Towards Deeper and Better Multi-VIew Feature Fusion for 3D Semantic Segmentation

Chaolong Yang, Yuyao Yan, Weiguang Zhao, Jianan Ye, Xi Yang, Amir Hussain, Bin Dong and Kaizhu Huang

January 9, 2024

Towards Deeper and Better Multi-view Feature Fusion for 3D Semantic Segmentation

Chaolong Yang[1], Yuyao Yan[2], Weiguang Zhao[1], Jianan Ye[2], Xi Yang[2], Amir Hussain[3], Bin Dong[4], and Kaizhu Huang[1](✉)

[1] Duke Kunshan University, Suzhou 215000, China
{chaolong.yang,weiguang.zhao,kaizhu.huang}@dukekunshan.edu.cn
[2] Xi'an Jiaotong-Liverpool University, Suzhou 215000, China
{yuyao.yan,xi.yang01}@xjtlu.edu.cn, Jianan.Ye20@student.xjtlu.edu.cn
[3] Edinburgh Napier University, Edinburgh, EH11 4BN, UK
A.Hussain@napier.ac.uk
[4] Ricoh Software Research Center (Beijing) Co., Ltd., Beijing 100000, China
Bin.Dong@srcb.ricoh.com

Abstract. 3D point clouds are rich in geometric structure information, while 2D images contain important and continuous texture information. Combining 2D information to achieve better 3D semantic segmentation has become a mainstream in 3D scene understanding. Albeit the success, it still remains elusive how to fuse and process the cross-dimensional features from these two distinct spaces. Existing state-of-the-art usually exploit bidirectional projection methods to align the cross-dimensional features and realize both 2D & 3D semantic segmentation tasks. However, to enable bidirectional mapping, this framework often requires a symmetrical 2D-3D network structure, thus limiting the network's flexibility. Meanwhile, such dual-task settings may distract the network easily and lead to over-fitting in the 3D segmentation task. As limited by the network's inflexibility, fused features can only pass through a decoder network, which affects model performance due to insufficient depth. To alleviate these drawbacks, in this paper, we argue that despite its simplicity, projecting unidirectionally multi-view 2D deep semantic features into the 3D space aligned with 3D deep semantic features could lead to better feature fusion. On the one hand, the unidirectional projection enforces our model focused more on the core task, i.e., 3D segmentation; on the other hand, unlocking the bidirectional to unidirectional projection enables a deeper cross-domain semantic alignment and enjoys the flexibility to fuse better and complicated features from very different spaces. In joint 2D-3D approaches, our proposed method achieves superior performance on the ScanNetv2 benchmark for 3D semantic segmentation.

Keywords: Point cloud · Semantic segmentation · Multi-view fusion.

1 Introduction

Semantic understanding of scenes is essential in numerous fields, including robot navigation, automatic driving systems, and medical diagnosis. While early researchers focused on 2D images to achieve scene understanding, fixed-view 2D images lack spatial structure information and suffer from object occlusion, limiting their use in spatially location-sensitive downstream tasks. In contrast, 3D point clouds offer a complete spatial structure without object occlusion. However, traditional 2D neural network-based methodologies cannot be used directly to deal with 3D data. To address this issue, point-based [21,22,29] and voxel-based [7,4,33] neural networks have been explored for 3D point cloud recognition and understanding. Nonetheless, 3D point clouds have low resolution and lack rich texture information. Thus, a promising solution to jointly understand complex scenes is to combine 2D images with detailed texture information and 3D point clouds with rich knowledge of geometric structures.

2D-3D fusion schemes for 3D semantic segmentation tasks can be categorized as bidirectional and unidirectional projection. A comparison of network frameworks between these two schemes is depicted in Fig. 1.

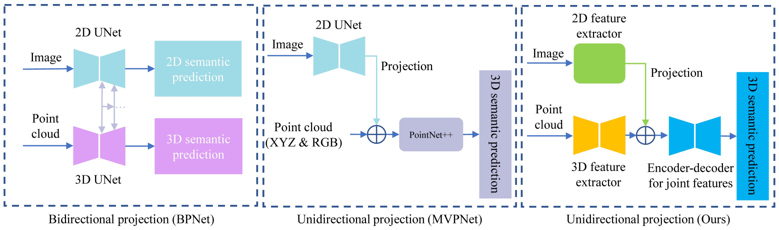

Fig. 1. Comparison of bidirectional & unidirectional projection

The pioneering BPNet [11] utilizes bidirectional projection to allow 2D and 3D features to flow between networks. Nevertheless, in order to mutually fuse information from 2D to 3D and 3D to 2D, it usually has to exploit a symmetrical decoder network. This makes its framework less flexible and could not take advantage of network depth, thus limiting its performance. Additionally, in complicated scenes, the 2D semantic component may distract from the core 3D task. To illustrate, we implement the idea of bidirectional projection on our proposed unidirectional projection framework. Namely, 3D features are also projected into 2D space combined with 2D features and then input to a complete 2D encoder-decoder network. On a large-scale complex indoor scene ScanNetv2 [5], we compare in Fig. 2 the 3D semantic loss on the validation set for unidirectional and bidirectional projection ideas during model training. Clearly, on the complicated scene of ScanNetv2, the bidirectional projection scheme causes distraction in the 3D task, where the loss goes up as the training continues. In comparison, the uni-projection implementation would lead to more stable performance with the focus mainly on the 3D task. Motivated by these findings, we argue that projecting unidirectionally multi-view 2D deep semantic features into the 3D space aligned with 3D deep semantic features can result in better feature fusion and more potential for downstream tasks like scene understanding.

Previous unidirectional projection methods [3,13] have been proposed in the literature to fuse 2D deep semantics and 3D shallow information (XYZ & RGB)

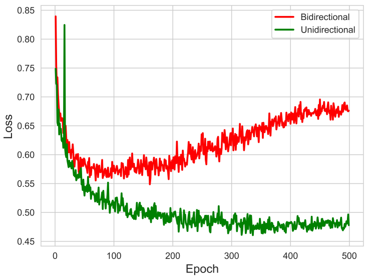

Fig. 2. Validation loss for unidirectional & bidirectional projection on validation set of the benchmark ScanNetv2 data

for 3D semantic segmentation tasks (see the middle graph in Fig. 1). Direct connection of deep and shallow semantics from different data domains could however result in the misalignment of the semantic space. To this end, we design a novel unidirectional projection framework, called the Deep Multi-view Fusion Network (DMF-Net), to more effectively fuse 2D & 3D features (see the right graph in Fig. 1). On the implementation front, we evaluate our model on the 3D semantic segmentation datasets: ScanNetv2 [5] and NYUv2 [25]. DMF-Net not only achieves top performance for joint 2D-3D methods on the ScanNetv2 benchmark, but also achieves state-of-the-art performance on the NYUv2 dataset. Our contributions can be summarized as follows.

- We argue that the unidirectional projection mechanism is not only more focused on 3D semantic understanding tasks than bidirectional projection but also facilitates deeper feature fusion. To this end, we design a method for uni-directional cross-domain semantic feature fusion to extract 2D & 3D deep features for alignment simultaneously.
- We propose a novel framework named Deep Multi-view Fusion Network (DMF-Net) for 3D scene semantic understanding. For the joint 2D-3D approaches, DMF-Net obtains top mIOU performance on the 3D Semantic Label Benchmark of ScanNetv2 [5], while it reaches the state-of-the-art on NYUv2 [25] datasets.
- We demonstrate the flexibility of DMF-Net, where all backbone modules in the network framework can be replaced. Specifically, 3D semantic segmentation performance will be stronger with a powerful backbone. Compared with common U-Net34 [23], the advanced Swin-UNet [2] shows a relative 4.4% improvement on mIOU in our framework.

2 Related Work

2.1 2D Semantic Segmentation

Image semantic segmentation has been significantly improved by the development of deep-learning [12] models. In the field of 2D semantic segmentation, Fully

Convolution Network (FCN) [18] is a landmark work although it has some limitations. Several Encoder-Decoder-based [23,30,26] structures combined multi-level information to fine segmentation. Besides, attention-based [2] models capable of extracting long-range contextual information were introduced into the image segmentation task. However, the lack of 3D spatial geometry information in 2D images hinders semantic comprehension of scenes.

2.2 3D Semantic Segmentation

To cope with the structuring problem of point cloud data, one popular research is to apply projection-based techniques [16,1,27]. However, multi-viewpoint projections are quite sensitive to the viewpoints chosen. Voxelized point clouds can be processed by 3D convolution in the same way as pixels in 2D neural networks. But, high-resolution voxels result in high memory and computational costs, while lower resolutions cause loss of detail. Consequently, 3D sparse convolutional networks [7,4] are designed to overcome these computational inefficiencies. Direct processing of point clouds [21,22] to achieve semantic segmentation has become a popular research topic in recent years. However, sparse 3D point clouds lack continuous texture information, resulting in limited recognition performance of 3D scenes.

2.3 3D Semantic Segmentation Based on Joint 2D-3D data

There has been some research in recent years on 2D and 3D data fusion, which can be broken down into unidirectional projection networks [6,3,13,15] and bidirectional projection networks [11]. The bidirectional projection network, typified by BPNet [11], focuses on both 2D and 3D semantic segmentation tasks. Due to the mutual flow of its 2D and 3D information, its network framework has to rely on a symmetrical decoding network. In addition to the inflexibility of its framework, the 2D task introduced by the bi-projection idea will distract the network from the 3D segmentation task. This is our motivation for choosing a framework based on the uni-projection idea.

In terms of view selection, 3DMV [6] and MVPNet [13] adopted a scheme with a fixed number of views, which means the views may not cover the entire 3D scene. In contrast, VMFusion [15] solves narrow viewing angle and occlusion issues by creating virtual viewpoints, but this approach has high computational costs that increase with the number of views. With an excellent balance, our work employs a dynamic view scheme that selects views based on the greedy algorithm of MVPNet until the view covers more than 90% of the 3D scene while keeping the scene uncut.

3 Methodology

3.1 Overview

An overview of our DMF-Net pipeline is illustrated in Fig. 3. Each scene data consist of a sequence of video frames and one point cloud scene. The input point cloud is called the original point cloud, while the point cloud formed by back-projecting all 2D feature maps into 3D space is called a back-projected point cloud with 2D features. DMF-Net consists of three U-shaped sub-networks, where the 2D feature extractor is 2D U-Net [23], and the 3D feature extractor is 3D MinkowskiUNet [4] (M.UNet). Moreover, the encoder-decoder for joint features is also the 3D M.UNet. Although we set a specific network backbone in our implementation, different network backbones can also be utilised in our DMF-Net. The view selection and back-projection modules will be elaborated in Sec. 3.2 and Sec. 3.3, respectively. As for the feature integration module, similar to MVPNet [13], it finds k nearest neighbour 2D features for each point of the original point cloud. Subsequently, it directly concatenates with the deep 3D semantic features of the original point cloud and input to 3D M.UNet for further learning to predict the semantic results of the entire scene.

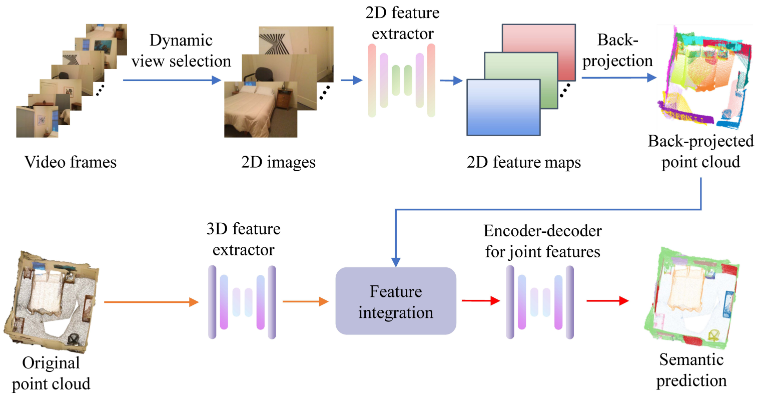

Fig. 3. Overview of the proposed DMF-Net

3.2 Dynamic View Selection

Previous work fixed the number of views, which would however cause insufficient overlaps between all the back-projected RGB-D frames and the scene point cloud. To make all the back-projected images cover as much of the scene point cloud as possible, many methods, e.g. MVPNet [13], generally choose to cut the point cloud scene. This will affect the recognition accuracy of cutting-edge objects. To this end, we propose a method for dynamic view selection, which sets a threshold for overlaps to ensure that the scene coverage is greater than 90% so that the number of views selected for each scene is different. The entire dynamic view selection algorithm is divided into two stages. The first stage is to construct the overlapping matrix. In the second stage, the view with the highest degree of overlap is dynamically selected according to the overlapping matrix cycle.

First, we define an overlapping matrix between the point cloud and video frames, as shown in Equation (1). This overlapping matrix indicates the relationship between each point and video frame. The first column lists the indices of the points, while the first row provides the indices of the frames. The entries in the matrix are either 0 or 1, representing non-overlapping and overlapping between points and frames respectively. Specifically, if a point of the original point cloud can find the back-projection point of the video frame within a certain range (1cm), it is determined that the point is an overlapping point. Considering that each point cloud contains hundreds of thousands of points, we randomly sample the original point cloud to reduce the computation.

$$\begin{array}{c} \\ Point\ 1 \\ Point\ 2 \\ Point\ 3 \\ \vdots \\ Point\ N \end{array} \begin{array}{c} \begin{array}{cccc} Frame\ 1 & Frame\ 2 & \cdots & Frame\ V \end{array} \\ \left(\begin{array}{cccc} 1 & 0 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & \cdots & 1 \end{array} \right) \end{array} \tag{1}$$

Second, we introduce the concept of scene overlap rate, which is the ratio between the number of overlap points corresponding to all selected video frames and the number of points in the down-sampled original point cloud. For each scene, the scene overlap rate is dynamically calculated after each video frame is selected. The overlapping matrix will be updated once one view is selected. If the scene overlap rate exceeds 90%, the selection of video frames is stopped to ensure excellent coverage of the scene while considering the number of views.

3.3 Unidirectional Projection Module

The video frames in the benchmark ScanNetv2 [5] dataset used in our experiments are captured by a fixed camera and reconstructed into a 3D scene point cloud. Therefore, we establish a mapping relationship between multi-view images and 3D point clouds based on depth maps, camera intrinsics, and poses. The world coordinate system is located where the point cloud scene is located, while the multi-view pictures belong to the pixel coordinate system. $(x_w, y_w, z_w)^T$ denotes a point in the world coordinate system and $(u, v)^T$ denotes a pixel point in the pixel coordinate system. Thus, the formula for converting the pixel coordinate system to the world coordinate system is shown as follows [32].

$$\begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} = Z_c K^{-1} \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}, \tag{2}$$

where Z_c is the depth value of the image, K is the camera internal parameter matrix, R is the orthogonal rotation matrix, and t is the translation vector.

To verify our unidirectional projection module, we back-project all the dynamically selected multi-view images into 3D space and put them together with the original point cloud. The visualization results are shown in Fig. 4. Each color in the back-projected point cloud represents the projected point set for each view. It is clear to see that all views dynamically selected basically cover the indoor objects in the whole point cloud scene.

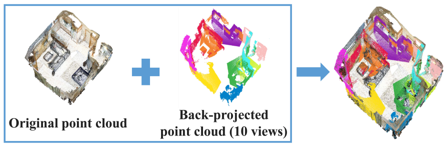

Fig. 4. Back-projection result

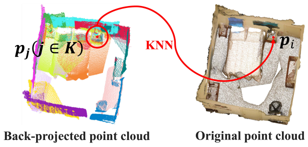

Fig. 5. Feature integration method

3.4 Feature Integration

The multi-view images are mapped to the space of the original point cloud using the unidirectional projection module in Sec. 3.3 to obtain the back-projected point cloud, each point of which contains 64-dimensional 2D deep semantic feature, denoted as f_j. Here, we define each point of the original point cloud as $p_i(x, y, z)$, as shown in Fig. 5. Specifically, each p_i utilizes the K-Nearest Neighbors (KNN) algorithm to find k back-projected points $p_j(j \in K)$ in the 3D Euclidean distance space. f_j is summed to obtain f_{2d} representing the 2D features of the p_i, while p_i has obtained the 64-dimensional 3D deep semantic feature f_{3d} through the 3D feature extractor. Finally, f_{2d} and f_{3d} are directly concatenated to become a 128-dimensional fusion feature F_i, which is calculated as follows.

$$F_i = \text{Concat}\left[f_{2d}, f_{3d}\right], \ f_{2d} = \sum_{j \in N_k(i)} f_j \tag{3}$$

4 Experiments

4.1 Datasets and Implementation Details

ScanNetv2 [5] is an indoor dataset including 706 different scenes, officially divided into 1201 training and 312 validation scans. Besides, the test set of 100 scans with hidden ground truth is used for benchmark. NYUv2 [25] contains 1449 densely labeled pairs of aligned RGB and depth images. We follow the official split of the dataset, using 795 for training and 654 for testing. Since this dataset has no 3D data, we need to use depth and camera intrinsics to generate 3D point clouds with 2D labels.

The training process can be divided into two stages. In the first stage, 2D images of ScanNetv2 were utilized to train a 2D feature extractor with 2D semantic labels. Noted that the original image resolution was downsampled to 320 × 240 for model acceleration and memory savings. In the second stage, a 3D network was trained with the frozen 2D feature extractor. The loss function used in the experiment is cross-entropy loss. As for the hyperparameter k in the feature integration module, we followed the previous practice, e.g. MVPNet and set it to 3. In the ablation study, the network structure of the proposed 3D feature extractor was set to M.UNet18A and the voxel size was set to 5 cm, which is consistent with the BPNet setup for a fair comparison. DMF-Net was trained for 500 epochs using the Adam [14] optimizer. The initial learning rate was set to 0.001, which decays with the cosine anneal schedule [19] at the 150th epochs. Besides, we conduct training on two RTX8000 cards with a mini-batch size 16.

4.2 Comparison with SoTAs on ScanNetv2 Benchmark

Quantitative results We compare our method with mainstream methods on the test set of ScanNetv2 to evaluate the 3D semantic segmentation performance of DMF-Net. The majority of these methods can be divided into point-based methods [22,10,28], convolution-based methods [31,17,29,4]), and 2D-3D fusion-based methods [6,13,11]. The results are reported in Table 1.

Table 1. Comparison with typical approaches on ScanNetv2 benchmark, including point-based, convolution-based and 2D-3D fusion-based (marked with *) methods

Methods	mIOU	bath	bed	bkshf	cab	chair	cntr	curt	desk	door	floor	other	pic	fridge	shower	sink	sofa	table	toilet	wall	window
P.Net++ [22]	33.9	58.4	47.8	45.8	25.6	36.0	25.0	24.7	27.8	26.1	67.7	18.3	11.7	21.2	14.5	36.4	34.6	23.2	54.8	52.3	25.2
3DMV* [6]	48.4	48.4	53.8	64.3	42.4	60.6	31.0	57.4	43.3	37.8	79.6	30.1	21.4	53.7	20.8	47.2	50.7	41.3	69.3	60.2	53.9
FAConv [31]	63.0	60.4	74.1	76.6	59.0	74.7	50.1	73.4	50.3	52.7	91.9	45.4	32.3	55.0	42.0	67.8	68.8	54.4	89.6	79.5	62.7
MCCNN [10]	63.3	86.6	73.1	77.1	57.6	80.9	41.0	68.4	49.7	49.1	94.9	46.6	10.5	58.1	64.6	62.0	68.0	54.2	81.7	79.5	61.8
FPConv [17]	63.9	78.5	76.0	71.3	60.3	79.8	39.2	53.4	60.3	52.4	94.8	45.7	25.0	53.8	72.3	59.8	69.6	61.4	87.2	79.9	56.7
MVPNet* [13]	64.1	83.1	71.5	67.1	59.0	78.1	39.4	67.9	64.2	55.3	93.7	46.2	25.6	64.9	40.6	62.6	69.1	66.6	87.7	79.2	60.8
DCM-Net [24]	65.8	77.8	70.2	80.6	61.9	81.3	46.8	69.3	49.4	52.4	94.1	44.9	29.8	51.0	82.1	67.5	72.7	56.8	82.6	80.3	63.7
KP-FCNN [28]	68.4	84.7	75.8	78.4	64.7	81.4	47.3	77.2	60.5	59.4	93.5	45.0	18.1	58.7	80.5	69.0	78.5	61.4	88.2	81.9	63.2
M.UNet [4]	73.6	85.9	81.8	**83.2**	70.9	**84.0**	52.1	85.3	66.0	64.3	95.1	54.4	28.6	73.1	**89.3**	67.5	77.2	68.3	87.4	85.2	72.7
BPNet* [11]	74.9	**90.9**	**81.8**	81.1	**75.2**	83.9	48.5	84.2	67.3	**64.4**	95.7	52.8	30.5	77.3	85.9	**78.8**	**81.8**	69.3	91.6	**85.6**	72.3
Ours*	**75.2**	90.6	79.3	80.2	68.9	82.5	**55.6**	**86.7**	**68.1**	60.2	**96.0**	**55.5**	**36.5**	**77.9**	85.9	74.7	79.5	**71.7**	**91.7**	**85.6**	**76.4**

DMF-Net achieves a significant mIOU performance improvement compared with point-based methods which are limited by their receptive field range and inefficient local information extraction. For convolution-based methods, such as stronger sparse convolution, M.UNet can expand the range of receptive fields. Our method outperforms M.UNet by a relative 2.2% on mIOU because 2D texture information was utilized. DMF-Net shows a relative improvement of 17.3% on mIOU compared to MVPNet, a baseline unidirectional projection scheme. Such improvement can be attributed to the fact that the feature alignment problem of MVPNet is alleviated. Especially, our unidirectional projection scheme DMF-Net is significantly better than the bidirectional projection method BPNet, one state-of-the-art in 2D-3D information fusion. The inflexibility of the BPNet framework limits its performance, while the high flexibility of our network framework enables further improvements.

Qualitative Results We compare the pure 3D sparse convolution M.UNet, the joint 2D-3D approach BPNet, and our method DMF-Net to conduct inference on the validation set of ScanNetv2. The visualization results are shown in Figure 6.

As indicated by the red boxes, the 3D-only method M.UNet does not discriminate well between smooth planes or objects with insignificant shape differences, such as windows, doors, pictures, and refrigerators. This may due to the low

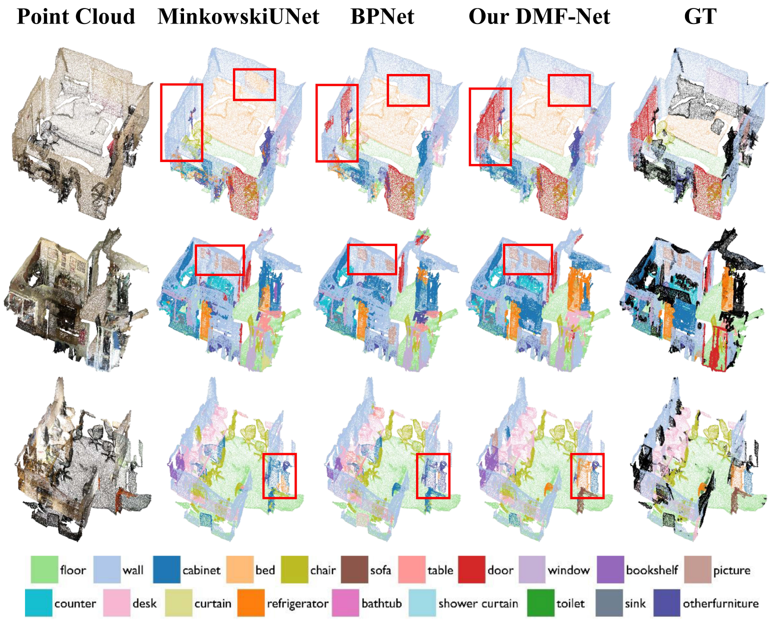

Fig. 6. Qualitative results of 3D semantic segmentation

resolution of the 3D point cloud and the lack of texture information for smooth planes. Despite the joint 2D-3D approach used by BPNet, the segmented objects are usually incomplete, owing that the bidirectional projection network distracts the core task, i.e. the 3D semantic segmentation.

4.3 Ablation Study and Analysis

Ablation for 2D-3D Fusion Effectiveness We first fuse 2D deep semantic features with 3D shallow semantic features (i.e. each point contains XYZ and RGB), followed by 3D sparse convolution. As shown in Table 2, the 3D semantic segmentation performance mIOU is improved from 66.4 to 70.8, indicating that 2D semantic features can benefit the 3D semantic segmentation task. The direct fusion of 2D deep semantic features with 3D shallow geometric features will cause misalignment in the semantic depth space affecting the network's performance. For this reason, our DMF-Net adds a 3D feature extractor based on the above framework so that 2D & 3D features are fused and aligned in semantic depth. As shown in Table 2, the feature-aligned model (V2) has a relative improvement of 1.3% on mIOU performance compared to the unaligned model (V1). In addition, we get a relative 4.4% on mIOU improvement with the stronger attention-based 2D backbone Swin-UNet [2] (V3) compared with the common U-Net34 model (V2). It is worth mentioning that the voxel size is sensitive to the performance of 3D sparse convolution. We adopt a deeper 3D sparse network, M.UNet34C, and set the voxel size to 2cm to obtain better results, as V4 shown in Table 2.

Table 2. 2D & 3D semantic segmentation on the validation set of ScanNetv2

Methods	Voxel Size	mIOU	
		2D	3D
U-Net34 [23]	-	60.7	-
Swin-UNet [2]	-	68.8	-
M.UNet18A [4]	5cm	-	66.4
Ours V1 (U-Net34 + XYZ & RGB)	5cm	-	69.9
Ours V2 (U-Net34 + M.UNet18A)	5cm	-	70.8
Ours V3 (Swin-UNet + M.UNet18A)	5cm	-	73.9
Ours V4 (Swin-UNet + M.UNet34C)	2cm	-	**75.6**

Table 3. 3D Semantic segmentation results on NYUv2 [25]

Methods	Mean Acc
SceneNet [8]	52.5
D3SM [9]	54.3
S.Fusion [20]	59.2
Scannet [5]	60.7
3DMV [6]	71.2
BPNet [11]	73.5
DMF-Net	**78.4**

Ablation for Projection Methods We conduct further ablative experiments to verify that the unidirectional projection scheme is more focused on the 3D semantic segmentation task than the bidirectional projection. Using the same framework in Figure 3, we project 3D features into the 2D deep semantic feature space. Essentially, we apply a projection method similar to Sec. 3.3, which is an opposite process. Meanwhile, the 2D-3D feature fusion is the same as in Sec. 3.4. After 3D features are fused with multi-view features, semantic labels are output through a U-Net34 Network. Hence 2D cross-entropy loss is introduced on the total loss of the model. To avoid focusing too much on the optimization of 2D tasks, we multiply the 2D loss by a weight of 0.1, the same with BPNet. At this time, the new 2D model parameters for training no longer freeze, and the learning rate of all 2D models is 10 times lower than that of the 3D model.

Our experiments show that the bidirectional projection model overfits when it is trained to the 200th epoch, as seen from the 3D validation loss in Figure 2. Meanwhile, the 3D mIOU of bidirectional projection only reaches 70.6, which is lower than the performance of the simple unidirectional projection (70.8). As the 2D task is also introduced, the increased learning parameters and the difficulty in adjusting the hyperparameters made it difficult for the model to focus more on 3D tasks. In this sense, unidirectional projection can focus more on 3D semantic segmentation tasks than bidirectional projection, leading to better flexibility.

4.4 DMF-Net on NYUv2

To verify the generalization ability, we conduct experiments on another popular RGB-D dataset, NYUv2 [25]. We report a dense pixel classification mean accuracy for DMF-Net, obtaining a significant performance improvement compared to other typical methods, especially joint 2D-3D methods, e.g. 3DMV [6] and BPNet [11]. As seen in Table 3, our DMF-Net gains a relative 6.6% performance improvement compared to the state-of-the-art BPNet [11]. This result demonstrates the strong generalization capability of the DMF-Net.

5 Conclusions and Future work

In our work, we propose a Deep Multi-view Fusion Network (DMF-Net) based on a unidirectional projection method to perform 3D semantic segmentation utilizing 2D continuous texture information and 3D geometry information. Compared with the previous 2D-3D fusion methods, DMF-Net enjoys a deeper and more flexible network. Thus DMF-Net enables improved segmentation accuracy for objects with little variation in shape, effectively compensating for the limitations of pure 3D methods. In addition, DMF-Net achieves the superior performance of the joint 2D-3D method in the ScanNetv2 benchmark. Moreover, we obtain significant performance gains over previous approaches on the NYUv2 dataset. Currently, the number of dynamically selected multi-view images in DMF-Net is relatively large in order to cover the full 3D scene. In the future, we will explore efficient view selection algorithms so that even a few image inputs could achieve the full coverage of the 3D scene.

Acknowledgements. The work was partially supported by the following: National Natural Science Foundation of China under no. 62376113; Jiangsu Science and Technology Programme (Natural Science Foundation of Jiangsu Province) under no. BE2020006-4, UK Engineering and Physical Sciences Research Council (EPSRC) Grants Ref. EP/M026981/1, EP/T021063/1, EP/T024917/.

References

1. Boulch, A., Le Saux, B., Audebert, N.: Unstructured point cloud semantic labeling using deep segmentation networks. In: 3DOR (2017)
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
3. Chiang, H.Y., Lin, Y.L., Liu, Y.C., Hsu, W.H.: A unified point-based framework for 3d segmentation. In: 3DV. pp. 155–163 (2019)
4. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: CVPR. pp. 3075–3084 (2019)
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR. pp. 5828–5839 (2017)
6. Dai, A., Nießner, M.: 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In: ECCV. pp. 452–468 (2018)
7. Graham, B., Engelcke, M., Van Der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: CVPR. pp. 9224–9232 (2018)
8. Handa, A., Patraucean, V., Badrinarayanan, V., Stent, S., Cipolla, R.: Scenenet: Understanding real world indoor scenes with synthetic data. arXiv preprint arXiv:1511.07041 (2015)
9. Hermans, A., Floros, G., Leibe, B.: Dense 3d semantic mapping of indoor scenes from rgb-d images. In: ICRA. pp. 2631–2638 (2014)
10. Hermosilla, P., Ritschel, T., Vázquez, P.P., Vinacua, À., Ropinski, T.: Monte carlo convolution for learning on non-uniformly sampled point clouds. ACM Transactions on Graphics (TOG) **37**(6), 1–12 (2018)

11. Hu, W., Zhao, H., Jiang, L., Jia, J., Wong, T.T.: Bidirectional projection network for cross dimension scene understanding. In: CVPR. pp. 14373–14382 (2021)
12. Huang, K., Hussain, A., Wang, Q., Zhang, R.: Deep learning: fundamentals, theory and applications, vol. 2. Springer (2019)
13. Jaritz, M., Gu, J., Su, H.: Multi-view pointnet for 3d scene understanding. In: ICCVW (2019)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
15. Kundu, A., Yin, X., Fathi, A., Ross, D., Brewington, B., Funkhouser, T., Pantofaru, C.: Virtual multi-view fusion for 3d semantic segmentation. In: ECCV. pp. 518–535 (2020)
16. Lawin, F.J., Danelljan, M., Tosteberg, P., Bhat, G., Khan, F.S., Felsberg, M.: Deep projective 3d semantic segmentation. In: CAIP. pp. 95–107 (2017)
17. Lin, Y., Yan, Z., Huang, H., Du, D., Liu, L., Cui, S., Han, X.: Fpconv: Learning local flattening for point convolution. In: CVPR. pp. 4293–4302 (2020)
18. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
19. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: ICLR (2017)
20. McCormac, J., Handa, A., Davison, A., Leutenegger, S.: Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In: ICRA. pp. 4628–4635 (2017)
21. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR. pp. 652–660 (2017)
22. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++ deep hierarchical feature learning on point sets in a metric space. In: NeurIPS. pp. 5105–5114 (2017)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
24. Schult, J., Engelmann, F., Kontogianni, T., Leibe, B.: Dualconvmesh-net: Joint geodesic and euclidean convolutions on 3d meshes. In: CVPR. pp. 8612–8622 (2020)
25. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV. pp. 746–760 (2012)
26. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR. pp. 5693–5703 (2019)
27. Tatarchenko, M., Park, J., Koltun, V., Zhou, Q.Y.: Tangent convolutions for dense prediction in 3d. In: CVPR. pp. 3887–3896 (2018)
28. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: ICCV. pp. 6411–6420 (2019)
29. Wu, W., Qi, Z., Fuxin, L.: Pointconv: Deep convolutional networks on 3d point clouds. In: CVPR. pp. 9621–9630 (2019)
30. Xiao, X., Lian, S., Luo, Z., Li, S.: Weighted res-unet for high-quality retina vessel segmentation. In: ITME. pp. 327–331 (2018)
31. Zhang, J., Zhu, C., Zheng, L., Xu, K.: Fusion-aware point convolution for online semantic 3d scene segmentation. In: CVPR. pp. 4534–4543 (2020)
32. Zhang, Z.: A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **22**(11), 1330–1334 (2000)
33. Zhao, W., Yan, Y., Yang, C., Ye, J., Yang, X., Huang, K.: Divide and conquer: 3d point cloud instance segmentation with point-wise binarization. In: ICCV (2023)