SCPNet: Semantic Scene Completion on Point Cloud

Zhaoyang Xia[1], **Youquan Liu**[1], **Xin Li**[2], **Xinge Zhu**[3], **Yuexin Ma**[4],
Yikang Li[1], **Yuenan Hou**[1]†, **and Yu Qiao**[1]
[1]Shanghai AI Laboratory [2]East China Normal University
[3]The Chinese University of Hong Kong [4]ShanghaiTech University
[1]{houyuenan, liyikang, qiaoyu}@pjlab.org.cn

Abstract

Training deep models for semantic scene completion (SSC) is challenging due to the sparse and incomplete input, a large quantity of objects of diverse scales as well as the inherent label noise for moving objects. To address the above-mentioned problems, we propose the following three solutions: 1) ***Redesigning the completion sub-network.*** *We design a novel completion sub-network, which consists of several Multi-Path Blocks (MPBs) to aggregate multi-scale features and is free from the lossy downsampling operations. 2)* ***Distilling rich knowledge from the multi-frame model.*** *We design a novel knowledge distillation objective, dubbed Dense-to-Sparse Knowledge Distillation (DSKD). It transfers the dense, relation-based semantic knowledge from the multi-frame teacher to the single-frame student, significantly improving the representation learning of the single-frame model. 3)* ***Completion label rectification.*** *We propose a simple yet effective label rectification strategy, which uses off-the-shelf panoptic segmentation labels to remove the traces of dynamic objects in completion labels, greatly improving the performance of deep models especially for those moving objects. Extensive experiments are conducted in two public SSC benchmarks,* i.e.*, SemanticKITTI and SemanticPOSS. Our SCPNet ranks* ***1****st on SemanticKITTI semantic scene completion challenge and surpasses the competitive S3CNet [3] by* ***7.2*** *mIoU. SCPNet also outperforms previous completion algorithms on the SemanticPOSS dataset. Besides, our method also achieves competitive results on SemanticKITTI semantic segmentation tasks, showing that knowledge learned in the scene completion is beneficial to the segmentation task.*

1. Introduction

Semantic scene completion (SSC) [20] aims at inferring both geometry and semantics of the scene from an incomplete and sparse observation, which is a crucial component in 3D scene understanding. Performing semantic scene completion in the outdoor scenarios is challenging due to the sparse and incomplete input, a large quantity of objects of diverse scales as well as the inherent label noise for those moving objects (See Fig. 1 (a)).

Recent years have witnessed an explosion of methods in the outdoor scene completion field [3, 19, 22, 25, 29, 32]. For example, S3CNet [3] performs 2D and 3D completion tasks jointly and achieves impressive performance on SemanticKITTI [1]. JS3C-Net [29] performs semantic segmentation first and then feeds the segmentation features to the completion sub-network. The coarse-to-fine refinement module is further put forward to improve the completion quality. Although significant progress has been achieved in this area, these methods heavily rely on the voxelwise completion labels and show unsatisfactory completion performance on the small, distant objects and crowded scenes. Moreover, the long traces of dynamic objects in original completion labels will hamper the learning of completion models, which is overlooked in the previous literature [20].

To address the preceding problems, we propose three solutions from the aspects of the completion sub-network redesign, distillation of multi-frame knowledge as well as completion label rectification. Specifically, we first make a comprehensive overhaul of the completion sub-network. We adopt the completion-first principle and make the completion module directly process the raw voxel features. Besides, we avoid the use of downsampling operations since they inevitably introduce information loss and cause severe misclassification for those small objects and crowded scenes. To improve the completion quality on objects of diverse scales, we design Multi-Path Blocks (MPBs) with varied kernel sizes, which aggregate multi-scale features and fully utilize the rich contextual information.

Second, to combat against the sparse and incomplete input signals, we make the single-scan student model distil knowledge from the multi-frame teacher model. However,

†: Corresponding author.

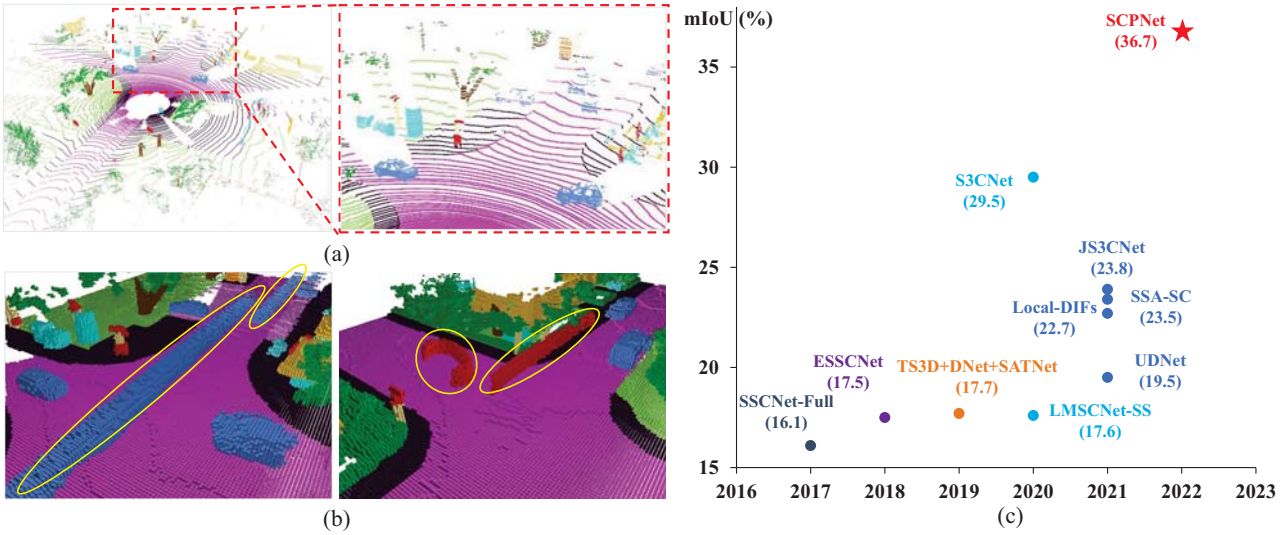

Figure 1. Left: challenges of semantic scene completion, *i.e.*, (a) sparse and incomplete input, varying completion difficulty for objects of diverse scales and (b) long traces of dynamic objects in completion labels. Traces of cars and persons are highlighted by yellow ellipses. Right: (c) performance comparison of various completion algorithms on SemanticKITTI semantic scene completion challenge.

mimicking the probabilistic knowledge of each point/voxel brings marginal gains. Instead, we propose to distill the pairwise similarity information. Considering the sparsity and unorderness of features, we align the features using their indices and then force the consistency between the pairwise similarity maps of student features and those of teacher features, make the student benefit from the relational knowledge of the teacher. The resulting Dense-to-Sparse Knowledge Distillation objective is termed DSKD, which is specifically designed for the scene completion task.

Finally, to address the long traces of those dynamic objects in the completion labels, we propose a simple yet effective label rectification strategy. The core idea is to use off-the-shelf panoptic segmentation labels to remove the traces of dynamic objects in completion labels. The rectified completion labels are more accurate and reliable, greatly improving the completion qualities of deep models on those moving objects.

We conduct experiments on two large-scale outdoor scene completion benchmarks, *i.e.*, SemanticKITTI [1] and SemanticPOSS [16]. Our SCPNet ranks **1**st on SemanticKITTI semantic scene completion challenge [1] and outperforms the S3CNet [3] by **7.2** mIoU. SCPNet also achieves better performance than other completion algorithms on the SemanticPOSS dataset. The learned knowledge from the completion task also benefits the segmentation task, making our SCPNet achieve superior performance on the SemanticKITTI semantic segmentation task.

Our contributions are summarized as follows.

- The comprehensive redesign of the completion sub-network. We unveil several key factors to building strong completion networks.
- To cope with the sparsity and incompleteness of the input, we propose to distill the dense relation-based knowledge from the multi-frame model. Note that we are the **first** to apply knowledge distillation to the semantic scene completion task.
- To address the long traces of moving objects in completion labels, we present the completion label rectification strategy.
- Our SCPNet ranks 1st on SemanticKITTI semantic scene completion challenge, outperforming the previous SOTA S3CNet [3] by **7.2** mIoU. Competitive performance is also shown in SemanticPOSS completion task and SemanticKITTI semantic segmentation task.

2. Related Work

Semantic scene completion. Early works on scene completion mainly concentrates on the indoor scenarios [5, 6, 15, 22]. The point cloud in indoor scenarios is dense, small-scale and has uniform density. By contrast, point cloud in outdoor scenes is sparse, large-scale and has varying density, which poses great challenges to the semantic scene completion algorithms [3, 20, 25, 29, 32]. Various algorithms have been proposed, for instance, LMSCNet [19] uses a mixture of 2D and 3D convolutions to build the efficient completion backbone. S3CNet [3] performs 2D and 3D scene completion simultaneously, and fuses the results by the proposed dynamic voxel fusion module. JS3C-Net [29] attaches the completion network to the segmentation backbone and refines the completion outputs via the point-voxel

[1] https://codalab.lisn.upsaclay.fr/competitions/7170#results till 2022-11-12 00:00 Pacific Time, and our method is termed SCPNet.

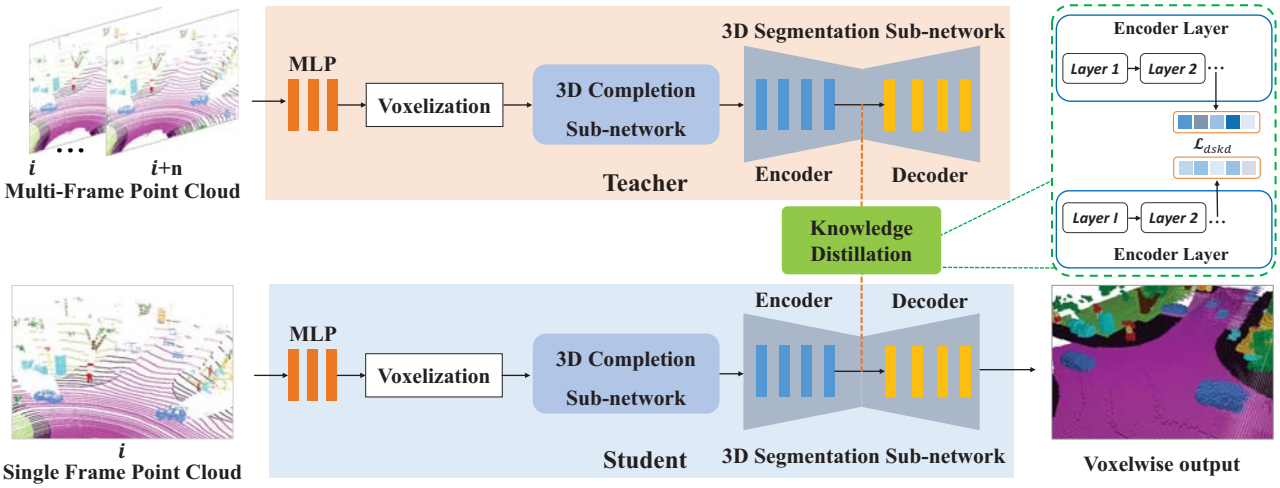

Figure 2. Framework overview of SCPNet. There are two SCPNets, one is teacher and the other is student, and they have the same architecture. The teacher takes multi-frame point cloud as input while the student takes the single-frame point cloud as input. The dense-to-sparse knowledge distillation loss is proposed to transfer the dense semantic knowledge from teacher to student. In each SCPNet, there are two sub-networks, *i.e.*, the completion sub-network and the segmentation sub-network. The point cloud is first processed by a stack of MLPs to extract point features. These point features are voxelized and then fed to the completion sub-network to produce denser voxel features. The produced voxel features are further fed to the segmentation sub-network to generate the ultimate voxelwise output. For the completion sub-network, it is comprised of several multi-path blocks, free from the lossy downsampling operations. For the segmentation sub-network, it is adapted from the Cylinder3D network.

interaction module. Compared with previous completion networks, our SCPNet is free from the lossy downsampling operations. Besides, our network is built on several MPBs that aggregate multi-scale features and can achieve high completion quality in objects of various sizes.

Knowledge distillation. Knowledge distillation (KD) originates from the pioneering work of G. Hinton *et al*. [8]. Its primary objective is to transfer the dark knowledge from the large over-parameterized teacher model to the small compact student model. A large number of methods have been proposed and various forms of knowledge [21, 24, 27, 33] have been designed, *e.g*., intermediate features [11, 21], visual attention maps [10, 33], region-level affinity scores [9], similarity scores of different samples [24, 26], *etc*. It is noteworthy that the majority of the distillation methods concentrate on the 2D tasks. Only a few distillation methods have focused on 3D domains, *e.g*., PVKD [12] and SparseKD [31]. To our knowledge, this is the first work that applies knowledge distillation to the semantic scene completion task. We propose to transfer the dense, relation-based semantic knowledge from the multi-frame model to the single-frame one.

3. Methodology

The objective of the semantic scene completion task is to infer the complete geometric and semantic layout given the incomplete and sparse input. Formally, given the input point cloud $\mathbf{X} \in \mathbb{R}^{N\times 3}$, the network needs to assign a label to each voxel of the $L\times W\times H$ voxel space to indicate whether it is empty or belongs to a specific class $c \in \{0, 1, 2, ..., C-1\}$, where C is the number of classes, L, W and H are the length, width and height of the voxel space, respectively. We denote the voxelwise output as $\mathbf{O} \in \mathbb{R}^{L\times W\times H\times C}$.

3.1. Framework overview

As shown in Fig. 2, our SCPNet is comprised of two sub-networks, *i.e*., the completion sub-network and the segmentation sub-network. The completion sub-network is designed based on several key design principles. The segmentation sub-network is built upon Cylinder3D [35, 36], with some minor modifications. In the following sections, we first detail the completion sub-network and introduce several design principles that are vital to building a strong completion sub-network. The knowledge distillation of multi-frame model and the completion label rectification will be explained thereafter.

3.2. Redesigning the Completion Sub-network

Take JS3C-Net [29] as example. The original JS3C-Net first performs semantic segmentation and then conducts completion upon the segmentation features. Although this pipeline can benefit from the segmentation outputs, there are several drawbacks in this framework. First, the parameters in the completion sub-network are much fewer than the segmentation sub-network, thus yielding unsatisfactory completion performance. Second, there are downsampling and upsampling blocks in the completion sub-network. The

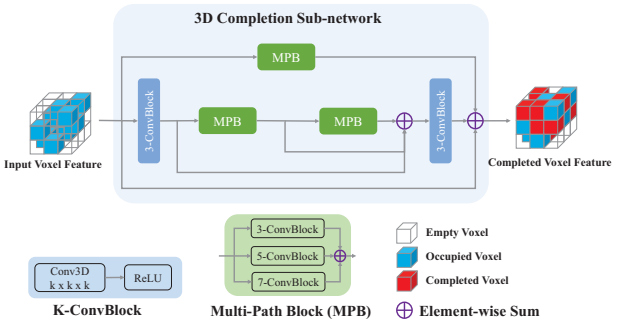

Figure 3. A schematic overview of the designed completion network. It is built upon the Multi-Path Blocks (MPBs) that contain 3×3×3, 5×5×5 and 7×7×7 convolution blocks (ConvBlock). In each convolution block, there is no convolution bias and no batch normalization to maintain sparsity during completion.

downsampling operations will inevitably lose the information of the original point cloud and the upsampling operation will cause over dilation and shape distortion.

To address the aforementioned problems, we make a comprehensive overhaul of the completion sub-network. More concretely, there are three principles in the design of the completion sub-network, *i.e.*, maintaining sparsity, no downsampling and aggregating multi-scale features.

Maintain sparsity. The completion sub-network needs the vanilla dense convolution for dilation while the segmentation sub-network uses sparse convolution for efficient processing. However, the bias of the convolution weight, running mean and beta of BN layers will break the sparsity of the original voxel features, thus substantially increasing the computation burden of the segmentation sub-network. Therefore, to reduce the overall computation cost and enjoy the high efficiency of sparse convolution, we remove all the convolution bias and the BN layers in the completion sub-network. In this condition, the voxel features produced by the completion sub-network can still keep the sparse property and the segmentation sub-network can use sparse convolution to process these sparse voxel features.

No downsampling. In popular completion networks such as S3CNet [3] and JS3C-Net [29], there are several downsampling and upsampling blocks in the completion part. The downsampling operations will inevitably lose the information of the original point cloud, causing severe completion and classification errors for small objects and crowded scenes. Therefore, we discard all downsampling and upsampling operations to relieve the information loss, maximally retaining the information of the raw point cloud. Besides, as opposed to JS3C-Net which takes the segmentation-first baseline, we adopt the completion-first principle. Concretely, we make the completion sub-network directly process the raw voxel features produced by the voxelization process. And the completion sub-network can also benefit from the large number of parameters of the segmentation sub-network.

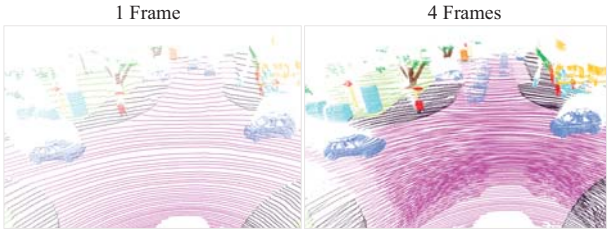

Figure 4. Comparison between single-frame and multi-frame point cloud. The multi-frame input is obviously denser than the single frame and the objects are easier to be identified, significantly reducing the completion difficulty.

Aggregate multi-scale features. To aggregate multi-scale features, we design the multi-path block which is comprised of 3×3×3, 5×5×5 and 7×7×7 convolution blocks. As shown in Fig. 3, there are three branches in the completion sub-network. The upper branch contains one MPB and the bottom branch is a residual connection. The middle branch is constructed by a 3×3×3 convolution block, two MPBs and a 3×3×3 convolution block. After the completion sub-network, we obtain the dense completed voxel features. We extract the non-empty voxel features as well as their voxel indices from the completed voxel features. The generated sparse voxel features are sent to the segmentation sub-network to produce the voxelwise segmentation output.

Modifications on the segmentation sub-network. Recall that, for the segmentation part, we take the Cylinder3D [36] as the backbone. Since the voxelwise completion labels are defined based on the cubic partition, we replace the cylindrical partition of Cylinder3D with conventional cubic partition. Besides, the original point refinement module consumes much GPU memory and brings limited gains, we discard this module to save memory usage.

3.3. Distilling Multi-frame Knowledge

Since single frame point cloud is sparse and incomplete in the outdoor scenarios, directly performing semantic scene completion from the single-frame input is extremely difficult. It is natural to wonder if we can construct a multi-frame model and then distil the dense semantic knowledge from this multi-frame network. From Fig. 4, it is evident that the multi-frame input significantly reduces the completion difficulty since multiple frames have much more points in the scene and objects are easier to be identified. The completion difficulty will gradually decrease as the number of input point cloud frames increases. Therefore, we construct the multi-frame teacher which takes denser point cloud as input and achieves better completion performance.

Inspired by [12], we make the single-frame model distil the relation-based structural knowledge from the multi-frame teacher network. Since the original voxel features

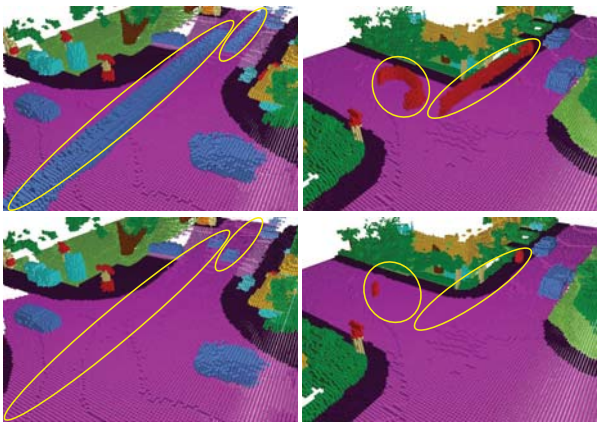

Figure 5. Top: original completion labels which have long traces of dynamic objects, *e.g.*, car and person. Bottom: rectified completion labels. The proposed label rectification operation can effectively remove the long traces of moving objects, making the completion labels more accurate.

are in the sparse form, we leverage the sparse features and their indices to perform knowledge distillation. We denote the voxel features and corresponding indices of the teacher and student as $F_T \in \mathbb{R}^{N_m \times C_f}$, $F_S \in \mathbb{R}^{N_s \times C_f}$, $\mathcal{I}_T \in \mathbb{R}^{N_m \times 3}$ and $\mathcal{I}_S \in \mathbb{R}^{N_s \times 3}$, respectively. N_m is the number of non-empty voxel features in the multi-frame, N_s is the number of non-empty voxel features in the single-frame, C_f is the number of channels of the voxel features. Note that the indices of teacher features and student features are sorted and $\mathcal{I}_S(i,j) = \mathcal{I}_T(i,j)$, where $i \in \{1, ..., N_s\}$ and $j \in \{1, 2, 3\}$. We first compute the pairwise relational knowledge of the student model:

$$\mathbf{P}_S(i,j) = \frac{F_S(i)^\mathsf{T} F_S(j)}{\|F_S(i)\|_2 \|F_S(j)\|_2}, i, j \in \{1, ..., N_s\} \quad (1)$$

The relational knowledge of the teacher model $\mathbf{P}_T$ is calculated in a similar way. The relational knowledge captures the similarity of each pair of voxel features and serve as important clues of the surrounding environment, which can be taken as high-level knowledge to be learned by the single-frame student model. The proposed Dense-to-Sparse Knowledge Distillation (DSKD) loss is given as below:

$$\mathcal{L}_{\text{dskd}}(\mathbf{P}_S, \mathbf{P}_T) = \frac{1}{N_s^2} \sum_{i=1}^{N_s} \sum_{j=1}^{N_s} \|\mathbf{P}_S(i,j) - \mathbf{P}_T(i,j)\|_2^2. \quad (2)$$

3.4. Completion Label Rectification

Generation of completion labels. As pointed out by [20], for outdoor semantic scene completion, the ground-truth completion labels are obtained by concatenating the segmentation labels of multiple consecutive point cloud frames. Specifically, for the t-th frame, the corresponding completion labels L_t^c are constructed in the following way:

$$L_t^c = \text{concat}[L_t^s; T_{t+1 \to t} L_{t+1}^s; ...; T_{t+T-1 \to t} L_{t+T-1}^s], \quad (3)$$

where L_t^s are the segmentation labels for the t-th frame, T is the number of frames used for concatenation, $T_{t+1 \to t}$ is the transformation matrix that transforms the coordinate from $(t+1)$-th frame to the t-th frame, and $\text{concat}[...;...]$ is the concatenating operation. The concatenation of multiple frames will lead to long traces for those moving objects, such as car and person. A vivid example is shown in Fig. 5. The long traces of those dynamic objects are obviously irrational and will hamper the learning of deep models.

Completion label rectification. To remove the long traces of moving objects in the completion labels, we resort to the panoptic segmentation labels. Specifically, given the panoptic labels of class i, we first voxelize the labels and obtain the voxelwise panoptic labels for class i. For each instance of class i, we calculate the bound of each instance, forming a cube. We union the cubes of all instances and use them to process the original voxelwise completion labels, filtering those voxels that are outside the cubes. The process is repeated for all classes that contain dynamic objects. The detailed information of the label rectification algorithm is shown in Algorithm 1. As shown in Fig. 5, the pro-

Algorithm 1 Pseudocode of Label Rectification.

```
# L_c: completion label (256x256x32)
# L_s: semantic segmentation label (Nx1)
# L_p: panoptic segmentation label (Nx1)
# S_c: classes of moving objects; u: unlabeled class

for C_i in S_c:
  # voxel position of completion label for class C_i
  P_c_i = getInd(L_c, C_i)
  L_p_i = L_p[L_s == C_i] # instance label of class C_i
  P_dif = []
  if len(L_p_i) == 0:
    P_dif = P_c_i
  elif len(L_p_i) > 0 and len(P_c_i) > 0:
    # mask of voxel position bound for instance d_j
    M_c = zeros([256, 256, 32])
    # point position of class C_i
    P_i = P[L_s == C_i, :]
    # voxel position of class C_i
    P_s_i = voxelize(P_i)
    D_i = unique(L_p_i) # all instance ids
    for d_j in D_i:
       # voxel position for instance d_j
       M_j = P_s_i[L_p_i == d_j]
       (x_min,x_max,y_min,y_max,z_min,z_max) = bound(
            M_j)
       M_c[x_min:x_max,y_min:y_max,z_min:z_max] = 1000
    # voxel position for instance label of class C_i
    P_p_i = getInd(M_c, 1000)
    P_dif = difference(P_c_i, P_p_i)
  if len(P_dif) > 0:
    for p_i in range(P_dif.shape[0]):
      L_c[P_dif[p_i,0],P_dif[p_i,1],P_dif[p_i,2]] = u
```

`getInd(A, b)`: get indices of b in A; `bound`: get bound of matrix; `difference(A, B)`: difference set of A minus B.

Table 1. Quantitative results of semantic scene completion algorithms on SemanticKITTI test set. Note that the online server still uses the **original** completion labels to evaluate algorithms. **Bold** - best in column for all single-frame methods.

Methods	mIoU	completion	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign
SSCNet-Full [22]	16.1	50.0	24.3	0.5	0.8	1.2	4.3	0.3	0.3	0.0	51.2	27.1	30.8	6.4	34.5	19.9	35.3	18.2	29.0	13.1	6.7
ESSCNet [34]	17.5	41.8	26.4	0.3	5.4	5.0	9.1	2.9	2.7	0.1	43.8	26.9	28.1	10.3	29.8	23.3	35.8	20.1	28.7	16.4	16.7
LMSCNet-SS [19]	17.6	56.7	30.9	0.0	0.0	1.5	0.8	0.0	0.0	0.0	64.8	29.0	34.7	4.6	38.1	21.3	41.3	19.9	32.1	15.0	0.8
TDS [7]	17.7	50.6	29.5	0.0	0.0	2.5	0.1	0.0	0.0	0.0	62.2	23.3	31.6	6.5	34.1	24.1	40.1	21.9	33.1	16.9	6.9
UDNet [37]	19.5	**59.4**	33.9	0.8	0.4	3.8	4.4	0.5	0.3	0.3	62.0	28.2	35.1	9.1	39.5	24.4	40.9	23.2	32.3	18.8	13.1
Local-DIFs [18]	22.7	57.7	34.8	3.6	2.4	4.4	4.8	2.5	1.1	0.0	67.9	40.1	42.9	11.4	40.4	29.0	42.2	26.5	39.1	21.3	17.5
SSA-SC [32]	23.5	58.8	36.5	13.9	4.6	5.7	7.4	4.4	2.6	0.7	72.2	37.4	43.7	10.9	43.6	30.7	43.5	25.6	41.8	14.5	6.9
JS3C-Net [29]	23.8	56.6	33.3	14.4	8.8	7.2	12.7	8.0	5.1	0.4	64.7	34.9	39.9	14.1	39.4	30.4	43.1	19.6	40.5	18.9	15.9
S3CNet [3]	29.5	45.6	31.2	**41.5**	**45.0**	6.7	16.1	**45.9**	**35.8**	**16.0**	42.0	17.0	22.5	7.9	**52.2**	31.3	39.5	34.0	21.2	31.0	24.3
SCPNet (#frame=1)	**36.7**	56.1	**46.4**	33.2	34.9	**13.8**	**29.1**	28.2	24.7	1.8	**68.5**	**51.3**	**49.8**	**30.7**	38.8	**44.7**	**46.4**	**40.1**	**48.7**	**40.4**	**25.1**
SCPNet (#frame=4)	47.5	68.5	59.0	48.5	51.1	21.0	37.8	47.6	35.0	10.5	79.7	57.7	60.0	32.8	50.2	54.1	56.9	47.1	58.0	48.3	48.0

posed label rectification operation can effectively remove the long traces of moving objects, making the completion labels more accurate.

3.5. Overall objective

The overall loss function is comprised of three terms, *i.e.*, the cross entropy loss, the lovasz-softmax loss [2] and the proposed distillation loss.

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \alpha\mathcal{L}_{\text{lovasz}} + \beta\mathcal{L}_{\text{dskd}}, \tag{4}$$

where α and β are the loss coefficients to balance the effect of each loss term.

4. Experiments

Datasets. Following the practice of popular scene completion models [3, 29], we conduct experiments on two popular LiDAR semantic scene completion benchmarks, *i.e.*, SemanticKITTI [1] and SemanticPOSS [16]. As to SemanticKITTI, it has 22 point cloud sequences. Sequences 00 to 10, 08 and 11 to 21 are used for training, validation and testing, respectively. 19 classes are chosen for training and evaluation after merging classes with distinct moving status and discarding classes with very few points. As for SemanticPOSS, it has 2, 988 frames and 11 classes are selected for evaluation. Although SemanticPOSS is smaller than SemanticKITTI in terms of dataset size, it is much more challenging since it contains a larger quantity of moving objects than SemanticKITTI, such as person and rider.

Evaluation metrics. Following [12, 36], we adopt the intersection-over-union (IoU) of each class and mIoU of all classes as the evaluation metric. The IoU of class i is calculated via: $IoU_i = \frac{TP_i}{TP_i+FP_i+FN_i}$, where TP_i, FP_i and FN_i denote the true positive, false positive and false negative of class i, respectively. For semantic scene completion, the dimension of the completion label is $256 \times 256 \times 32$. We also report the completion mIoU which is the class-agnostic version of mIoU. Note that mIoU is the major evaluation metric for the semantic scene completion task.

Implementation details. The output size of Cylinder3D is set as $256 \times 256 \times 32$ to adapt to the completion task. The number of training epochs is set as 30 and the initial learning rate is set as 0.0015. We use Adam [14] as the optimizer. Gradient norm clip is set 10 to stabilize the training process. α and β are set as 1 and 3, 000, respectively. We filter points that are outside the point cloud range of the voxelwise completion labels. For SemanticKITTI, we first train the model on the training set and then finetune it on both training and validation sets before submitting the predictions of test set to the online server.

Baseline KD algorithms. We take the vanilla knowledge distillation [8], FitNets [21], NST [13], PKT [17] and PVKD [12] as baseline distillation algorithms. Vanilla knowledge distillation takes the softened logits as the distilled knowledge. FitNets directly mimics the teacher features. NST adopts the maximum mean discrepancy to minimize the distance between student features and teacher features. PKT models the teacher knowledge as a probability distribution and then forces the consistency of the probability distribution between the teacher and the student. PVKD distils the voxelwise output and inter-voxel affinity knowledge. And we discard the original pointwise output distillation and inter-point affinity distillation of PVKD since they consume much GPU memory and bring marginal gains.

4.1. Results

Quantitative results. We summarize the performance of SCPNet and state-of-the-art semantic scene completion methods in Table 1 and 2. On SemanticKITTI, our SCPNet significantly outperforms other scene completion algorithms in terms of mIoU. For example, our SCPNet is **7.2** mIoU higher than S3CNet [3]. On classes such as car, other-vehicle, road, parking, sidewalk, fense, terrain and other-ground, the performance gap between SCPNet and S3CNet is more than **10** IoU. Our SCPNet also achieves superior performance on SemanticPOSS val set. On classes such as car, trunk, pole, fence and bike, SCPNet is at least **5** IoU higher than JS3C-Net [29]. The impressive performance on

Table 2. Quantitative results of semantic scene completion algorithms on SemanticPOSS val set. Note that the completion labels are processed by the proposed rectification algorithm. **Bold** - best in column for all single-frame methods.

Methods	mIoU	completion	person	rider	car	trunk	plants	traffic-sign	pole	building	fence	bike	ground
SSCNet-Full [22]	15.2	53.5	5.3	0.3	1.3	5.6	39.6	1.0	2.6	28.7	3.4	26.0	43.1
LMSCNet-SS [19]	16.5	52.3	7.7	0.3	0.6	4.0	37.7	1.9	8.2	36.8	13.8	25.8	45.1
MotionSC [25]	17.6	52.7	7.8	0.5	0.5	3.9	39.8	2.2	8.5	39.2	13.1	30.8	47.0
JS3C-Net [29]	22.7	**58.1**	**18.9**	0.2	7.1	3.6	47.8	2.2	0.0	46.3	26.6	43.4	**53.2**
SCPNet (#frame=1)	**26.3**	56.3	9.2	**3.3**	**12.4**	**11.0**	**49.6**	**3.1**	**11.1**	**50.1**	**40.8**	**48.7**	49.8
SCPNet (#frame=4)	33.6	60.7	21.4	5.0	34.5	8.9	55.3	7.8	31.0	50.3	47.9	55.1	52.3

Table 3. Quantitative results of our SCPNet and state-of-the-art LiDAR semantic segmentation methods on SemanticKITTI test set. C3D + SCPNet denotes Cylinder3D initialized from the trained weight of SCPNet. **Bold** - best in column.

Methods	**mIoU**	Latency (ms)	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic
SPVNAS [23]	66.4	259	97.3	51.5	50.8	59.8	58.8	65.7	65.2	43.7	90.2	67.6	75.2	16.9	91.3	65.9	86.1	73.4	71.0	64.2	66.9
AF2S3Net [4]	69.7	–	94.5	65.4	86.8	39.2	41.1	**80.7**	80.4	**74.3**	91.3	68.8	72.5	**53.5**	87.9	63.2	70.2	68.5	53.7	61.5	71.0
RPVNet [28]	70.3	–	**97.6**	**68.4**	68.7	44.2	61.1	75.9	74.4	73.4	93.4	70.3	80.7	33.3	**93.5**	72.1	**86.5**	**75.1**	71.7	64.8	61.4
PVKD [12]	71.2	**76**	97.0	67.9	**69.3**	53.5	60.2	75.1	73.5	50.5	91.8	70.9	77.5	41.0	92.4	69.4	**86.5**	73.8	**71.9**	64.9	65.8
2DPASS [30]	**72.9**	–	97.0	63.6	63.4	**61.1**	61.5	77.9	**81.3**	74.1	89.7	67.4	74.7	40.0	93.5	**72.9**	86.2	73.9	71.0	65.0	70.4
Cylinder3D [36]	68.9	170	97.1	67.6	63.8	50.8	58.5	73.7	69.2	48.0	92.2	65.0	77.0	32.3	90.7	66.5	85.6	72.5	69.8	62.4	66.2
C3D + SCPNet	71.5		97.5	60.9	56.3	58.6	**65.9**	70.7	71.8	58.7	**93.6**	**72.1**	**80.9**	36.2	93.3	72.1	86.2	74.1	71.6	**66.7**	**71.8**

Table 4. Comparison between different knowledge distillation algorithms and the proposed DSKD on the SemanticKITTI val set.

Methods	mIoU
SCPNet w/o DSKD	34.4
+ KD [8]	33.8
+ FitNets [21]	33.8
+ PKT [17]	34.6
+ PVKD [12]	36.2
+ NST [13]	36.3
+ DSKD	**37.2**

two large-scale benchmarks strongly demonstrate the superiority of our SCPNet .

Besides, we use the trained weight of the segmentation sub-network as initialization to train Cylinder3D on the SemanticKITTI semantic segmentation task. From Table 3, Cylinder3D initialized from trained weight of the completion task outperforms the original Cylinder3D model by **2.6** mIoU, and achieves impressive segmentation performance among various competitive LiDAR segmentation models such as 2DPASS [30], PVKD [12] and RPVNet [28]. The encouraging results show that knowledge learned in the completion task is also beneficial to the segmentation task.

Comparison with baseline KD algorithms. From Table 4, it is evident that the proposed DSKD method can bring more gains than conventional knowledge distillation algorithms. For instance, compared with FitNets [21] which directly mimics the teacher features, our DSKD can bring more than **2** mIoU, showing the effectiveness of the proposed relation-based distillation algorithm. The vanilla KD objective and FitNets hamper the performance of the base model, indicating that directly mimicking the logits or features can not boost the completion performance.

Qualitative results. We also provide visual comparison between JS3C-Net [29], SCPNet (single-frame) and SCPNet (multi-frame). As can be seen from Fig. 6, our SCPNet (single-frame) make more accurate completion predictions than JS3C-Net on road and vegetation. On long, thin objects such as poles, our single-frame model also yields high-quality completion results compared with JS3C-Net. The predictions of our single-frame model also resemble those of the multi-frame network, demonstrating the efficacy of the proposed DSKD algorithm.

4.2. Ablation studies

Experiments are conducted in SemanticKITTI val set.

Effect of completion label rectification. We report the performance of our SCPNet on completion labels with and without rectification. From Table 5, it is evident that the proposed rectification strategy greatly enhances the performance of SCPNet on those dynamic objects, *e.g.*, car and person. For example, the completion label rectification can bring **8.1**, **24.5**, **26.1** and **17.6** IoU improvement on car, person, bicyclist and motorcyclist, respectively. The impressive performance gains strongly demonstrate the effectiveness of the label rectification algorithm.

Effect of the completion sub-network. To examine the

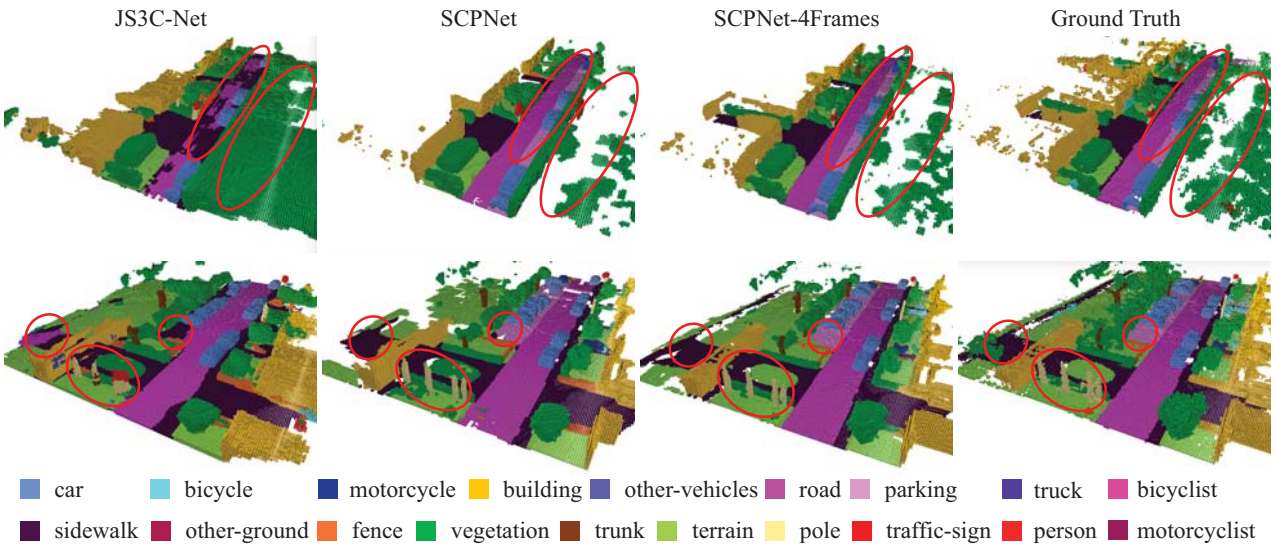

Figure 6. Visual comparison of different methods on the SemanticKITTI validation set. From left to right: predictions of JS3C-Net [29], SCPNet (single frame), SCPNet (multi-frame) and ground-truth. Regions that have large prediction errors are highlighted by red ellipses.

Table 5. Impact of completion label rectification on the performance.

Methods	mIoU	completion	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign
SCPNet	37.2	49.9	50.5	**28.5**	31.7	**58.4**	41.4	19.4	19.9	0.2	70.5	**60.9**	52.0	**20.2**	**34.1**	33.0	35.3	**33.7**	51.9	38.3	27.5
SCPNet + label rectification	**40.8**	**52.1**	**58.6**	27.0	**34.7**	54.2	**41.9**	**43.9**	**46.0**	**17.8**	**70.6**	60.2	**53.1**	7.7	33.9	32.2	**41.9**	32.0	**52.4**	**38.7**	**28.9**

Table 6. Impact of the completion sub-network, DSKD and downsampling operation on the performance.

(a) Completion sub-network

	mIoU
JS3C-Net [29]	24.0
JS3C-Net + completion-subnet	**30.8**

(b) DSKD

	mIoU
SCPNet w/o DSKD	34.4
SCPNet w/ DSKD	**37.2**

(c) Downsampling operation

	mIoU
SCPNet w/o downsampling	**37.2**
SCPNet w/ downsampling	33.1

effect of our completion sub-network, we add it to JS3C-Net. The detailed performance is shown in Table 6 (a). Our completion sub-network can bring **6.8** mIoU improvement to JS3C-Net, which strongly demonstrates the effectiveness and generalization of the proposed completion sub-network.

Effect of DSKD. We compare the performance of our SCPNet with and without the proposed DSKD in Table 6 (b). The proposed distillation method can bring **2.8** mIoU improvement to our SCPNet , showing the benefit of distilling relation-based knowledge from the multi-frame model.

Effect of the downsampling operation. We add the downsampling operation to our completion sub-network and examine its effect. As reported in Table 6 (c), the downsampling operation hampers the completion performance of our SCPNet . Specifically, the completion performance of SCPNet decreases from 37.2 mIoU to 33.1 mIoU. The negative results show that the no-downsampling principle is vital to the success of the completion sub-network redesign.

5. Conclusion

To address the challenges of the semantic scene completion task, we propose three solutions from the aspects of the completion network redesign, dense-to-sparse knowledge distillation as well as completion label rectification. The resulting completion network, termed SCPNet , achieves superior completion performance in two large-scale semantic scene completion benchmarks, *i.e.*, SemanticKITTI and SemanticPOSS. The learned knowledge on the completion task is also beneficial to the semantic segmentation task.

Acknowledgements. This work is partially supported by the National Key R&D Program of China (No. 2022ZD0160100),and in part by Shanghai Committee of Science and Technology (Grant No. 21DZ1100100).

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of Lidar Sequences. In *IEEE International Conference on Computer Vision*, pages 9297–9307, 2019. 1, 2, 6

[2] Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. The Lovasz-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2018, pages 4413–4421, 2018. 6

[3] Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Liu Bingbing. S3CNet: A Sparse Semantic Scene Completion Network for LiDAR Point Clouds. In *Conference on Robot Learning*, pages 2148–2161. PMLR, 2021. 1, 2, 4, 6

[4] Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, and Bingbing Liu. (af)2-S3Net: Attentive Feature Fusion with Adaptive Feature Selection for Sparse Semantic Segmentation Network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12547–12556, 2021. 7

[5] Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, and Matthias Nießner. Scancomplete: Large-scale Scene Completion and Semantic Segmentation for 3D Scans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2018. 2

[6] Ben Fei, Weidong Yang, Wenming Chen, Zhijun Li, Yikang Li, Tao Ma, Xing Hu, and Lipeng Ma. Comprehensive Review of Deep Learning-based 3D Point Clouds Completion Processing and Analysis. *arXiv preprint arXiv:2203.03311*, 2022. 2

[7] Martin Garbade, Yueh-Tung Chen, Johann Sawatzky, and Juergen Gall. Two Stream 3D Semantic Scene Completion. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 6

[8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *Statistics*, 1050:9, 2015. 3, 6, 7

[9] Yuenan Hou, Zheng Ma, Chunxiao Liu, Tak-Wai Hui, and Chen Change Loy. Inter-Region Affinity Distillation for Road Marking Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12486–12495, 2020. 3

[10] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning Lightweight Lane Detection CNNs by Self Attention Distillation. In *IEEE International Conference on Computer Vision*, pages 1013–1021, 2019. 3

[11] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning to Steer by Mimicking Features from Heterogeneous Auxiliary Networks. In *Association for the Advancement of Artificial Intelligence*, volume 33, pages 8433–8440, 2019. 3

[12] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-Voxel Knowledge Distillation for LiDAR Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8479–8488, 2022. 3, 4, 6, 7

[13] Zehao Huang and Naiyan Wang. Like What You Like: Knowledge Distill via Neuron Selectivity Transfer. *arXiv preprint arXiv:1707.01219*, 2017. 6, 7

[14] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015. 6

[15] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and Think: Disentangling Semantic Scene Completion. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[16] Yancheng Pan, Biao Gao, Jilin Mei, Sibo Geng, Chengkun Li, and Huijing Zhao. SemanticPOSS: A Point Cloud Dataset with Large Quantity of Dynamic Instances. In *IEEE Intelligent Vehicles Symposium*, pages 687–693. IEEE, 2020. 2, 6

[17] Nikolaos Passalis and Anastasios Tefas. Learning Deep Representations with Probabilistic Knowledge Transfer. In *European Conference on Computer Vision*, pages 268–284, 2018. 6, 7

[18] Christoph B Rist, David Emmerichs, Markus Enzweiler, and Dariu M Gavrila. Semantic Scene Completion using Local Deep Implicit Functions on LiDAR Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7205–7218, 2021. 6

[19] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. LMSCNet: Lightweight Multiscale 3D Semantic Completion. In *International Conference on 3D Vision (3DV)*, pages 111–119. IEEE, 2020. 1, 2, 6, 7

[20] Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. 3d Semantic Scene Completion: A Survey. *International Journal of Computer Vision*, pages 1–28, 2022. 1, 2, 5

[21] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for Thin Deep Nets. In *International Conference on Learning Representations*, 2015. 3, 6, 7

[22] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic Scene Completion from a Single Depth Image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1754, 2017. 1, 2, 6, 7

[23] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution. In *European Conference on Computer Vision*, pages 685–702. Springer, 2020. 7

[24] Fred Tung and Greg Mori. Similarity-Preserving Knowledge Distillation. In *IEEE International Conference on Computer Vision*, pages 1365–1374, 2019. 3

[25] Joey Wilson, Jingyu Song, Yuewei Fu, Arthur Zhang, Andrew Capodieci, Paramsothy Jayakumar, Kira Barton, and Maani Ghaffari. MotionSC: Data Set and Network for Real-Time Semantic Mapping in Dynamic Environments. *arXiv preprint arXiv:2203.07060*, 2022. 1, 2, 7

[26] Xiaohan Xing, Yuenan Hou, Hang Li, Yixuan Yuan, Hongsheng Li, and Max Q-H Meng. Categorical Relation-preserving Contrastive Knowledge Distillation for Medical

Image Classification. In *Medical Image Computing and Computer Assisted Intervention*, pages 163–173. Springer, 2021. 3

[27] Guodong Xu, Yuenan Hou, Ziwei Liu, and Chen Change Loy. Mind the Gap in Distilling StyleGANs. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 3

[28] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu. RPVnet: A Deep and Efficient Range-Point-Voxel Fusion Network for Lidar Point Cloud Segmentation. In *IEEE International Conference on Computer Vision*, pages 16024–16033, October 2021. 7

[29] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse Single Sweep LiDAR Point Cloud Segmentation via Learning Contextual Shape Priors from Scene Completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3101–3109, 2021. 1, 2, 3, 4, 6, 7, 8

[30] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2DPASS: 2D Priors Assisted Semantic Segmentation on LiDAR Point Clouds. In *European Conference on Computer Vision (ECCV)*, 2022. 7

[31] Jihan Yang, Shaoshuai Shi, Runyu Ding, Zhe Wang, and Xiaojuan Qi. Towards Efficient 3D Object Detection with Knowledge Distillation. *arXiv preprint arXiv:2205.15156*, 2022. 3

[32] Xuemeng Yang, Hao Zou, Xin Kong, Tianxin Huang, Yong Liu, Wanlong Li, Feng Wen, and Hongbo Zhang. Semantic Segmentation-assisted Scene Completion for LiDAR Point Clouds. In *IEEE International Conference on Intelligent Robots and Systems*, pages 3555–3562. IEEE, 2021. 1, 2, 6

[33] Sergey Zagoruyko and Nikos Komodakis. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *International Conference on Learning Representations*, 2017. 3

[34] Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient semantic scene completion network with spatial group convolution. In *European Conference on Computer Vision*, pages 733–749, 2018. 6

[35] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Wei Li, Yuexin Ma, Hongsheng Li, Ruigang Yang, and Dahua Lin. Cylindrical and Asymmetrical 3D Convolution Networks for Lidar-based Perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[36] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and Asymmetrical 3D Convolution Networks for Lidar Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9939–9948, 2021. 3, 4, 6, 7

[37] Hao Zou, Xuemeng Yang, Tianxin Huang, Chujuan Zhang, Yong Liu, Wanlong Li, Feng Wen, and Hongbo Zhang. Up-to-Down Network: Fusing Multi-Scale Context for 3D Semantic Scene Completion. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 16–23. IEEE, 2021. 6