fuse the 3D and 2D latent features. Considering the inherent domain gap between the two modalities, we only fuse the latent features of the paired points and pixels. And the point-to-pixel correspondences could be easily calculated according to the pre-calibrated intrinsic and extrinsic parameters of the sensors. The coordinates of the paired points and pixels are denoted as $\{p_i, x_i\}_{i=1}^N$, where $p_i \in \mathbb{R}^3$ is the coordinate of the point, $x_i \in \mathbb{Z}^2$ is the coordinate of the pixel, and N is the number of matching pairs. The DMF module takes the paired 3D feature $f(p_i)$ and 2D feature $f(x_i)$ from the current 3D and 2D decoder layers as input, and outputs the fused 3D feature $g(p_i)$ and fused 2D feature $g(x_i)$, which are further fed into the next 3D and 2D decoder layers, respectively.

Learning 3D fused features: Multi-head attention-based fusion mechanism is employed to fuse the paired latent features in the DMF module. Figure 3 illustrates the calculation process of the 3D fused feature $g(p_i)$. Specifically, Minkowski convolution [50] operation and convolution operation are

performed on $f(p_i)$ and $f(x_i)$ respectively to extract their corresponding key feature $K(\cdot)$, query feature $Q(\cdot)$, and value feature $V(\cdot)$. Then, the dot product, summation, and Softmax operations are performed on the 3D key feature $K(p_i)$ and 2D query feature $Q(x_i)$ to obtain the 3D attention map $A(p_i)$. Weighted summation is performed on $A(p_i)$ and 3D value feature $V(p_i)$ to obtain the multi-head attention feature, which is extended to the same dimension with $f(p_i)$ by a Minkowski convolution layer and concatenated with $f(p_i)$. Finally, the concatenated feature passes through a Minkowski convolution layer to output the 3D fused feature.

The above-mentioned calculation process of the 3D fused feature $g(p_i)$ could be formulated as:

$$g(p_i) = \mathbf{M} \left(\mathbf{M} \left(\sum (A(p_i) \odot V(p_i)) \right) \oplus f(p_i) \right), \quad (2)$$

where $A(p_i) = \text{Softmax} \left(\sum (K(p_i) \odot Q(x_i)) \right)$, \mathbf{M} denotes the Minkowski convolution, \odot denotes the dot product, and \oplus denotes the concatenation.

Learning 2D fused features: Similarly, the calculation process of the 2D fused feature $g(x_i)$ is formulated as:

$$g(x_i) = \mathbf{C} \left(\mathbf{C} \left(\sum (A(x_i) \odot V(x_i)) \right) \oplus f(x_i) \right), \quad (3)$$

where $A(x_i) = \text{Softmax} \left(\sum (K(x_i) \odot Q(p_i)) \right)$, and \mathbf{C} denotes the commonly-used convolution.

We utilize multiple DMF modules to fuse the 3D and 2D latent features in multiple scales. Compared with direct feature concatenation, the proposed multi-head attention-based fusion mechanism could extract more discriminative and informative features from the two modalities, which would be demonstrated in Subsection IV-C.

C. Consistency Loss

In order to constrain the consistency between the learned dual-modal features in the output feature spaces of the original stream, we propose the consistency loss term.

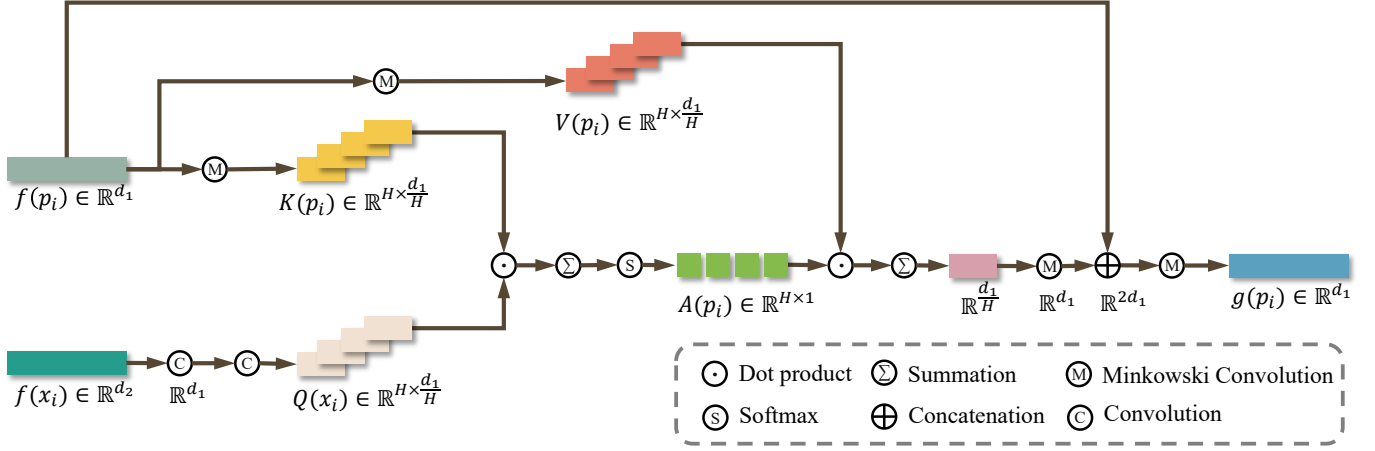


Fig. 3. The calculation process of the 3D fused feature $g(p_i)$ in the dual-modal fusion module. The dimensions of the key feature $K(\cdot)$, query feature $Q(\cdot)$, and value feature $V(\cdot)$ are the results of dividing the dimensions of their corresponding latent feature $f(\cdot)$ by the head number H . d_1 and d_2 denote the dimensions of 3D features and 2D features respectively. The attention-based mechanism in the dual-modal fusion module facilitates adaptively learning complementary information from dual-modal data.

The paired output features of the 3D and 2D branches in the original stream are denoted as $\{\mathbf{y}(p_i), \mathbf{y}(x_i)\}_{i=1}^N$. Then, the proposed consistency loss term $Loss_c$ is formulated as:

$$\mathcal{L}_c = \frac{1}{N} \sum_{i=1}^N \|\mathbf{y}(p_i) - \mathbf{y}(x_i)\|_2^2, \quad (4)$$

where $\|\cdot\|_2$ denotes the L2-norm.

Overall, the dual-modal fusion module and the consistency loss term fuse the dual-modal features in different levels of feature space.

D. Pseudo-label Optimization Module

Due to the inherent limitations of the two modalities (i.e., the lack of texture information in point clouds and the lack of depth information in images), the 3D and 2D encoder-decoder

branches tend to predict pseudo labels for objects according to their geometric structures and textures respectively.

In order to guarantee the effectiveness of the self-training of unlabeled data in the original stream, we propose the Pseudo-label Optimization (PLO) module based on a voting mechanism to improve the reliability of the pseudo labels. The PLO module is only utilized for the paired unlabeled points and pixels, whose coordinates are denoted as $\{p_i^u, x_i^u\}_{i=1}^{N_u}$, where N_u is the number of the unlabeled matching pairs. It takes the coarse pseudo labels of the unlabeled points and pixels as input and outputs their corresponding optimized pseudo labels.

Specifically, the coarse 3D and 2D pseudo labels of the paired unlabeled point p_i^u and pixel x_i^u , which are output by the pseudo-label prediction stream, are denoted as $c_{3D}(p_i^u)$ and $c_{2D}(x_i^u)$ respectively. And the process of optimizing 3D (top) and 2D (bottom) pseudo labels by the PLO module is shown in Figure 4.

Optimization of 3D pseudo labels: Firstly, the coarse 2D pseudo label $c_{2D}(x_i^u)$ is projected to its paired point to obtain the projected 3D pseudo label $c_{3D}(x_i^u)$. The coarse 3D pseudo label $c_{3D}(p_i^u)$ is retained if it is consistent with $c_{3D}(x_i^u)$. Otherwise, a confidence-based filtering mechanism is utilized. The confidence of the coarse 3D pseudo label is simply the value of the $c_{3D}(p_i^u)$ -th dimension of the 3D output feature in the pseudo-label prediction stream, which is denoted as $\gamma_c(p_i^u, c_{3D}(p_i^u))$. The coarse 3D pseudo label $c_{3D}(p_i^u)$ is retained if its confidence is larger than the predetermined confidence threshold t_{conf} . Otherwise, the coarse 3D pseudo label is deleted. The above process could be formulated as:

$$\hat{c}_{3D}(p_i^u) = \begin{cases} c_{3D}(p_i^u) & , c_{3D}(x_i^u) = c_{3D}(p_i^u) \text{ or} \\ & \gamma_c(p_i^u, c_{3D}(p_i^u)) > t_{conf}, \\ \text{deleted} & , \text{other,} \end{cases} \quad (5)$$

where $\hat{c}_{3D}(p_i^u)$ is the optimized 3D pseudo label.

Optimization of 2D pseudo labels: As seen in the bottom part of Figure 4, the projected 2D pseudo label $c_{2D}(p_i^u)$ is

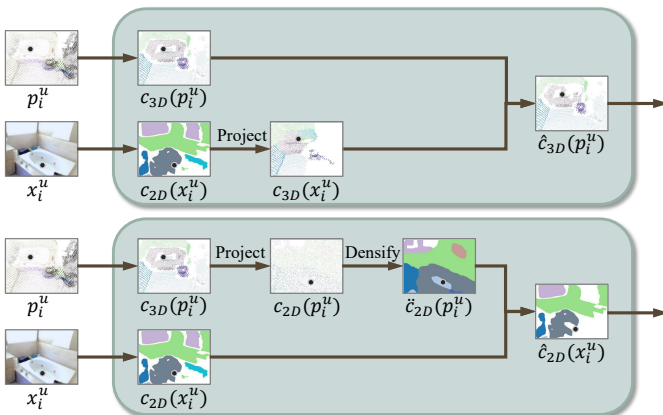


Fig. 4. The optimization process of 3D (top) and 2D (bottom) pseudo-labels. The coarse 2D pseudo labels are projected to point clouds to obtain the projected 3D pseudo labels. The coarse 3D pseudo labels are densified after being projected to the image plane to obtain the projected 2D pseudo labels. The black point denotes the pseudo label that is deleted by the pseudo-label optimization module.

sparse, which leaves the majority of the pixels unprojected. To address this issue, we project each 3D output feature in the pseudo-label stream into the image plane and perform average pooling in the local areas for the unprojected pixels. The dimension with the largest value in the pooled output feature is selected as the dense 2D pseudo label $\hat{c}_{2D}(p_i^u)$.

Similarly, the optimization process of the 2D pseudo label is formulated as:

$$\hat{c}_{2D}(x_i^u) = \begin{cases} c_{2D}(x_i^u) & , \hat{c}_{2D}(p_i^u) = c_{2D}(x_i^u) \text{ or} \\ & \gamma_c(x_i^u, c_{2D}(x_i^u)) > t_{conf}, \\ \text{deleted} & , \text{other,} \end{cases} \quad (6)$$

where $\hat{c}_{x_D}(x_i^u)$ is the optimized 2D pseudo label.

E. Total Loss Function

As depicted in Figure 2a, four cross-entropy loss terms are employed for the labeled point clouds and their corresponding images, and the unlabeled point clouds and their corresponding images, which are denoted as \mathcal{L}_{3D}^l , \mathcal{L}_{2D}^l , \mathcal{L}_{3D}^u , and \mathcal{L}_{2D}^u . And their targets are the ground truth 3D labels, projected 2D labels, optimized 3D pseudo labels, and optimized 2D pseudo labels, respectively. Since a normalization is performed during the calculation process of the cross-entropy loss by dividing the length of the input vector, the four losses have a same order of magnitude, and we thus assign a same weight to them.

Combined with the consistency loss term \mathcal{L}_c in Subsection III-C, the total loss function \mathcal{L}_{total} is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{3D}^l + \mathcal{L}_{2D}^l + \mathcal{L}_{3D}^u + \mathcal{L}_{2D}^u + \lambda_c \mathcal{L}_c, \quad (7)$$

where λ_c is the weight of the consistency loss term.

IV. EXPERIMENTS

A. Experimental Setup

Dataset: We evaluate the proposed PD-Net on the ScanNet dataset [51], which contains 1613 indoor point clouds recon-

structed from depth images. In addition, the ScanNet dataset contains more than 2.5×10^6 RGB images, and each point cloud corresponds to more than 5000 images. The intrinsic and extrinsic parameters of the sensors are also provided, which enables the calculation of the point-to-pixel correspondences. Both the 3D point clouds and images in the ScanNet dataset are annotated with 20 semantic categories.

Implementation details and metrics: In this work, the 3D and 2D encoder-decoder branches, which are based on MinkowskiNet18A [50] and ResNet34 [52] respectively, both use the U-Net [55] architectures. For each 3D point cloud, we randomly sample 3 images from its corresponding image set for dual-modal training. The weight threshold t_{ema} in the EMA method is set to 0.999, the head number in the dual-modal fusion module is set to 4, the confidence threshold t_{conf} in the pseudo-label optimization module is set to 0.9 for deleting the false labels with low confidences, and the consistency loss weight λ_c is set to 5. The voxel size is set to 5cm for efficient training. We apply the Stochastic Gradient Descent (SGD) optimizer with a base learning rate of 0.01. The batch size and epoch number is set as 16 and 150 respectively. The data augmentation techniques include random flip, random rotation, and color jittering.

For evaluating the performance of semi-supervised segmentation, we split the training point clouds into a labeled set and an unlabeled set. Specifically, we randomly sample the labeled point clouds from the training point clouds with two different ratios (i.e., 20% and 10%). Only the labeled point clouds and their corresponding images are trained in the first 100 epochs for a more stable semi-supervised training. The unlabeled point clouds and their corresponding images are incorporated into training in the last 50 epochs.

We use mean Intersection over Union (mIoU), mean Accuracy (mAcc), and Overall Accuracy (OA) as the evaluation metrics for both 3D and 2D semantic segmentation.

TABLE I
EVALUATION RESULTS ON THE VALIDATION SET OF THE SCANNET DATASET [51]. THE BEST RESULTS ARE IN BOLD IN EACH METRIC.

	Method	Point Cloud			Image		
		mIoU	mAcc	OA	mIoU	mAcc	OA
20%	MinkowskiNet18A [50]	59.31	67.92	84.13	-	-	-
	ResNet34 [52]	-	-	-	45.04	59.20	74.96
	Deng <i>et al.</i> [21]	55.12	63.61	82.43	-	-	-
	TCSM-V2 [25]	-	-	-	52.65	61.08	80.17
	CPS [53]	-	-	-	55.23	64.97	81.86
	π -Model [23]	60.41	69.08	84.34	51.09	60.08	78.81
	Mean Teacher [24]	61.12	69.47	84.72	51.82	60.56	79.68
	Pseudo-Labels [54]	60.64	69.27	84.51	53.01	62.17	80.68
	PD-Net	63.38	71.61	86.28	60.17	70.78	83.29
10%	MinkowskiNet18A [50]	52.27	61.19	80.81	-	-	-
	ResNet34 [52]	-	-	-	43.00	55.09	72.75
	Deng <i>et al.</i> [21]	52.38	60.76	81.18	-	-	-
	TCSM-V2 [25]	-	-	-	46.98	57.61	74.25
	CPS [53]	-	-	-	48.01	58.76	76.02
	π -Model [23]	54.84	63.45	81.54	47.72	57.56	75.57
	Mean Teacher [24]	55.24	63.70	81.79	46.63	57.39	74.10
	Pseudo-Labels [54]	54.47	63.38	81.40	46.83	57.51	74.12
	PD-Net	58.38	67.23	83.68	50.80	60.77	79.38

TABLE II

EVALUATION RESULTS ON THE VALIDATION SET OF THE SCANNet. * DENOTES THAT POINT CLOUDS AND IMAGES ARE TRAINED JOINTLY. † DENOTES THAT ONLY DEPTH IMAGES ARE USED FOR TRAINING. § DENOTES THAT ONLY RGB IMAGES ARE USED FOR TRAINING. # DENOTES THAT RGB-D IMAGES ARE USED FOR TRAINING.

Method	3D mIoU	Method	2D mIoU
Pointnet++ [56]	53.5	ERFNetEnc § [57]	51.7
PoinConv [58]	61.0	AdaptNet++ § [59]	52.9
PointASNL [60]	63.5	AdaptNet++ † [59]	53.8
MVPNet [10]	65.0	Deeplabv3 § [61]	56.1
Minkowski42 [50]	68.0	ERFNetEnc † [57]	56.7
KPConv [62]	69.2	SSMA # [63]	61.1
JointPointBased [45]	69.2	RFBNet # [64]	62.6
PointTransformer [5]	70.6	GRBNet # [65]	62.6
BPNet * [47]	73.9	MCA-Net # [66]	64.3
StratifiedPT [67]	74.3	BPNet * [47]	71.9
PD-Net * (20%)	63.4	PD-Net * (20%)	60.2
PD-Net * (10%)	58.4	PD-Net * (10%)	50.8

B. Comparative Evaluation

Considering the 3D and 2D encoder-decoder branches are based on MinkowskiNet18A [50] and ResNet34 [52] respectively, we evaluate the performances of the baseline models (i.e., MinkowskiNet18A and ResNet34) by training on the labeled set. Then, we compare the proposed PD-Net with several semi-supervised uni-modal semantic segmentation methods for point clouds [21] and images [25], [53]. In addition, several typical semi-supervised learning methods are extended to tackle the semi-supervised dual-modal semantic segmentation task, including π -Model [23], Mean Teacher [24], and Pseudo-Labels [54]. We evaluate these semi-supervised learning methods based on MinkowskiNet18A and ResNet34 while retaining their other experimental settings for a fair comparison. All these comparative methods utilize the same labeled set and unlabeled set.

Table I reports the quantitative results of the proposed PD-Net and comparative methods on the validation set of the ScanNet [51]. As seen from this table, in two different labeled-ratio settings, the proposed PD-Net outperforms all the comparative methods in point cloud segmentation and image segmentation tasks. The proposed PD-Net outperforms Pseudo-Labels, Deng *et al.* [21], TCSM-V2 [25], and CPS [53], because it could effectively utilize the complementary information from the point clouds and images, and mitigate the negative impact brought by the falsely predicted pseudo labels to some extent. And the proposed PD-Net outperforms the π -Model and Mean Teacher, probably because the consistency constraints between the latent features from different scales are more powerful than the constraints between the output features from different transformations.

Figure 5 and Figure 6 visualize the qualitative results of point cloud segmentation and image segmentation respectively. As seen from these figures, the proposed PD-Net predicts more accurately than MinkowskiNet18A and ResNet34 in both two labeled-ratio settings. We highlight the key regions in the dark blue boxes. The visualization results demonstrate that the proposed PD-Net yields promising performances in 3D and 2D segmentation with only a small number of labeled point clouds required.

In addition, we compare the proposed PD-Net, which is trained in a semi-supervised manner, with several typical fully-supervised semantic segmentation methods for point clouds [5], [10], [45], [47], [50], [56], [58], [60], [62], [67] and for images [47], [57], [59], [61], [63]–[66] on the validation set of the ScanNet. The corresponding results are reported in Table II. As seen from this table, the PD-Net trained under the 20%-labeled setting achieves comparable results with the comparative fully-supervised methods, which further demonstrates the effectiveness of the proposed PD-Net.

C. Ablation Study

The effectiveness of each key element in the proposed PD-Net is verified by conducting ablation studies on the validation set of ScanNet dataset [51]. The following models under two labeled-ratio settings are compared:

- Baseline: The 3D and 2D encoder-decoder branches (based on MinkowskiNet18A [50] and ResNet34 [52]) trained on the labeled point clouds and their corresponding images.
- Model A: Based on Baseline, the pseudo-label supervision for unlabeled data is added.
- Model B: Based on Model A, the EMA method [24] is utilized to update the parameters of the pseudo-label prediction stream.
- Model C: Based on Model B, the Pseudo-label Optimization (PLO) module and the consistency loss term are added.
- PD-Net (the whole model): Based on Model C, the Dual-modal Fusion (DMF) module is added.

The corresponding results are reported in Table III. As seen from this table, Model A performs better than the Baseline, indicating that the coarse pseudo labels generated by the pseudo-label prediction stream could supervise the unlabeled data to some extent. Model B makes further progress based on Model A, demonstrating that using the historical information to update the parameters of the pseudo-label prediction stream is superior to the common updating strategy. The performances of Model C are promoted based on Model B, which is attributed to the consistency constraints between the 3D and 2D output features and the optimization of the pseudo labels.

TABLE III

ABLATION STUDIES OF THE INVOLVED COMPONENTS.

	Model	Point Cloud			Image		
		mIoU	mAcc	OA	mIoU	mAcc	OA
20%	Baseline	59.31	67.92	84.13	45.04	59.20	74.96
	Model A	60.64	69.27	84.51	53.01	62.17	80.68
	Model B	61.46	70.20	85.03	55.02	64.59	80.86
	Model C	62.26	72.11	85.56	58.30	69.83	82.56
	PD-Net	63.38	71.61	86.28	60.17	70.78	83.29
10%	Baseline	52.27	61.19	80.81	43.00	55.09	72.75
	Model A	54.47	63.38	81.40	46.83	57.51	74.12
	Model B	54.72	67.66	81.52	47.07	57.94	75.34
	Model C	56.87	66.49	82.90	50.14	59.61	79.18
	PD-Net	58.38	67.23	83.68	50.80	60.77	79.38

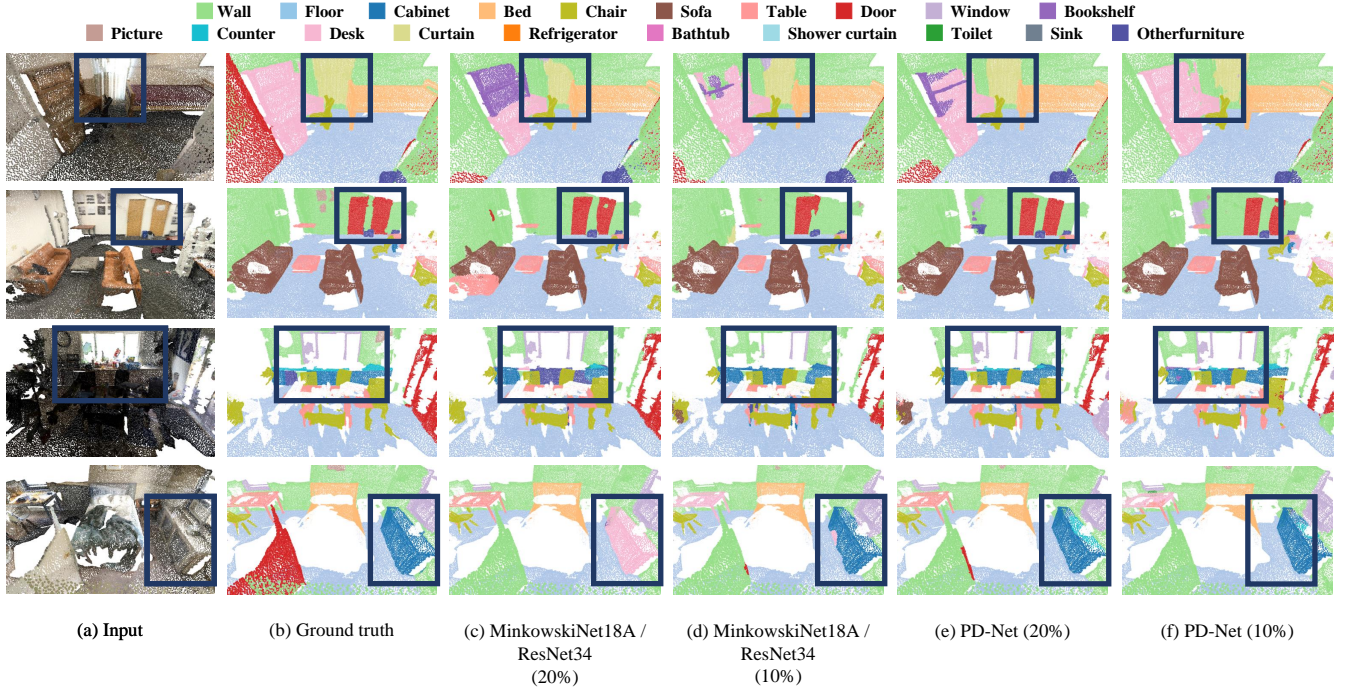


Fig. 5. Qualitative results of point cloud segmentation on the validation set of the ScanNet [51]. The segmentation results of the baseline model (MinkowskiNet18A [50] and ResNet34 [52]) and our proposed PD-Net in two different labeled-ratio settings (20% and 10%) are visualized.

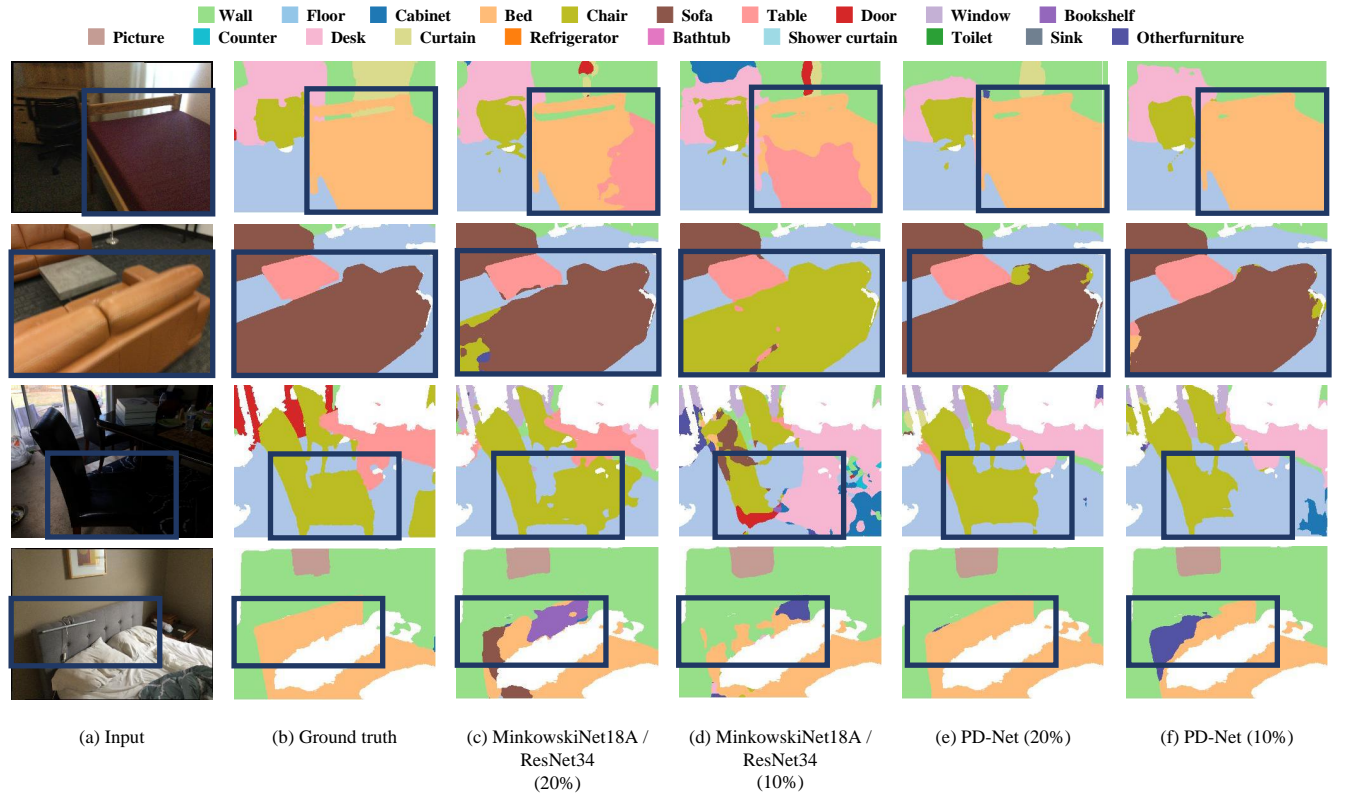


Fig. 6. Qualitative results of image segmentation on the validation set of the ScanNet [51]. The segmentation results of the baseline model (MinkowskiNet18A [50] and ResNet34 [52]) and our proposed PD-Net in two different labeled-ratio settings (20% and 10%) are visualized.

TABLE IV
RESULTS OF PD-NET WITH DIFFERENT FUSION MODULES.

	Module	Point Cloud			Image		
		mIoU	mAcc	OA	mIoU	mAcc	OA
20%	BP [47]	62.63	71.04	85.80	58.40	69.81	82.10
	DMF	63.38	71.61	86.28	60.17	70.78	83.29
10%	BP [47]	57.19	66.61	83.01	50.20	59.74	79.02
	DMF	58.38	67.23	83.68	50.80	60.77	79.38

TABLE V
RESULTS OF PD-NET WITH DIFFERENT t_{conf} .

	t_{conf}	Point Cloud			Image		
		mIoU	mAcc	OA	mIoU	mAcc	OA
20%	0.60	62.31	70.35	85.42	58.37	68.66	81.47
	0.85	62.83	71.04	85.72	58.89	69.12	82.45
	0.90	63.38	71.61	86.28	60.17	70.78	83.29
	0.95	62.95	71.23	85.95	59.10	69.30	82.78
10%	0.60	56.14	65.87	82.36	48.79	58.61	78.05
	0.85	57.47	66.62	83.14	49.30	58.92	78.72
	0.90	58.38	67.23	83.68	50.80	60.77	79.38
	0.95	57.55	66.82	83.18	50.24	60.22	78.98

The whole PD-Net achieves the best results in most cases, probably because the DMF module could effectively fuse the dual-modal latent features.

To further verify the superiority of the DMF module, we replace the DMF module with a similar module for dual-modal feature fusion, while keeping the experimental settings and other modules unchanged. Specifically, we choose the Bidirectional Projection (BP) module in BPNet [47].

The comparison results are reported in Table IV. As seen from this table, the model with DMF module achieves better segmentation performances, indicating that the multi-head attention-based mechanism has stronger fusion ability than the view fusion strategy in the BP module which simply learns the impact factors for each view at every point.

D. Discussion

As seen from Tables I and III on the ScanNet [51] dataset, the following points could be revealed:

- Comparison between the proposed PD-Net and its two baseline methods (MinkowskiNet18A [50] and ResNet34 [52]): Table III reports the ablation results on the main components of PD-Net, where the two baselines are trained with only labeled data. As seen from this table, Models {A, B, C}, which introduce different operations on pseudo labels of unlabeled data gradually, perform better than each baseline. These results demonstrate that the usage of unlabeled data for semi-supervised learning is one reason of why PD-Net outperforms its baselines. Moreover, the whole model PD-Net, which integrates Model C and the explored dual-modal fusion module, performs better than Models {A, B, C} and the baselines, demonstrating that the dual-modal fusion module is another reason of PD-Net's superiority to its baselines.

- Comparison between PD-Net and the uni-modal methods [21], [25], [53]: On one hand, compared with the 3D segmentation method [21], it is noted from Table I that the baseline [50] of PD-Net performs better, demonstrating that a strong baseline is one reason of why PD-Net outperforms [21]. According to this result and the above discussion on the baselines, it could be further speculated that the two reasons of why PD-Net outperforms the baselines are also the reasons of why PD-Net outperforms the 3D segmentation method [21]. On the other hand, as seen from Tables I and III, the image segmentation methods [25], [53] perform better than the baseline [52], but worse than Model C (with the pseudo-label optimization module and the consistency loss) and the whole model (Model C + the dual-modal fusion module), demonstrating that the pseudo-label optimization module, the consistency loss, and the dual-modal fusion module are the reasons of PD-Net's superiority to the image segmentation methods [25], [53].
- Comparison between PD-Net and the dual-modal methods [23], [24], [54]: As seen from Tables I and III, both Model C (with the pseudo-label optimization module and the consistency loss) and the whole model (Model C + the dual-modal fusion module) always perform better than the three dual-modal methods, demonstrating that the pseudo-label optimization module, the consistency, and the dual-modal fusion module are the reasons of PD-Net's superiority.

It is noted from the above discussion that compared with different existing methods, the reasons of PD-Net's superiority are not always exactly identical, however, the pseudo-label optimization module, the consistency loss, and the dual-modal fusion module in PD-Net are always the factors for guaranteeing PD-Net's superiority to all the comparative methods, which support the claimed contributions in Section I.

E. Analysis on Hyper-parameters

In this section, we provide more analysis on some hyper-parameters, including the confidence threshold t_{conf} in the

TABLE VI
RESULTS OF PD-NET WITH DIFFERENT λ_c .

	λ_c	Point Cloud			Image		
		mIoU	mAcc	OA	mIoU	mAcc	OA
20%	0	61.72	70.65	84.87	55.28	64.63	80.52
	0.2	62.91	71.14	85.96	56.86	67.82	81.08
	1	63.21	71.42	86.08	59.31	69.53	83.14
	5	63.38	71.61	86.28	60.17	70.78	83.29
	10	62.82	70.92	85.99	59.17	68.82	83.66
	50	61.98	69.84	84.31	56.50	64.92	80.69
10%	0	55.11	66.07	81.96	47.56	57.08	75.69
	0.2	57.90	67.82	83.29	48.88	58.12	78.12
	1	57.71	66.94	83.22	50.20	60.18	78.83
	5	58.38	67.23	83.68	50.80	60.77	79.38
	10	58.03	67.12	83.29	50.71	60.25	79.23
	50	54.87	66.47	82.37	47.33	56.76	75.48

TABLE VII
RESULTS OF PD-NET WITH DIFFERENT HEAD NUMBERS.

	H	Point Cloud			Image		
		mIoU	mAcc	OA	mIoU	mAcc	OA
20%	3	62.08	70.11	85.58	57.70	67.52	81.87
	4	63.38	71.61	86.28	60.17	70.78	83.29
	5	62.16	70.19	85.57	56.44	65.53	82.04
10%	3	56.20	66.19	82.46	49.23	58.89	78.75
	4	58.38	67.23	83.68	50.80	60.77	79.38
	5	56.47	66.43	82.51	48.85	58.19	78.23

pseudo-label optimization module, the weight of consistency loss λ_c , and the voxel size. The experiments are conducted on the validation set of ScanNet [51].

Effect of confidence threshold. As seen in (5) and (6), the confidence threshold t_{conf} affects the quality of the optimized pseudo labels. We evaluate the proposed PD-Net with $t_{conf} = \{0.6, 0.85, 0.9, 0.95\}$ to estimate the insensitive range of t_{conf} . The corresponding results are reported in Table V, which indicate that our model achieves relatively stable performances when t_{conf} ranges in $[0.85, 0.95]$ and t_{conf} with a lower value (i.e., $t_{conf} = \{0.6\}$) may impair the quality of pseudo labels, and thus deteriorate the performances. The model achieves the best performances when $t_{conf} = 0.9$.

Effect of the weight for consistency loss. As seen in (7), the loss weight λ_c affects the balance between the cross-entropy loss terms and \mathcal{L}_c . We evaluate the proposed PD-Net with $\lambda_c = \{0, 0.2, 1, 5, 10, 50\}$ to estimate the insensitive range of λ_c . The corresponding results are reported in Table VI, which indicate that our method is relatively insensitive to λ_c when λ_c ranges in $[0.2, 10]$ and the performances drop evidently when λ_c is set to extreme values (i.e., $\lambda_c = \{0, 50\}$). The model achieves the best performances in most cases when $\lambda_c = 5$.

Effect of head number. As stated in [71], multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions, which indicates that more heads could enhance the representation ability of the model. However, as revealed in [72], the majority of attention heads can be removed without deviating too much from the original performance and most heads are redundant given the rest of the model at test time. And too many heads may result in overfitting considering the strong representation

ability on the training set. Thus, the head number H in the DMF module affects the quality of the fused features and the performance of the model.

Here, we evaluate the proposed PD-Net with $H = \{3, 4, 5\}$. And the corresponding results are reported in Table VII. As seen from this table, compared with the results when H is set as 4, the performances of 3D and 2D segmentation degrade when H is set as 3 or 5. This phenomenon is consistent with the revealed points in [71] and [72], which indicates that an appropriate head number needs to be set.

Effect of voxel size. In previous experiments, we set the voxel size to 5cm for efficient training. We evaluate PD-Net and baseline models with voxel size = $\{5\text{cm}, 2\text{cm}\}$, and the corresponding results are reported in Table VIII. The results show that decreasing voxel size could simultaneously improve the performances of 3D and 2D segmentation, demonstrating that fine-grained voxels could provide higher-quality 3D information and better boost the 2D semantic segmentation. But in the meanwhile, smaller voxel size inevitably brings higher computational cost and causes longer forward time, as seen in the last column of Table VIII.

F. PD-Net on NYUv2

The NYUv2 dataset [68] is a widely-used RGB-D dataset, which contains 1449 densely annotated pairs of aligned RGB and depth images. Following 3DMV [9], BPNet [47], and SemAffiNet [48], we additionally evaluate PD-Net on NYUv2 dataset by converting the depth images to pseudo point clouds according to the camera's pose matrix. We adopt the 13-class configuration for a fair comparison with the comparative methods [9], [28], [47], [51], [69], [70].

We utilize the pixel-level classification accuracy metric and report the results in Table IX. As seen from this table, our proposed PD-Net achieves comparable results with the comparative fully-supervised RGB-D and joint 2D-3D methods. The results on the NYUv2 dataset demonstrate the effectiveness and generality of PD-Net.

G. Limitations

Figure 7 reports comparative failure cases on the ScanNet [51] dataset between the proposed PD-Net and the Mean-Teacher [24] that has a relatively better performance than the other comparative methods. As seen from the red rectangles

TABLE VIII
RESULTS OF PD-NET AND BASELINE MODELS WITH DIFFERENT VOXEL SIZES.

	Model	Point Cloud			Image			Time
		mIoU	mAcc	OA	mIoU	mAcc	OA	
20%	Baseline (5cm)	59.31	67.92	84.13	45.04	59.20	74.96	1.3s
	Baseline (2cm)	59.45	70.35	83.73	49.84	61.00	76.52	3.1s
	PD-Net (5cm)	63.38	71.61	86.28	60.17	70.78	83.29	2.5s
	PD-Net (2cm)	64.72	75.68	88.17	62.42	73.54	86.63	7.3s
10%	Baseline (5cm)	52.27	61.19	80.81	43.00	55.09	72.75	1.3s
	Baseline (2cm)	55.00	65.16	81.22	44.68	56.53	72.73	3.1s
	PD-Net (5cm)	58.38	67.23	83.68	50.80	60.77	79.38	2.5s
	PD-Net (2cm)	59.05	71.01	85.50	58.16	68.63	82.80	7.3s

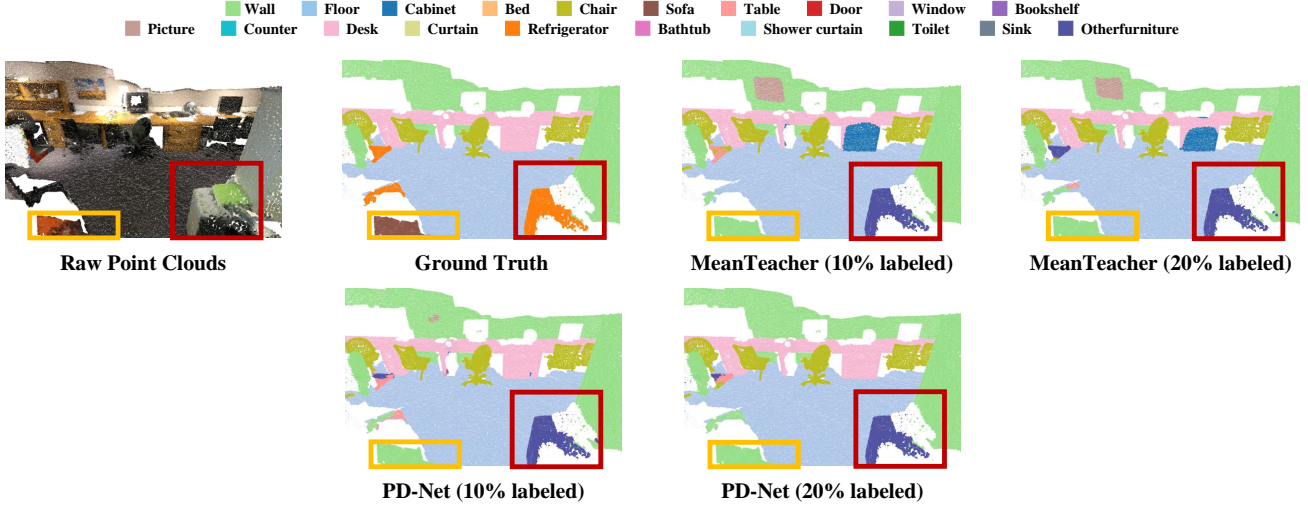


Fig. 7. Failure cases of the proposed method and MeanTeacher [24] on the ScanNet [51] dataset with 10% and % labeled data. The red and yellow rectangles indicate the failure cases.

TABLE IX
SEMANTIC SEGMENTATION RESULTS ON NYUV2 [68] USING DENSE
PIXEL-LEVEL CLASSIFICATION ACCURACY METRIC.

NYUv2	Accuracy
SceneNet [28]	52.5
Hermans <i>et al.</i> [69]	54.3
SemanticFusion [70]	59.2
ScanNet [51]	60.7
3DMV [9]	71.2
BPNet [47]	73.5
SemAffiNet [48]	78.3
PD-Net (20%)	71.7

of this figure, the proposed PD-Net wrongly recognizes a refrigerator as other furniture, while MeanTeacher also makes a similar mistake. This is mainly due to the small amount of training data for refrigerators. Additionally, as highlighted by the yellow rectangles, the proposed method fails to accurately recognize a small portion of a sofa, and the comparative method faces the same issue. This is because the local part of the sofa has a significantly different structure compared to a complete sofa.

In summary, although the proposed method has shown its superiority to some comparative methods, it is still hard for both the proposed method and comparative method to make accurate predictions for (i) objects with less training data and (ii) local object parts that differ significantly from their corresponding complete objects. It would be one of our future works to alleviate the above issues.

V. CONCLUSIONS

We propose a parallel dual-stream network, called PD-Net, to handle the semi-supervised dual-modal semantic segmentation task. The proposed PD-Net consists of two parallel streams (i.e., original stream and pseudo-label prediction

stream), in which the 3D and 2D encoder-decoder branches are used to extract 3D and 2D features respectively, and multiple dual-modal fusion modules are used to fuse the multi-scale dual-modal latent features. The pseudo-label optimization module is explored to improve the quality of the pseudo labels output by the pseudo-label prediction stream. Experimental results demonstrate that the proposed PD-Net not only outperforms the comparative semi-supervised methods but also achieves competitive performances with some fully-supervised methods in most cases.

REFERENCES

- [1] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3d point clouds: A survey," *IEEE TPAMI*, vol. 43, no. 12, pp. 4338–4364, 2021.
- [2] S. Deng and Q. Dong, "Ga-net: Global attention network for point cloud semantic segmentation," *SPL*, vol. 28, pp. 1300–1304, 2021.
- [3] S. Fan, Q. Dong, F. Zhu, Y. Lv, P. Ye, and F.-Y. Wang, "Scf-net: Learning spatial contextual features for large-scale point cloud segmentation," in *CVPR*, 2021, pp. 14 504–14 513.
- [4] J. Li and Q. Dong, "Open-set semantic segmentation for point clouds via adversarial prototype framework," in *CVPR*, 2023, pp. 9425–9434.
- [5] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *ICCV*, 2021, pp. 16 259–16 268.
- [6] A. Xiao, D. Guan, X. Zhang, and S. Lu, "Domain adaptive lidar point cloud segmentation with 3d spatial consistency," *IEEE TMM*, vol. 26, pp. 5536–5547, 2024.
- [7] H. Zhang, C. Wang, L. Yu, S. Tian, X. Ning, and J. Rodrigues, "Pointgt: A method for point-cloud classification and segmentation based on local geometric transformation," *IEEE TMM*, vol. 26, pp. 8052–8062, 2024.
- [8] A. Du, T. Zhou, S. Pang, Q. Wu, and J. Zhang, "Pcl: Point contrast and labeling for weakly supervised point cloud semantic segmentation," *IEEE TMM*, vol. 26, pp. 8902–8914, 2024.
- [9] A. Dai and M. Nießner, "3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation," in *ECCV*, 2018, pp. 452–468.
- [10] M. Jaritz, J. Gu, and H. Su, "Multi-view pointnet for 3d scene understanding," in *ICCVW*, 2019, pp. 1–9.
- [11] G. Krissel, M. Opitz, G. Waltner, H. Possegger, and H. Bischof, "Fuseseg: Lidar point cloud segmentation fusing multi-modal data," in *WACV*, 2020, pp. 1874–1883.
- [12] X. Yan, J. Gao, C. Zheng, C. Zheng, R. Zhang, S. Cui, and Z. Li, "2dpass: 2d priors assisted semantic segmentation on lidar point clouds," in *ECCV*, 2022.

- [13] Z. Zhuang, R. Li, Y. Li, K. Jia, Q. Wang, and M. Tan, "Perception-aware multi-sensor fusion for 3d lidar semantic segmentation," in *ICCV*, 2021, pp. 16 260–16 270.
- [14] L. Zhao, H. Zhou, X. Zhu, X. Song, H. Li, and W. Tao, "Lif-seg: Lidar and camera image fusion for 3d lidar semantic segmentation," *IEEE TMM*, vol. 26, pp. 1158–1168, 2023.
- [15] Y. Wu, J. Liu, M. Gong, P. Gong, X. Fan, A. K. Qin, Q. Miao, and W. Ma, "Self-supervised intra-modal and cross-modal contrastive learning for point cloud understanding," *IEEE TMM*, vol. 26, pp. 1626–1638, 2024.
- [16] T.-H. Wu, Y.-C. Liu, Y.-K. Huang, H.-Y. Lee, H.-T. Su, P.-C. Huang, and W. H. Hsu, "Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation," in *ICCV*, 2021, pp. 15 510–15 519.
- [17] X. Shi, X. Xu, K. Chen, L. Cai, C. S. Foo, and K. Jia, "Label-efficient point cloud semantic segmentation: An active learning approach," *arXiv preprint: 2101.06931*, 2021.
- [18] H. Li, Z. Sun, Y. Wu, and Y. Song, "Semi-supervised point cloud segmentation using self-training with label confidence prediction," *Neurocomputing*, vol. 437, pp. 227–237, 2021.
- [19] L. Jiang, S. Shi, Z. Tian, X. Lai, S. Liu, C.-W. Fu, and J. Jia, "Guided point contrastive learning for semi-supervised point cloud semantic segmentation," in *ICCV*, 2021, pp. 6423–6432.
- [20] L. Kong, J. Ren, L. Pan, and Z. Liu, "Lasermix for semi-supervised lidar semantic segmentation," in *CVPR*, June 2023, pp. 21 705–21 715.
- [21] S. Deng, Q. Dong, B. Liu, and Z. Hu, "Superpoint-guided semi-supervised semantic segmentation of 3d point clouds," in *ICRA*, 2021.
- [22] L. Li, H. P. H. Shum, and T. P. Breckon, "Less is more: Reducing task and model complexity for 3d point cloud semantic segmentation," in *CVPR*, June 2023, pp. 9361–9371.
- [23] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *ICLR*, 2017, pp. 1–13.
- [24] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NeurIPS*, 2017, pp. 1195–1204.
- [25] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE TNNLS*, vol. 32, pp. 523–534, 2020.
- [26] L. Tang, Z. Chen, S. Zhao, C. Wang, and D. Tao, "All points matter: entropy-regularized distribution alignment for weakly-supervised 3d segmentation," *NeurIPS*, vol. 36, 2024.
- [27] Q. Meng, W. Wang, T. Zhou, J. Shen, Y. Jia, and L. Van Gool, "Towards a weakly supervised framework for 3d point cloud object detection and annotation," *IEEE TPAMI*, vol. 44, no. 8, pp. 4454–4468, 2021.
- [28] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla, "Understanding realworld indoor scenes with synthetic data," in *CVPR*, 2016, pp. 4077–4085.
- [29] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high- and low-level consistency," *IEEE TPAMI*, vol. 43, pp. 1369–1379, 2019.
- [30] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *CVPR*, 2020, pp. 12 671–12 681.
- [31] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *CVPR*, 2021, pp. 2613–2622.
- [32] X. Zhao, R. Vemulapalli, P. A. Mansfield, B. Gong, B. Green, L. Shapira, and Y. Wu, "Contrastive learning for label efficient semantic segmentation," in *ICCV*, 2020, pp. 10 603–10 613.
- [33] I. Alonso, A. Sabater, D. Ferstl, L. Montesano, and A. C. Murillo, "Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank," in *ICCV*, 2021, pp. 8199–8208.
- [34] X. Wang, B. Zhang, L. Yu, and J. Xiao, "Hunting sparsity: Density-guided contrastive learning for semi-supervised semantic segmentation," in *CVPR*, 2023, pp. 3114–3123.
- [35] Y. Zhong, B. Yuan, H. Wu, Z. Yuan, J. Peng, and Y.-X. Wang, "Pixel contrastive-consistent semi-supervised semantic segmentation," in *ICCV*, 2021, pp. 7253–7262.
- [36] C. Liang, W. Wang, J. Miao, and Y. Yang, "Logic-induced diagnostic reasoning for semi-supervised semantic segmentation," in *ICCV*, 2023, pp. 16 197–16 208.
- [37] W. Wang, G. Sun, and L. Van Gool, "Looking beyond single images for weakly supervised semantic segmentation learning," *IEEE TPAMI*, vol. 46, no. 3, pp. 1635–1649, 2022.
- [38] L. Hoyer, D. J. Tan, M. F. Naeem, L. Van Gool, and F. Tombari, "Semivl: semi-supervised semantic segmentation with vision-language guidance," in *ECCV*. Springer, 2025, pp. 257–275.
- [39] H. Mai, R. Sun, T. Zhang, and F. Wu, "Rankmatch: Exploring the better consistency regularization for semi-supervised semantic segmentation," in *CVPR*, 2024, pp. 3391–3401.
- [40] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *BMVC*, 2018.
- [41] S. Liu, S. Zhi, E. Johns, and A. J. Davison, "Bootstrapping semantic segmentation with regional contrast," in *ICLR*, 2022.
- [42] Y. Zhou, H. Xu, W. Zhang, B.-B. Gao, and P.-A. Heng, "C3-semiseg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing," in *ICCV*, 2021, pp. 7016–7025.
- [43] Z. Tian, X. Zhang, P. Zhang, and K. Zhan, "Improving semi-supervised semantic segmentation with dual-level siamese structure network," in *MM*, 2023, pp. 4200–4208.
- [44] H. Su, V. Jampani, D. Sun, S. Maji, E. Kalogerakis, M.-H. Yang, and J. Kautz, "Splatnet: Sparse lattice networks for point cloud processing," in *CVPR*, 2018, pp. 2530–2539.
- [45] H.-Y. Chiang, Y.-L. Lin, Y.-C. Liu, and W. H. Hsu, "A unified point-based framework for 3d segmentation," in *3DV*, 2019, pp. 155–163.
- [46] M. Jaritz, T.-H. Vu, R. d. Charette, E. Wirbel, and P. Pérez, "xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation," in *CVPR*, 2020, pp. 12 605–12 614.
- [47] W. Hu, H. Zhao, L. Jiang, J. Jia, and T.-T. Wong, "Bidirectional projection network for cross dimension scene understanding," in *CVPR*, 2021, pp. 14 373–14 382.
- [48] Z. Wang, Y. Rao, X. Yu, J. Zhou, and J. Lu, "Semaffinet: Semantic-affine transformation for point cloud segmentation," in *CVPR*, 2022, pp. 11 809–11 819.
- [49] J. Li, H. Dai, H. Han, and Y. Ding, "Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving," in *CVPR*, 2023, pp. 21 694–21 704.
- [50] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *CVPR*, 2019, pp. 3075–3084.
- [51] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3d reconstructions of indoor scenes," in *CVPR*, 2017, pp. 2432–2443.
- [52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
- [53] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *CVPR*, 2021, pp. 2613–2622.
- [54] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICMLW*, 2013, pp. 896–901.
- [55] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.
- [56] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *NeurIPS*, 2017, pp. 5099–5108.
- [57] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE TPAMI*, vol. 19, pp. 263–272, 2018.
- [58] W. Wu, Z. Qi, and L. Fuxin, "Pointconv: Deep convolutional networks on 3d point clouds," in *CVPR*, 2019, pp. 9621–9630.
- [59] A. Valada, R. Mohan, and W. Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *IJCV*, vol. 128, pp. 1239–1285, 2018.
- [60] X. Yan, C. Zheng, Z. Li, S. Wang, and S. Cui, "Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling," in *CVPR*, 2020, pp. 5589–5598.
- [61] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.
- [62] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *ICCV*, 2019, pp. 6411–6420.
- [63] A. Valada, R. Mohan, and W. Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *IJCV*, vol. 128, pp. 1239–1285, 2018.
- [64] L. Deng, M. Yang, T. Li, Y. He, and C. Wang, "Rfbnet: Deep multimodal networks with residual fusion blocks for rgb-d semantic segmentation," *ArXiv*, vol. abs/1907.00135, 2019.
- [65] Y. Qian, L. Deng, T. Li, C. Wang, and M. Yang, "Gated-residual block for semantic segmentation using rgb-d data," *TITS*, vol. 23, pp. 11 836–11 844, 2022.

- [66] W. Shi, D. Zhu, G. Zhang, J. Xu, X. Wang, L. Chen, J. Li, and X. Zhang, "Multilevel cross-aware rgbd indoor semantic segmentation for bionic binocular robot," *IEEE TMRB*, vol. 2, pp. 382–390, 2020.
- [67] X. Lai, J. Liu, L. Jiang, L. Wang, H. Zhao, S. Liu, X. Qi, and J. Jia, "Stratified transformer for 3d point cloud segmentation," in *CVPR*, 2022, pp. 8490–8499.
- [68] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.
- [69] A. Hermans, G. Floros, and B. Leibe, "Dense 3d semantic mapping of indoor scenes from rgb-d images," in *ICRA*, 2014, pp. 2631–2638.
- [70] J. McCormac, A. Handa, A. J. Davison, and S. Leutenegger, "Semantic-fusion: Dense 3d semantic mapping with convolutional neural networks," in *ICRA*, 2017, pp. 4628–4635.
- [71] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.
- [72] P. Michel, O. Levy, and G. Neubig, "Are sixteen heads really better than one?" in *NeurIPS*, 2019.