2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
September 27 - October 1, 2021. Prague, Czech Republic

2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-6654-1714-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/IROS51168.2021.9635888

Up-to-Down Network: Fusing Multi-Scale Context for 3D Semantic Scene Completion

Hao Zou[1], Xuemeng Yang[1], Tianxin Huang[1], Chujuan Zhang[1], Yong Liu[1,*],
Wanlong Li[2], Feng Wen[2], and Hongbo Zhang[2]

***Abstract*— An efficient 3D scene perception algorithm is a vital component for autonomous driving and robotics systems. In this paper, we focus on semantic scene completion, which is a task of jointly estimating the volumetric occupancy and semantic labels of objects. Since the real-world data is sparse and occluded, this is an extremely challenging task. We propose a novel framework, named Up-to-Down network (UDNet), to achieve the large-scale semantic scene completion with an encoder-decoder architecture for voxel grids. The novel up-to-down block can effectively aggregate multi-scale context information to improve labeling coherence, and the atrous spatial pyramid pooling module is leveraged to expand the receptive field while preserving detailed geometric information. Besides, the proposed multi-scale fusion mechanism efficiently aggregates global background information and improves the semantic completion accuracy. Moreover, to further satisfy the needs of different tasks, our UDNet can accomplish the multi-resolution semantic completion, achieving faster but coarser completion. Detailed experiments in the semantic scene completion benchmark of SemanticKITTI illustrate that our proposed framework surpasses the state-of-the-art methods with remarkable margins and a real-time inference speed by using only voxel grids as input.**

I. INTRODUCTION

The understanding of the large-scale outdoor scene is one of the fundamental functions of the autonomous driving system. The most commonly used sensors for autonomous driving are the LiDAR sensors that generate the sparse point cloud data. However, the sparseness and occlusion of the point cloud data make it difficult to realize the robust 3D perception tasks such as object detection [1]–[3] and tracking [4]–[6]. In order to make better use of the point cloud data in various downstream applications, it is vital to accomplish an effective 3D data understanding algorithm. In this paper, we focus on the large-scale semantic completion task for voxel grids, as illustrated in Fig. 1.

The indoor semantic scene completion methods have developed rapidly and achieved satisfactory performance in benchmark datasets such as SUNCG [7] and NYU [8]. Since the release of the SemanticKITTI completion dataset [9], researchers turn their attention to the outdoor semantic scene completion. Most existing large-scale outdoor semantic scene completion methods can be classified into two categories in terms of the input representation, the grid-based methods [7], [10], [11] and the point-based methods [12]–[14]. Given the voxel grids as input, the grid-based methods directly encode the occupancy grid and utilize a convolution neural network (CNN) to extract features for achieving lightweight but effective semantic completion. In this way, the inevitable information loss decreases the semantic completion accuracy. The point-based methods usually encode the point clouds as voxel grids and use a sparse convolution network [15] or CNN to extract features for semantic completion. Generally, the point-based methods can preserve the geometric structure of the 3D scene, but cannot fully explore the multi-scale context and have higher computation cost. Moreover, they need the additional data for training since the SemanticKITTI semantic scene completion benchmark only provides the voxel grids for training and inference.

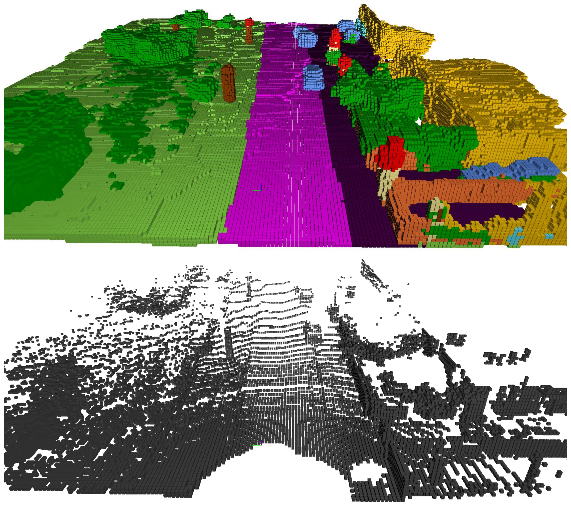


Fig. 1. An illustration of the semantic scene completion result from the SemanticKITTI validation set. The lower half is the input voxel grid and the upper half is the semantic completion result of our method.

In this work, we observe that even the same object categories have different physical sizes in the 3D scene, which has a significant impact on completion accuracy. We find that aggregating multi-scale context and using a large receptive field can effectively alleviate this problem, which is ignored by previous works. In other words, the local geometric details and the global background information of the scene play an essential role in large-scale semantic completion tasks. The local detailed information assists the network to reconstruct the fine-grained structure, and the multi-scale global background information assists the network to infer the occlusion and incomplete parts. Capturing the local

[1] Hao Zou, Xuemeng Yang, Tianxin Huang, Chujuan Zhang and Yong Liu are with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, 310027, China. (Yong Liu* is the corresponding author, email: yongliu@iipc.zju.edu.cn)

[2] Wanlong Li, Feng Wen and Hongbo Zhang are with the Huawei Noah's Ark lab.

and global features with a large receptive field can further improve the robustness of algorithms and the coherence of completion labels. This multi-scale context provides rich supervisions to learn discriminative 3D features for semantic completion but was seldom explored in previous works.

Motivated by the above concerns, we propose a novel semantic scene completion framework, named Up-to-Down network (UDNet), which is an encoder-decoder structure connected by the proposed up-to-down block. The main idea of our method is to fuse the local geometric details and global background information and further enlarge the receptive field without a loss of resolution. Specifically, given a 3D occupancy grid, our method leverages a 3D convolution based encoder to directly extract the voxel-wise feature without any preprocessing. For utilizing the multi-scale context information and eliminating the semantic gap of the features in different layers, the proposed up-to-down block effectively aggregates the features from the current, the previous, and the next layer of the encoder. We leverage the atrous spatial pyramid pooling (ASPP) module to capture the context of various receptive fields in the last layer of the encoder, which can exponentially enlarge the receptive field while preserving detailed geometric information. Additionally, for fully exploring multi-scale context features, we propose a multi-scale fusion mechanism to fuse the discriminative features from each layer of the decoder. Finally, due to the encoder-decoder structure, we accomplish a multi-resolution semantic completion task with multi-scale losses. Experiments on the SemanticKITTI validation set demonstrate that our UDNet can achieve the best performance on the multi-resolution semantic completion task(~10% improvement for scene completion and ~5% improvement for semantic completion over state-of-the-art) and outperform the baseline a large margin in overall performance.

The main contributions of this paper can be summarized as follows:

1 We propose a novel Up-to-Down network (UDNet) that efficiently takes advantage of multi-scale context features, leading to increased performance of 3D semantic scene completion.

2 We propose an efficient UD block to fully explore context information as well as strengthen the feature representation, and the ASPP module is utilized to dig the feature representation power of the 3D CNN with a large receptive field.

3 For satisfying the needs of various tasks, we achieve the semantic completion task at different resolutions.

4 Our proposed 3D semantic completion framework outperforms state-of-the-art methods with remarkable margins and ranks first among all publish works in terms of the scene completion task by using only occupancy grids as input.

II. Related Work

A. Semantic Scene Completion

Indoor Semantic Scene Completion: SSCNet [7] combines semantic segmentation and completion tasks for the first time, and proposes an end-to-end network to achieve indoor semantic scene completion. Similar to SSCNet [7], the subsequent indoor semantic scene completion network [16]–[19] converts the depth map into a truncated signed distance function (TSDF) [20] representation of voxels, and then uses a 3D CNN to extract features and generate completion and semantic labels. VVNet [17] effectively reduces the computational complexity by combining 2D view and 3D volume convolution and improves the accuracy of the semantic completion results. ForkNet [18] builds different geometric and semantic expressions based on a single encoder and three independent decoders in the same latent space, and realizes scene completion. CCPNet [21] improves the coherence of labels through a cascaded background pyramid. The proposed Guided Residual Refinement (GRR) module stores the fine-grained structure in the low-level features to solve the problem of the loss of target details and the ignoring of multi-scale spatial background information in previous methods. DDRNet [22] effectively reduces network parameters through a novel factorized convolution layer and simultaneously performs a multi-scale fusion of depth and image to improve the completion and segmentation performance. SGNN [23] proposes a new sparse generative neural network architecture, which uses a self-supervision mechanism to solve the problem of difficulty in obtaining ground truth in the completion task and predicts high-resolution geometry in a fully-convolution fashion.

Outdoor Semantic Scene Completion: For the first time, SemanticKITTI [9] introduces semantic scene completion in large-scale outdoor scenes and provides a benchmark dataset. Noted that the SemanticKITTI directly provides voxel grids rather than point clouds as the data representation for the semantic completion task. Most exist large-scale outdoor semantic scene completion can be classified into two categories in terms of the input representation, the grid-based methods [11] and the point-based methods [12]–[14]. LMSCNet [11] takes the voxel grid as input and utilizes a 2D UNet [24] with multi-scale skip connection to encode features on the spatial axis, and uses a 3D segmentation head to generate the completion result. Local-DIFs [12] proposes a continuous representation method based on local deep implicit functions to achieve scene completion. S3CNet [14] leverages a flipped TSDF computed from a spherical range image as a spatial encoding to differentiate the free, occupied, and occluded space of a scene and proposes a 2D sparse completion network based on bird's-eye-view to support the construction of 3D scene completion. JS3CNet [13] assists the semantic segmentation task to learn the background shape prior through the completion network and improves the segmentation performance. Meanwhile, it applies the cascade structure to use the segmentation result as the completion network input and proposes the Point-Voxel Interaction (PVI) module to fuse the segmentation and completion results to further improve the semantic completion performance. As a comparison, our UDNet directly takes the voxel grid as input, where each voxel is marked as empty or occupied, depending on whether or not it contains a laser measurement.

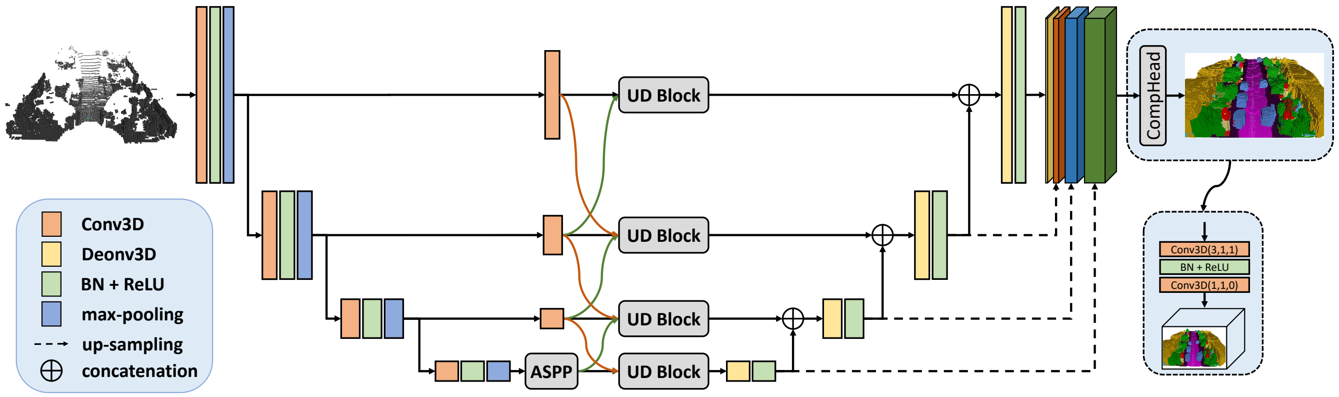


Fig. 2. The overall framework of our proposed UDNet, which follows an encoder-decoder structure. In the down-sampling stage, 3D convolution with batch normalization and ReLU are used to extract voxel-wise features, and max-pooling is leveraged to down-sample the features. In the up-sampling stage, 3D deconvolution with batch normalization and ReLU are utilized for refinement. Before skip connection operation, the UD block are proposed for fusing multi-scale context information and ASPP module are used to fuse the features of different receptive fields. Finally, a multi-scale fusion mechanism is used to fully explore multi-scale contexts and a simple completion head are proposed for predicting the final outputs.

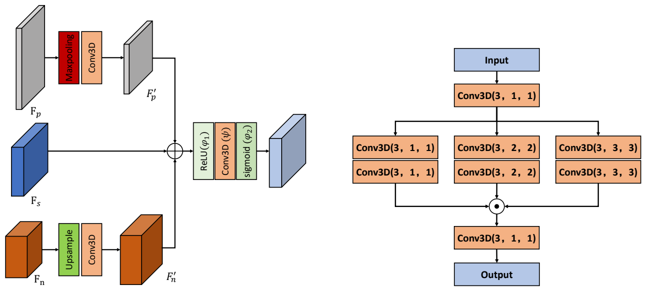
III. Method

A. Overview

In this paper, we propose the Up-to-Down network (UDNet), which is an end-to-end framework aiming at the semantic scene completion task for voxel grids. Given an occupancy grid as input, the goal of our UDNet is developing the 3D scene completion with a semantic label of each voxel. The semantic labels are noted as $\mathcal{L} = \{l_i\}_{i=0}^{N}$, where N represents the number of categories and l_0 is the empty voxel. We express the semantic completion result as a one-hot code with N+1 feature dimensions.

The critical observation is that the local geometric details and the global background information of the scene play an important role in the semantic completion task because of the fact that reconstructing the tiny objects needs more fine-grained information and the global context feature implies the latent completion for the occluded and incomplete parts. Those multi-scale contexts provide rich supervisions for learning the discriminative feature representations from voxel grids but were seldom explored in previous works.

Towards the above concerns, we propose an Up-to-Down network (UDNet) for effectively achieving 3D semantic scene completion. We propose a novel up-to-down block (UD block), which aggregates features of the current layer with low-level and high-level features to take advantage of local geometric details and multi-scale global background information for improving completion accuracy. Besides, to further enlarge the receptive field without a loss of resolution, we utilize the ASPP module to extract features in the last layer of the encoder. Furthermore, the features of different scales are aggregated to predict the semantic completion results in the decoder for fully exploring multi-scale contexts. In addition, we aim to achieve multi-resolution semantic completion for satisfying the need of different tasks, especially for robotics that needs coarser but faster scene completion. The detailed experiments demonstrate that our method surpasses the state-of-the-art with remarkable margins in terms of the scene completion task as well as the multi-resolution semantic completion task.



(a) UD block (b) ASPP module

Fig. 3. Illustrations of the UD block and ASPP module. The parameters of 3D convolution are kernel size, padding size and dilation size. The ⊕ denotes the concatenation and the ⊙ denotes the element-wise sum.

B. Network Architecture

As illustrated in Fig. 2, we design a UNet-like architecture for learning voxel-wise feature representations with the 3D convolution and 3D deconvolution. The whole framework can be divided into four stages: down-sampling stage, feature aggregation stage, up-sampling stage, and output stage. In the down-sampling stage, we utilize four 3D convolutions with kernel size 3^3, padding 1, and stride 2 to extract features and the generated four spatial feature dimensions are 16-32-64-128, respectively. Note that a batch normalization [25] and ReLU are appended after each convolution. We leverage max-pooling with kernel size 2^3 to down-sample the features and the spatial resolution of the features is down-sampled 16 times totally. In the up-sampling stage, there are four up-sample layers to gradually decrease feature dimensions as 128-64-32-16. Each layer is composed of deconvolution with kernel size 4^3, padding 1, and stride 2 and followed by a batch normalization [25] and ReLU. Since there are numerous physical sizes of different object categories in outdoor scenes, the network needs to capture different scale information. In order to make full use of the multi-scale context information in the encoder, we aggregate the feature from the same level of the encoder and the feature from the previous and next level of the encoder using the proposed UD block in the feature aggregation stage. Note that before

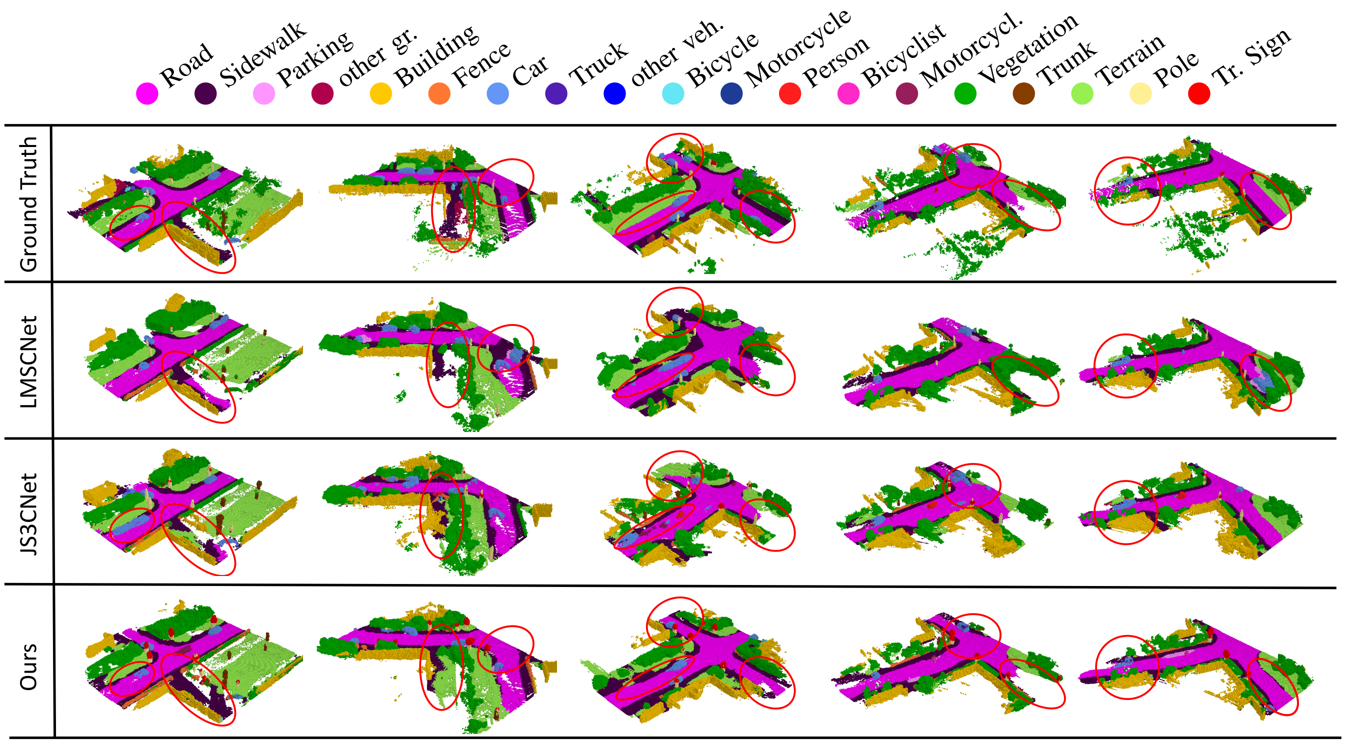


Fig. 4. Qualitative result of 3D semantic scene completion at full scale on the semanticKITTI validation set. Red circles show that our method preforms better in many details than state-of-the-art methods [11], [13]. It can be observed that our completion results contain more accurate and completer compared to the state-of-the-art methods.

UD block, we use a 1^3 convolution and ASPP module in the last layer to further expand the receptive field as well as learn the discriminative feature representation. Then, the aggregated features are simply concatenated with the feature from the previous layer of the decoder, since the voxel-wise features of decoder preserve more discriminative semantic features for 3D scene completion. In the output stage, to further utilize global background information, we present the multi-scale fusion mechanism to fuse discriminative features from different layers of the decoder and predict semantic completion results. Since the feature of each layer has different spatial resolutions, we leverage the trilinear interpolation to up-sample different features and concatenate them. Finally, an easy but efficient completion head is used to predict the final completion results.

UD block: The traditional UNet [24] structure directly fuses the feature from the current layer of the encoder and the previous layer of the decoder. However, the receptive fields of various layers are different, resulting in a large difference in learned semantic features. To cope with the semantic gap and fuse multi-scale global context information, we propose a novel up-to-down block (UD block) as shown in Fig. 3. Before skip connection operation, we aggregate the feature from the same layer of the encoder denoted as F_s with the feature from the previous and next layer of the encoder denoted as F_p and F_n, where $F_s \in \mathbb{R}^{C\times H\times W\times D}$, $F_p \in \mathbb{R}^{\frac{C}{2}\times 2H\times 2W\times 2D}$ and $F_s \in \mathbb{R}^{2C\times \frac{H}{2}\times \frac{W}{2}\times \frac{D}{2}}$. Specifically, F_p is down-sampled by max-pooling with kernel size 2^3 and convolution with kernel size 1^3, formulated as $F_p^{'} = \psi(down(F_p))$, where $\psi(x)$ represents a 3D convolution. F_n is up-sampled by trilinear interpolation and convolution with kernel size 1^3, formulated as $F_n^{'} = \psi(up(F_n))$. The final result can be computed as follow:

$$F_r = \varphi_2(\psi(\varphi_1([F_s, F_p^{'}, F_n^{'}]))), \qquad (1)$$

where $\varphi_2(x) = \frac{1}{1+exp(-x)}$ correspond to sigmoid activation function, $\varphi_2(x)$ is ReLU, and $F_r \in \mathbb{R}^{C\times H\times W\times D}$.

ASPP module: Since there are numerous physical sizes of different object categories in outdoor scenes, the network needs a large receptive field to capture different scale information. In order to dig the feature representation power of the 3D CNN with a large receptive field, we leverage the atrous spatial pyramid pooling (ASPP) module [26] to fuse the features of different receptive fields. Due to the dilation convolutions with different dilation rates, the network can enlarge the receptive field without a loss of resolution. As shown in Fig. 3, the input features are abstracted by a 1^3 convolution and three 3D dilation convolutions with kernel size $\{3^3, 3^3, 3^3\}$, padding $\{1, 2, 3\}$ and dilation rate $\{1, 2, 3\}$. The features extracted from different dilation rates are concatenated and abstracted by a 1^3 convolution to generate the final results.

Multi-resolution semantic completion: Follow [11], [23], we aim to develop a multi-resolution semantic completion task for meeting the needs of different tasks named UDNet-MR, as well as achieving coarser but faster semantic completion with a lower resolution, as shown in Fig. 5. Specifically, we directly append a completion head after the

TABLE I

Performance comparison with published methods on the SemanticKITTI test split by submitting to the official test server. Our method performs state-of-the-art performance in terms of scene completion. The bold value indicates the top performance.TD denotes TS3D + DarkNet53Seg. TDS denotes TS3D + DarkNet53Seg + SATNet. * results originate from [9].

Method	IoU	*Road* ●	*Sidewalk* ●	*Parking* ●	*other gr.* ●	*Building* ●	*Car* ●	*Truck* ●	*Bicycle* ●	*Motorcycle* ●	*other veh.* ●	*Vegetation* ●	*Trunk* ●	*Terrain* ●	*Person* ●	*Bicyclist* ●	*Motorcycl.* ●	*Fence* ●	*Pole* ●	*Tr. Sign* ●	mIoU	FPS
Point-based:																						
LDIFs [12]	57.7	**67.9**	**42.9**	**40.1**	11.4	40.4	**34.8**	4.4	3.6	2.4	4.8	42.2	26.5	39.1	2.5	1.1	0.0	29.0	21.3	17.5	22.7	-
JS3CNet [13]	56.6	64.7	39.9	34.9	**14.1**	39.4	33.3	**7.2**	14.4	8.8	12.7	**43.1**	19.6	**40.5**	8.0	5.1	0.4	30.4	18.9	15.9	23.8	1.7
S3CNet [14]	45.6	42.0	22.5	17.0	7.9	**52.2**	31.2	6.7	**41.5**	**45.0**	**16.1**	39.5	**34.0**	21.2	**45.9**	**35.8**	**16.0**	**31.3**	**31.0**	**24.3**	**29.5**	1.8
Grid-based:																						
*SSCNet [7]	29.8	27.6	17.0	15.6	6.0	20.9	10.4	1.8	0.0	0.0	0.1	25.8	11.9	18.2	0.0	0.0	0.0	14.4	7.9	3.7	9.5	**45.9**
*TS3D [10]	29.8	28.0	17.0	15.7	4.9	23.2	10.7	2.4	0.0	0.0	0.2	24.7	12.5	18.3	0.0	0.0	0.0	13.2	7.0	3.5	9.5	9.8
*TD [9]	25.0	27.5	18.5	18.9	6.6	22.1	8.0	2.2	0.0	0.0	3.9	19.5	12.9	20.2	**2.3**	**0.6**	0.0	17.8	7.6	7.0	10.2	8.7
*TDS [9]	50.6	62.2	31.6	23.3	6.5	34.1	30.7	4.85	0.0	0.0	0.0	40.1	21.9	**33.1**	0.0	0.0	0.0	24.1	16.9	6.9	17.7	1.3
LMSCNet [11]	55.3	**64.0**	33.1	24.9	3.2	38.7	29.5	2.5	0.0	0.0	0.1	40.5	19.0	30.8	0.0	0.0	0.0	20.5	15.7	0.5	17.0	21.3
UDNet(ours)	**59.4**	62.0	**35.1**	**28.2**	**9.1**	**39.5**	**33.9**	**3.8**	**0.8**	**0.4**	**4.4**	**40.9**	**23.2**	32.3	0.5	0.3	**0.3**	**24.4**	**18.8**	**13.1**	19.5	13.7

remaining 3 layers of the decoder, thus outputs at input relative scale of $\frac{1}{2^s}$, where $s \in \{1,2,3\}$. For obtaining multi-resolution semantic labels, we follow [27] and take the first value for every $2{\times}2{\times}2$, $4{\times}4{\times}4$ and $8{\times}8{\times}8$ non-overlapping area in the original label to get different resolution of labels.

C. Loss

Given the training data, our method can be trained using an end-to-end fashion. Specifically, for each scale s, the total loss $\mathcal{L}_{total}$ can be formulated as

$$\mathcal{L}_{total} = w_{full} {*} \mathcal{L}_{full} + w_s {*} \mathcal{L}_s, \quad (2)$$

where losses $\mathcal{L}_{full}$ for full scale completion are lovasz loss [28] and weighted cross-entry loss with same weight, losses $\mathcal{L}_s$ for $\frac{1}{2^s}$ resolution completion are weighted cross-entry loss and the w_{full} and w_s are the weights of each loss. We set $w_{full} = [\frac{5}{8}]^T$ and $w_s = [\frac{1}{8}, \frac{1}{8}, \frac{1}{8}]^T$ for UDNet-MR and $w_{full} = [1]^T$ and $w_s = [0,0,0]^T$ for UDNet. The weighted cross-entry loss can be formulated as:

$$\mathcal{L}_{wce} = -\sum_i \alpha_i P(y_i) log P(\hat{y}_i), \quad \alpha_i = 1/\sqrt{f_i}, \quad (3)$$

where f_i is the frequency of each category, and $P(y_i)$ and $P(\hat{y}_i)$ are the corresponding ground truth and predicts probability. The lovasz loss can be formulated as:

$$\mathcal{L}_{ls} = \frac{1}{|C|} \sum_{c \in C} J(\mathbf{e}(c)), \quad (4)$$

where J is the lovasz extension of IoU, $\mathbf{e}(c)$ is the vector of errors for class c.

IV. Experiment

In this section, we describe the experimental details of our proposed UDNet and compare it with the state-of-the-art methods on the scene semantic completion benchmark of SemanticKITTI [9] dataset, which provides 3D occupancy grids with semantic labels. Meanwhile, we conduct detailed ablation experiments to verify the effectiveness of our proposed modules. Finally, we visualize some qualitative results, as shown in Fig. 4.

A. Experiment Setup

Dataset: SemanticKITTI [9] is a large-scale LiDAR point cloud dataset, which utilizes Velodyne HDL-64 laser scanner to collect the data. This dataset is based on the KITTI odometry dataset [29] that contains 22 sequences. For the semantic scene completion task, SemanticKITTI provides 8728 voxelized LiDAR scans with ground truth labels. The size of each occupancy grid is 256 × 256 × 32 with a 0.2m voxel size. We use the public train/validation/test split protocol defined in the SemanticKITTI, and Sequence 0-7, 9-10 (3834 scans) for training, Sequence 8 (815 scans) for validation, Sequence 11-21 (3901 scans) for testing. To further expand the dataset, we adopt the x-y flipping augmentation for data diversification in the training set.

Implementation details: We train our model using Adam optimizer with a learning rate of 0.002 and a momentum of 0.9 for 60 epochs and use the exponential schedule with a decay rate of 0.98 each epoch. The method is trained using 4 NVIDIA GTX 1080Ti GPUs in parallel with a batch size of 4. Follow LMSCNet [11] we use mean IoU as the semantic completion metric and IoU, Precision, Recall as the scene completion metric. The mean IoU are defined as:

$$mIoU = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c + FN_c}, \quad (5)$$

where TP_c, FP_c, and FN_c denote the number of true positive, false positive, and false negative predictions for class c and C denotes the number of classes.

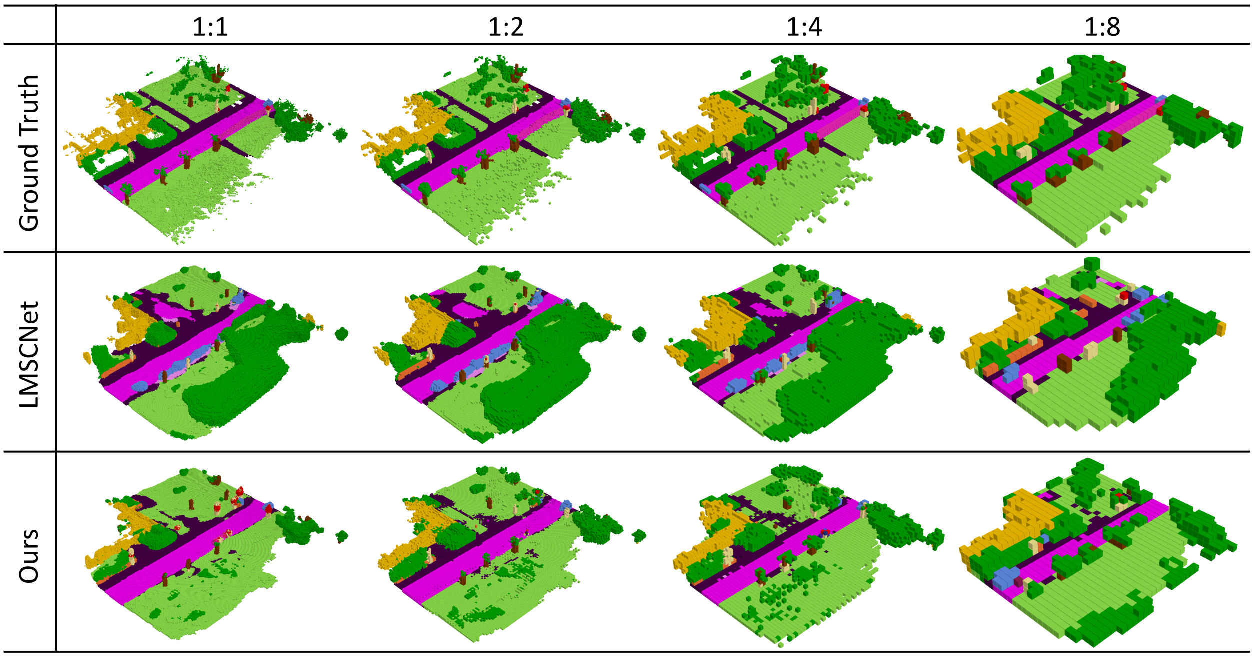


Fig. 5. Qualitative result of multi-resolution completion on the SemanticKITTI validation set. Compared to the results of LMSCNet, our method achieves more accurate semantic scene completion such as the $cars$ and $vegetation$.

B. Comparisons with the state-of-the-art

In this subsection, we present a quantitative evaluation of the proposed UDNet on outdoor large-scale semantic scene completion dataset: SemanticKITTI [9] for the 3D semantic completion task. It is worth noting that the SemanticKITTI benchmark only provides voxel grids for the semantic scene completion task, where each voxel is marked as empty or occupied, depending on whether or not it contains a laser measurement.

Quantitative Analysis: We list our results with all other published works in Table I and categorize them into two classes of point-based and grid-based. In each category, the best IoU per class is indicated as the bold value. As shown in Table I, our UDNet achieves the state-of-the-art performance in terms of IoU and reaches competitive performance on the semantic completion task in the SemanticKITTI benchmark. From the results, we can see that our proposed UDNet achieves **1.7%** improvement from the previous state-of-the-art [12] and **4.1%** improvement from the LMSCNet [11] with the same input data in terms of scene completion. As for the semantic completion task, our method achieves a more balanced performance and reaches a satisfactory result not only in big objects but also in tiny objects, such as the class of $vegetation$ and car. In addition, the point-based methods are usually difficult to achieve real-time completion, while our method can reach **13.7** FPS - ~8 × faster than JS3CNet [13] and S3CNet [14]. It is worth noting that our UDNet uses only the data from the SemanticKITTI provided, whereas the methods like point-based methods [12]–[14] use the raw point cloud data as the additional supplement. While the proposed method achieves high completion accuracy, it is also faster than most methods, making it applicable to real-time systems.

TABLE II

QUANTITATIVE RESULTS OF MULTI-SCALE SEMANTIC SCENE COMPLETION ON THE SEMANTICKITTI VALIDATION SET. THE BOLD VALUE INDICATES THE TOP PERFORMANCE.

Scale	Method	IoU	Precision	Recall	mIoU	FPS
1:1	LMSCNet	46.7	64.9	62.6	14.8	9.1
	UDNet-MR(ours)	**55.8**	**79.2**	**65.3**	**19.7**	**13.4**
1:2	LMSCNet	50.6	70.8	63.9	15.2	**51.5**
	UDNet-MR(ours)	**57.5**	**82.6**	**65.4**	**19.0**	28.3
1:4	LMSCNet	56.4	77.9	67.2	16.0	**153.7**
	UDNet-MR(ours)	**61.6**	**84.4**	**69.5**	**19.0**	35.4
1:8	LMSCNet	64.8	84.7	73.4	16.0	**230.8**
	UDNet-MR(ours)	**68.2**	**86.2**	**76.4**	**18.4**	41.6

Qualitative Analysis: As shown in Fig. 4, we visualize the prediction results of our method and the state-of-the-art open-source methods [11], [13] on the SemanticKITTI [9] validation set and ground truth is also provided as a reference. We utilize the pre-trained models and inference code published by LMSCNet[1] and JS3CNet[2], and reproduce the results, which represent the grid-based and point-based methods, respectively. As can be seen, the scene completion results of our method are more abundant in details and the semantic completion results of our method are less error-prone.

As for the scene completion task, it can be easily seen that our method performs better in the consistency of completion results and completeness of details. For example, JS3CNet will cause some redundant completion results of the car in the first column, and LMSCNet and JS3CNet will cause some insufficient completion results of the $road$ in the third column. As for the semantic completion task, it can be easily seen that our method is less error-prone. For example,

[1] https://github.com/cv-rits/LMSCNet

[2] https://github.com/yanx27/JS3C-Net

TABLE III

QUANTITATIVE RESULTS OF THE BASELINE AND ABLATIONS ON THE SEMANTICKITTI VALIDATION SET. MS DENOTES THE MULTI-SCALE FUSION MECHANISM. UNDERLINE MARKS RESULTS THAT 5% HIGHER THAN THE BASELINE. THE BOLD VALUE INDICATES THE TOP PERFORMANCE. THE W/O SEM. DENOTES WITHOUT SEMANTIC SUPERVISION SIGNAL.

Method	*IoU*	*Precision*	*Recall*	*mIoU*	*Car* ●	*Bicycle* ●	*Motorcycle* ●	*Truck* ●	*other veh.* ●	*Person* ●	*Bicyclist* ●	*Motorcycl.* ●	*Road* ●	*Parking* ●	*Sidewalk* ●	*other gr.* ●	*Building* ●	*Fence* ●	*Vegetation* ●	*Trunk* ●	*Terrain* ●	*Pole* ●	*Tr. Sign* ●
UNet(baseline)	55.4	83.6	62.2	18.0	36.4	0.0	0.0	10.2	6.9	0.0	0.0	0.0	65.6	14.9	36.8	0.0	34.4	11.5	39.7	12.4	44.9	21.4	6.2
UD block	55.5	84.2	61.9	19.2	39.8	0.0	0.0	21.3	8.6	0.2	0.2	0.0	65.4	15.6	36.2	0.7	34.3	11.7	39.7	15.6	44.2	22.2	**8.4**
UD block+MS	56.5	78.7	66.7	19.5	42.2	1.5	**4.1**	17.9	10.4	**2.9**	1.0	0.0	66.1	18.2	36.7	0.0	36.1	11.4	39.6	**19.3**	44.7	15.3	3.1
UD block+ASPP	56.7	75.9	69.8	19.9	41.6	1.0	2.8	25.5	**14.9**	2.4	0.0	0.0	66.8	18.3	35.7	1.1	**36.7**	10.6	39.4	17.4	45.0	15.7	2.9
UDNet	58.9	78.5	70.9	**20.7**	**42.1**	**1.8**	2.3	**25.7**	11.2	2.5	**1.2**	0.0	**67.0**	**20.3**	**37.2**	**2.2**	36.0	**11.9**	**40.1**	18.3	**45.8**	**23.0**	3.8
UDNet w/o sem.	**59.3**	78.5	71.3	(59.3)	59.3																		

LMSCNet will cause wrong classifications for the *road*, *sidewalk*, and *terrain* in the second and fourth columns. Since our method fuses the multi-scale context information, those problems are rarely happening.

C. Multi-resolution semantic completion

In this subsection, we provide the quantitative evaluation of the UDNet-MR on the SemanticKITTI validation set. We utilize the pre-trained model and inference code published by LMSCNet, and reproduce the results. Note that the results in LMSCNet are reported by using the SemanticKITTI completion dataset V1.0, while the official updates the dataset to V1.1, which causes the performance decrease of LMSCNet.

Quantitative Analysis: Table II presents the multi-resolution semantic scene completion results on the SemanticKITTI validation set with comparison to the state-of-the-art method [11]. From the table, we can see that our method outperforms the previous state-of-the-art by a significant margin in overall performance, that are nearly **10%** gains in scene completion and **5%** gains in semantic completion at full scale completion. Since LMSCNet appends four time-consuming segmentation heads for the multi-resolution completion, the running speed drops sharply with the increase of resolution. Even we leverage the 3D convolution instead of the 2D convolution, our method runs faster than LMSCNet at the full-resolution completion task with nearly 1.5 times the margin. In addition, the performance of the semantic completion decreases as the scene resolution decreases. The main reason is that the geometric details and semantic information cannot be preserved at the low-resolution scene.

Qualitative Analysis: Fig. 5 shows the multi-resolution semantic completion qualitative results of our method and LMSCNet on the SemanticKITTI validation set and the ground truth is also provided as a reference. It can be easily seen that our method performs better for objects such as the *cars*, *vegetation*, and *road*. Part of the reason that our network can better capture multi-scale context information with a large receptive field. As for the scene completion task, we can see that our method can achieve smoother and more complete prediction results. For example, LMSCNet will cause some redundant prediction results for the objects such as the *vegetation* and *car*. As for the semantic completion task, our method achieves more accurate label classifications than LMSCNet, such as the *trunk* and *terrain*.

D. Ablation study

In this subsection, we conduct detailed ablation studies to analyze the effect of the proposed modules. All models are trained on the train set and evaluated on the validation set. A series of quantitative results are shown in Table III.

Effect of the UD block: For fully exploring the context information and eliminating semantic ambiguity in skip connection operations, the proposed UD block can effectively improve the semantic completion task compared to the baseline, as shown in the first and second rows. It can be easily seen that the baseline with the UD block achieves 1.2% improvement from the baseline in terms of mIoU.

Effects of the multi-scale fusion mechanism: To validate the contribution of the multi-scale fusion mechanism to the results, we train baseline with the UD block and multi-scale fusion mechanism, as shown in the third row. Compared to the baseline, it increases 1.1% and 1.5% for scene completion and semantic completion, respectively, which demonstrates that the multi-scale information will effectively help the UDNet to learn the fine-grained and global context features.

Effects of the ASPP module: The effects of the ASPP module have been verified in the 2D semantic segmentation task [26]. We utilize the 3D ASPP module for fusing the information of different receptive fields and train baseline with the UD block and ASPP module, as shown in the fourth row. Compared to the baseline, it increases 1.3% and 1.9% for the scene completion and semantic completion, which demonstrates that the ASPP module enables better capture the richer semantic information of different scales.

Impact of without semantic supervision signal: For exploring the impact of the semantic information for scene completion tasks, we train our method with and without semantic supervision signal, as shown in the fifth and sixth rows. Without the semantic supervision signal, the performance of completion only increases 0.4% compared to jointly training segmentation and completion tasks. The performance of scene completion is unaffected by semantic supervision signal. This implies that semantic completion and scene completion are actually related.

V. CONCLUSIONS

In this paper, we have presented the UDNet for large-scale semantic scene completion from voxel grids. The proposed UD block fully explores context information as well as strengthens the feature representation, and the ASPP module is utilized to dig the feature representation power of the 3D CNN with a large receptive field. The multi-scale fusion mechanism has been proposed to aggregate the multi-scale context feature. Furthermore, we have proposed the multi-resolution completion for satisfying different needs. Extensive experiments demonstrate that our UDNet significantly improves the outdoor large-scale scene completion accuracy with real-time inference speed on the SemanticKITTI test set. In future work, we will explore the relationship between indoor and outdoor semantic scene completion and expand our UDNet to indoor semantic scene completion.

REFERENCES

[1] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[2] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 770–779.

[3] C. He, H. Zeng, J. Huang, X.-S. Hua, and L. Zhang, "Structure aware single-stage 3d object detection from point cloud," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 873–11 882.

[4] S. Giancola, J. Zarzar, and B. Ghanem, "Leveraging shape completion for 3d siamese tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1359–1368.

[5] H. Qi, C. Feng, Z. Cao, F. Zhao, and Y. Xiao, "P2b: Point-to-box network for 3d object tracking in point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6329–6338.

[6] H. Zou, J. Cui, X. Kong, C. Zhang, Y. Liu, F. Wen, and W. Li, "F-siamese tracker: A frustum-based double siamese network for 3d single object tracking," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 8133–8139.

[7] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1746–1754.

[8] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[9] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9297–9307.

[10] M. Garbade, Y.-T. Chen, J. Sawatzky, and J. Gall, "Two stream 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[11] L. Roldão, R. de Charette, and A. Verroust-Blondet, "Lmscnet: Lightweight multiscale 3d semantic completion," *arXiv preprint arXiv:2008.10559*, 2020.

[12] C. B. Rist, D. Emmerichs, M. Enzweiler, and D. M. Gavrila, "Semantic scene completion using local deep implicit functions on lidar data," *arXiv preprint arXiv:2011.09141*, 2020.

[13] X. Yan, J. Gao, J. Li, R. Zhang, Z. Li, R. Huang, and S. Cui, "Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion," *arXiv preprint arXiv:2012.03762*, 2020.

[14] R. Cheng, C. Agia, Y. Ren, X. Li, and L. Bingbing, "S3cnet: A sparse semantic scene completion network for lidar point clouds," *arXiv preprint arXiv:2012.09242*, 2020.

[15] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3075–3084.

[16] S. Liu, Y. Hu, Y. Zeng, Q. Tang, B. Jin, Y. Han, and X. Li, "See and think: Disentangling semantic scene completion," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 261–272.

[17] Y.-X. Guo and X. Tong, "View-volume network for semantic scene completion from a single depth image," *arXiv preprint arXiv:1806.05361*, 2018.

[18] Y. Wang, D. J. Tan, N. Navab, and F. Tombari, "Forknet: Multi-branch volumetric semantic completion from a single depth image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8608–8617.

[19] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5828–5839.

[20] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison *et al.*, "Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011, pp. 559–568.

[21] P. Zhang, W. Liu, Y. Lei, H. Lu, and X. Yang, "Cascaded context pyramid for full-resolution 3d semantic scene completion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7801–7810.

[22] J. Li, Y. Liu, D. Gong, Q. Shi, X. Yuan, C. Zhao, and I. Reid, "Rgbd based dimensional decomposition residual network for 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7693–7702.

[23] A. Dai, C. Diller, and M. Nießner, "Sg-nn: Sparse generative neural networks for self-supervised scene completion of rgb-d scans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 849–858.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[27] W. Liu, Y. Sun, and Q. Ji, "Mdan-unet: Multi-scale and dual attention enhanced nested u-net architecture for segmentation of optical coherence tomography images," *Algorithms*, vol. 13, no. 3, p. 60, 2020.

[28] M. Berman, A. R. Triki, and M. B. Blaschko, "The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4413–4421.

[29] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.