Label-efficient Semantic Scene Completion with Scribble Annotations

Song Wang[1] , **Jiawei Yu**[1] , **Wentong Li**[1] , **Hao Shi**[1] , **Kailun Yang**[3] ,
Junbo Chen[2*] and **Jianke Zhu**[1*]

[1]Zhejiang University
[2]Udeer.ai
[3]Hunan University
jkzhu@zju.edu.cn, junbo@udeer.ai

Abstract

Semantic scene completion aims to infer the 3D geometric structures with semantic classes from camera or LiDAR, which provide essential occupancy information in autonomous driving. Prior endeavors concentrate on constructing the network or benchmark in a fully supervised manner. While the dense occupancy grids need point-wise semantic annotations, which incur expensive and tedious labeling costs. In this paper, we build a new label-efficient benchmark, named ScribbleSC, where the sparse scribble-based semantic labels are combined with dense geometric labels for semantic scene completion. In particular, we propose a simple yet effective approach called Scribble2Scene, which bridges the gap between the sparse scribble annotations and fully-supervision. Our method consists of geometric-aware auto-labelers construction and online model training with an offline-to-online distillation module to enhance the performance. Experiments on SemanticKITTI demonstrate that Scribble2Scene achieves competitive performance against the fully-supervised counterparts, showing 99% performance of the fully-supervised models with only 13.5% voxels labeled. Both annotations of ScribbleSC and our full implementation are available at https://github.com/songw-zju/Scribble2Scene.

1 Introduction

Semantic scene completion (SSC), also known as semantic occupancy estimation, aims to predict 3D *geometric* and *semantic* information about the whole scene. Recent studies have shown that occupancy grid can model the environments in a general representation [Tian *et al.*, 2023] and provide essential guidance for downstream tasks [Sima *et al.*, 2023; Liu *et al.*, 2024], which motivate a series of camera-based methods [Cao and de Charette, 2022; Li *et al.*, 2023c].

Current research on semantic scene completion mainly focuses on designing effective network structures under the fully supervised settings [Cao and de Charette, 2022; Li *et*

---

*Corresponding authors.

al., 2023c; Xia *et al.*, 2023], while few research work provides the solution to learn from less expensive sparse labels. Meanwhile, existing semantic scene completion [Behley *et al.*, 2019; Li *et al.*, 2023b] or occupancy estimation benchmarks [Wang *et al.*, 2023c; Wei *et al.*, 2023] all heavily rely on fully point-wise annotations for semantic segmentation on LiDAR point cloud, which not only incurs the expensive and tedious manual labeling but also limits their application in new scenarios. In this work, we firstly revisit the annotation on 3D semantic scene completion, and then propose the scribble-supervised paradigm for this task.

The ground truth for semantic scene completion contains both *geometric* and *semantic* parts that are obtained by accumulating multi-scan point clouds with point-wise semantic labels [Behley *et al.*, 2019; Wang *et al.*, 2022]. The *geometric* part can be directly extracted from raw LiDAR scans while the *semantic* one needs densely annotated labels. We aim to construct a label-efficient benchmark with only sparse semantic labels. Besides, previous efforts [Li *et al.*, 2023c; Xia *et al.*, 2023; Shi *et al.*, 2024] indicate that it is challenging for fully-supervised models to estimate geometry. It is even more severe to learn from sparse semantic labels, as most of the non-empty voxels are unlabeled without semantic information. To tackle this issue, we investigate the potential of *dense* geometry offline, thus significantly mitigating the reliance on dense semantic labels.

In this work, we make full use of the sparse annotations in ScribbleKITTI [Unal *et al.*, 2022] to generate scribble-based semantic occupancy labels combined with the dense geometric structure to construct a new benchmark called ScribbleSC. Specially, we develop a simple yet effective approach, dubbed as Scribble2Scene, which is the first weakly-supervised scheme for 3D semantic occupancy estimation. More importantly, it achieves similar performance compared with the existing fully supervised methods. Our proposed Scribble2Scene consists of two stages, including geometry-aware auto-labelers construction (Stage-I) and online model training with distillation (Stage-II).

At Stage-I, we construct geometry-aware auto-labelers with scribble annotations, including Dean-Labeler and Teacher-Labeler. Dean-Labeler treats the complete geometric structure as input, which converts this task into an easier semantic segmentation problem to obtain high-quality voxel-wise segmentation results. Teacher-Labeler is also trained in

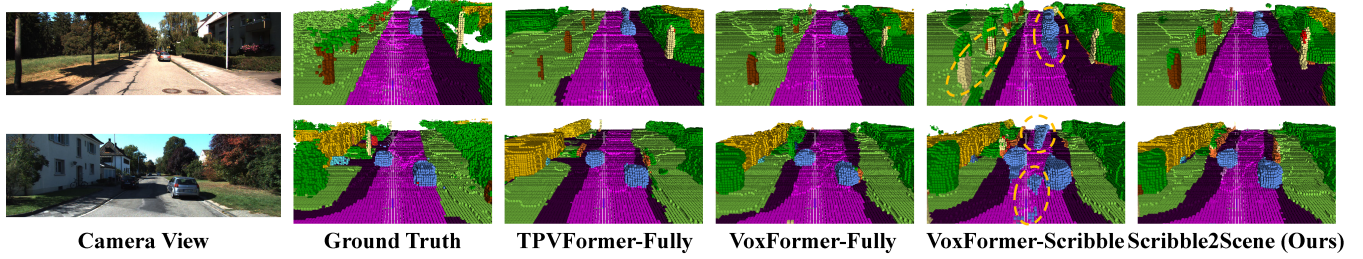| Camera View | Ground Truth | TPVFormer-Fully | VoxFormer-Fully | VoxFormer-Scribble | Scribble2Scene (Ours) |

Figure 1: **Comparisons between fully supervised methods and our proposed weakly scribble-supervised Scribble2Scene approach** for semantic scene completion on SemanticKITTI dataset. Due to limited pages, we provide full figure in comparison and related Supplementary Material in the arXiv version.

offline mode with both input image and complete geometry, which has the same network architecture as the online model. It has the capability to extract more accurate features and semantic logits for the online model. At Stage-II, we train the online completion network in a fully-supervised manner based on the pseudo labels provided by Dean-Labeler. In particular, a new range-guided offline-to-online distillation scheme is proposed for large-scale semantic scene understanding, which enhances the performance of the online model with the features from the trained Teacher-Labeler. Fig. 1 shows some qualitative results and comparisons.

Our main contributions are summarized as below:

- We revisit the annotation of semantic scene completion and propose a scribble-based label-efficient benchmark named ScribbleSC, which provides both sparse semantic annotations and dense geometric labels.

- We propose Scribble2Scene, the first weakly-supervised approach for semantic scene completion, designed to handle sparse scribble annotations. Geometry-aware auto-labelers construction and offline-to-online distillation training are devised to accurately predict 3D semantic occupancy.

- Under our presented Scribble2Scene framework, the camera-based scribble-supervised model achieves up to a competitive 99% performance of the fully-supervised one on SemanticKITTI without incurring the computational cost during inference. Additional experiments on SemanticPOSS demonstrate the generalization capability and robustness of our proposed scheme.

2 Related Work

Semantic Scene Completion. Semantic scene completion (SSC) is firstly proposed in SSCNet [Song *et al.*, 2017] to construct the complete 3D occupancy with voxel-wise semantic labels from a single-view observation. At the early stage, researchers mainly focus on the indoor scenarios [Liu *et al.*, 2018; Zhang *et al.*, 2018; Li *et al.*, 2020; Cai *et al.*, 2021] with RGB image or depth map as input. SemanticKITTI [Behley *et al.*, 2019] provides the first large dataset and benchmark in the outdoor for autonomous driving. The subsequent works mainly utilize the occupancy grid voxelized from the current LiDAR frame [Roldao *et al.*, 2020; Wilson *et al.*, 2022] or point cloud directly [Yan *et al.*, 2021;

Cheng *et al.*, 2021; Xia *et al.*, 2023; Mei *et al.*, 2023] as input and obtain promising performance. Recently, camera-based methods [Cao and de Charette, 2022; Huang *et al.*, 2023; Li *et al.*, 2023c; Zhang *et al.*, 2023; Yao *et al.*, 2023] attract more research attention due to their lower sensor costs. Vox-Former [Li *et al.*, 2023c] estimates the coarse geometry firstly and adopts the non-empty proposals to perform deformable cross-attention [Zhu *et al.*, 2021b] with image features, which achieves the best performance among camera-based models. Along this line, we mainly focus on *vision only* methods.

Sparsely Annotated Learning. Sparse annotations for 2D image segmentation are widely explored including scribble [Lin *et al.*, 2016; Liang *et al.*, 2022; Li *et al.*, 2024b], box [Tian *et al.*, 2021; Li *et al.*, 2022; Li *et al.*, 2024a], point [Bearman *et al.*, 2016; Fan *et al.*, 2022; Li *et al.*, 2023a] and etc. In 3D scene understanding, ScribbleKITTI [Unal *et al.*, 2022] re-annotates the KITTI Odometry dataset [Geiger *et al.*, 2012] and provides the scribble-supervised benchmark for LiDAR segmentation on SemanticKITTI [Behley *et al.*, 2019]. Box2Mask [Chibane *et al.*, 2022] adopts 3D bounding boxes to train dense segmentation models and achieves 97% performance of current fully-supervised models. We explore the potential of sparse annotations on the geometrically and semantically challenging task of semantic scene completion.

Teacher-Student Network. Knowledge Distillation (KD) is proposed initially to transfer the dark knowledge from a large trained teacher model to a small student one for model compression [Hinton *et al.*, 2015]. Following researchers achieve this goal in 2D image at not only logit-level [Cho and Hariharan, 2019; Furlanello *et al.*, 2018; Zhao *et al.*, 2022; Liu *et al.*, 2022] but also the feature-level [Romero *et al.*, 2015; Heo *et al.*, 2019a; Heo *et al.*, 2019b; Yang *et al.*, 2022] for in-depth exploration. Teacher-Student architectures in KD are widely adopted in various tasks and applications [Ye and Bors, 2022; Ye and Bors, 2023; Wang *et al.*, 2023b; Wang *et al.*, 2024]. In SSC, SCPNet [Xia *et al.*, 2023] is proposed to distill dense knowledge from a multi-scan model to a single one with pairwise relational information while its design is exclusively tailored for LiDAR-based methods. CleanerS [Wang *et al.*, 2023a] generates a perfect visible surface with ground truth voxels and trains a teacher model having cleaner knowledge in indoor scenarios. In this paper, we present a new offline-to-online distillation scheme, which is specially designed for 3D semantic scene completion in self-driving environments.

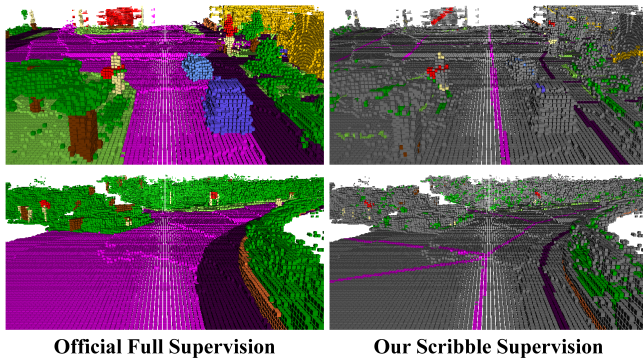**Official Full Supervision**   **Our Scribble Supervision**

Figure 2: Examples of the fully-annotated ground truth from SemanticKITTI (left) and scribble-annotated supervision from our constructed ScribbleSC (right).
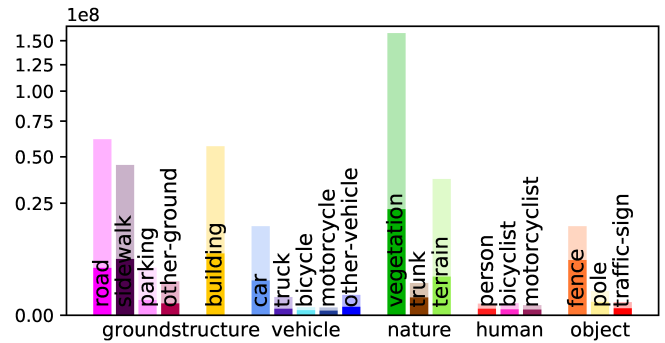

Figure 3: Quantity on each category of voxels labeled within ScribbleSC (deep color) in comparison to the fully-annotated SemanticKITTI dataset (light color). The total number of labeled voxels in ScribbleSC is only 13.5% over SemanticKITTI.

3 The ScribbleSC Benchmark

The supervision of semantic scene completion can be split into two parts, including *geometric* structure and *semantic* label. The *geometric* information can be easily obtained by accumulating exhaustive LiDAR scans and voxelizing the points that fall on the predefined region in front of the car. Meanwhile, the *semantic* label of each voxel is determined by the majority of labeled points within the voxel. The annotation of *semantic* part is highly dependent on the dense point-wise semantic segmentation labels, which requires an expensive and complicated labeling process. To construct a label-efficient benchmark, we make use of sparse annotations from ScribbleKITTI [Unal *et al.*, 2022] to replace the original full annotations provided by SemanticKITTI [Behley *et al.*, 2019] and achieve the construction of the *semantic* part. In ScribbleKITTI [Unal *et al.*, 2022], line-scribbles are adopted to label the accumulated point clouds and cover only 8.06% labeled points of the total training set including 10 sequences, which contribute to a 90% time saving[1].

Label Construction. Like SemanticKITTI [Behley *et al.*, 2019], we superimpose 70 future LiDAR scans to get the dense geometric structure and choose the volume of $51.2\,\mathrm{m}$ in the forward of car, 25.6 m to the left/right side and $6.4\,\mathrm{m}$ in height with an off-the-shelf voxelizer tool[2]. The voxel resolution is set to $0.2\,\mathrm{m}$ and a volume $\mathcal{V}$ of $256 \times 256 \times 32$ voxels can be obtained. We assign the *empty* label to voxels that are devoid of any points. For non-empty voxels, the semantic label is ascertained by conducting a majority vote across the scribble labels of the points situated within the voxel. If no labeled points exist in a non-empty voxel, we annotate it as *unlabeled*. Fig. 2 provides typical examples from ScribbleSC compared with the fully-annotated SemanticKITTI [Behley *et al.*, 2019]. Further information on the label construction for ScribbleSC is included in the Supplementary Material.

Label Usage Instruction. The label $\mathcal{V} \in \mathbb{R}^{X \times Y \times Z}$ $(X, Y, Z = 256, 256, 32)$ in ScribbleSC contains 19 semantic classes, an empty class, and an unlabeled class, which can

be split into full geometric annotation $\mathcal{G}$ and sparse semantic annotation $\mathcal{S}$. The geometric annotation $\mathcal{G} \in \mathbb{R}^{X \times Y \times Z}$ is a binary voxel grid map. Each voxel is marked as 0 if it is empty else 1 for occupied no matter with labeled or unlabeled points. The semantic annotation $\mathcal{S} \in \mathbb{R}^{X \times Y \times Z}$ contains a small number of voxels that are labeled with semantic classes. The vast majority of voxels in $\mathcal{S}$ are 0, including empty and unlabeled ones. Training directly with ScribbleSC inevitably encounters an imbalance between geometric and semantic supervision. We will introduce our solution in Sec. 4.

Voxel Labeling Statistics. We have conducted the distribution analysis of semantic labels in ScribbleSC and compared it with the fully annotated SemanticKITTI Benchmark as shown in Fig. 3. Our proposed semantic annotations only contain 13.5% labeled voxels over SemanticKITTI. ScribbleSC is a more challenging benchmark as it not only contains a substantial count of empty voxels but also a large number of unlabeled voxels among non-empty ones. More statistical analyses are given in the Supplementary Material.

4 Proposed Method

4.1 Overview of Scribble2Scene

Fig. 4 provides the overview of our proposed Scribble2Scene method for scribble-supervised semantic scene completion, which can be divided into two stages, *i.e.* geometry-aware auto-labelers construction (Stage-I) and online model training with distillation (Stage-II). At Stage-I, we make full use of complete geometric structure $\mathcal{G}$ and image $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$ observed from the current frame to construct Dean-Labeler and Teacher-Labeler with sparse scribble annotations, as shown in the left of Fig. 4. At Stage-II, we then adopt the pseudo labels generated by Dean-Labeler and perform the presented range-guided offline-to-online distillation with Teacher-Labeler to train the online model, as illustrated in the right of Fig. 4. All the models need to predict the complete 3D occupancy information $\mathcal{O} \in \mathbb{R}^{X \times Y \times Z \times C}$ in the predefined voxel space, where C is the number of total categories including empty and semantic classes. VoxFormer [Li *et al.*, 2023c] is employed as the baseline model, which is the state-of-the-art (SOTA) semantic scene completion network with only camera input. We delve into the specifics of Stage-I in Sec. 4.2

---

[1]Scribble labels cost around 10-25 minutes per tile, while full annotations cost 1.5-4.5 hours.

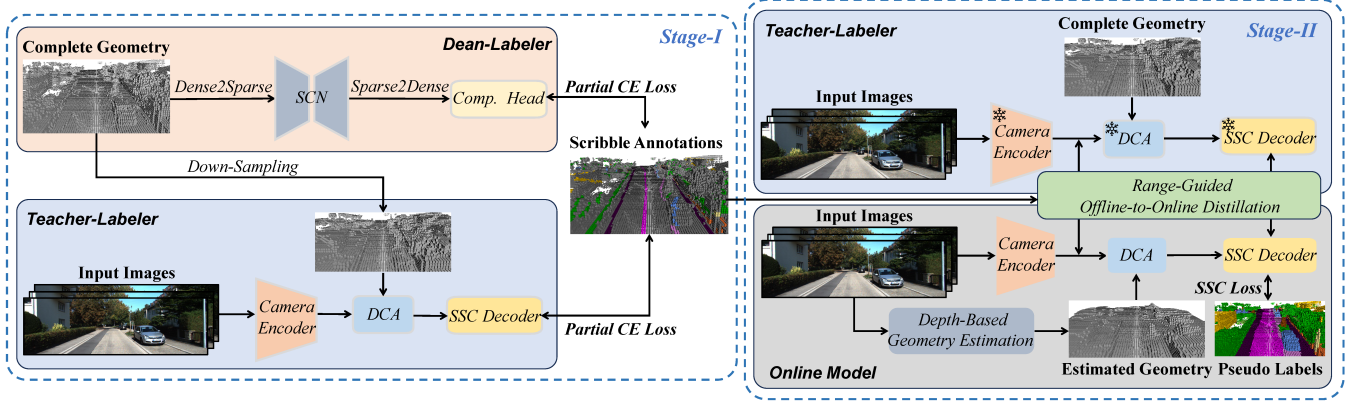[2]https://github.com/jbehley/voxelizer, MIT License

Figure 4: Overview of Scribble2Scene. The left half illustrates the offline **geometry-aware auto-labelers construction** at Stage-I. The right half shows the **online model training with distillation** at Stage-II. The accurate pseudo labels from Dean-Labeler and the well-trained Teacher-Labeler are fully leveraged for online model optimization.

and Stage-II in Sec. 4.3.

4.2 Geometry-Aware Auto-Labelers

To fully investigate the potential of the dense geometric structure and sparse semantic labels, we construct two geometry-aware auto-labelers (GA^{2}L) in the offline mode at Stage-I. The offline mode means that we can leverage the complete geometry $\mathcal{G}$ from the whole sequences to train a more performant model.

Dean-Labeler. Existing SOTA semantic scene completion models [Li *et al.*, 2023c; Mei *et al.*, 2023; Xia *et al.*, 2023] employ different branches to process geometric structure and semantic information, respectively. Moreover, their overall performance is often greatly limited by the inaccuracy of geometry estimation. If we directly treat the complete geometry $\mathcal{G}$ as input, the semantic scene completion can be converted into a voxel-wise semantic segmentation problem. Motivated by this, we adopt the sparse convolutional network (*SCN*) as the backbone of Dean-Labeler to obtain voxel-wise semantic prediction from the complete geometry, as illustrated in the top left of Fig. 4. Since the *SCN* does not change the geometric structure of the input, we only need to process the semantic part with scribble supervision. The partial cross-entropy loss function $\texttt{CE}(\cdot, \cdot)$ to train Dean-Labeler is formulated as

$$\mathcal{L}_{\text{partial\_ce}} = \sum_{i=1}^{X} \sum_{j=1}^{Y} \sum_{k=1}^{Z} \mathcal{S}_{i,j,k} \cdot \texttt{CE}(\mathcal{O}_{i,j,k}^{(D)}, \mathcal{S}_{i,j,k}), \quad (1)$$

where $\mathcal{O}^{(D)}$ is the predicted occupancy information from Dean-Labeler. We only perform optimization on voxels that are non-empty and labeled.

Leveraging complete geometry input, Dean-Labeler obtains the promising 3D semantic predictions $\mathcal{P}^{(D)}$(Sec. 5.3), which eliminates the imbalance between geometry (*dense*) and scribble-based semantic supervision (*sparse*) with ScribbleSC. The high-quality semantic predictions serve as the pseudo-labels for online model training at Stage-II.

Teacher-Labeler. To further explore the role of complete geometry $\mathcal{S}$, we design Teacher-Labeler with the modality-specific model VoxFormer [Li *et al.*, 2023c]. We replace

the noisy coarse geometry $\hat{\mathcal{G}}_{\frac{1}{2}} \in \mathbb{R}^{\frac{X}{2} \times \frac{Y}{2} \times \frac{Z}{2}}$ estimated from depth prediction as in VoxFormer with the down-sampling complete geometry $\mathcal{G}_{\frac{1}{2}} \in \mathbb{R}^{\frac{X}{2} \times \frac{Y}{2} \times \frac{Z}{2}}$. The deformable cross attention (*DCA*) is employed to sample image features with the precise non-empty proposals, as shown in the bottom left of Fig. 4. The *SSC Decoder* is made of a deformable self-attention (*DSA*) module and a completion head. The partial cross entropy loss in Eq. 1 is also adopted to train Teacher-Labeler. Additionally, we employ the geometric loss from MonoScene [Cao and de Charette, 2022] with $\mathcal{G}$ to alleviate the geometry change from the non-empty proposals when dense convolutions are used in the completion head.

Teacher-Labeler is trained with the precise non-empty proposals provided by complete geometry, which is able to extract more accurate features from input images. Therefore, the completion model can focus on the semantic part to make this task easier with only sparse scribble semantic annotations. In our experiments (Sec. 5.3), Teacher-Labeler achieves extremely higher performance compared to online models using noisy coarse geometry, which is further leveraged for online model training at Stage-II.

4.3 Online Model Training with Distillation

At Stage-II, we perform the online model training, which only employs currently observed information as inference input. The accurate pseudo labels provided by Dean-Labeler are utilized to replace the scribble labels so that we can optimize the model in a fully-supervised paradigm.

The online model as the student network has the same architecture as Teacher-Labeler while it cannot use the complete geometry $\mathcal{G}$ as input. To take advantage of the features and predictions obtained from models with the full geometry, we propose a novel range-guided offline-to-online distillation (RGO^{2}D) module that instructs the online model to learn auxiliary modality-specific knowledge at Stage-II. As shown in the right half of Fig. 4, we adopt the well-trained Teacher-Labeler as the offline teacher model and freeze the network weights to perform offline-to-online distillation for the online model, which alleviates the interference of inaccu-
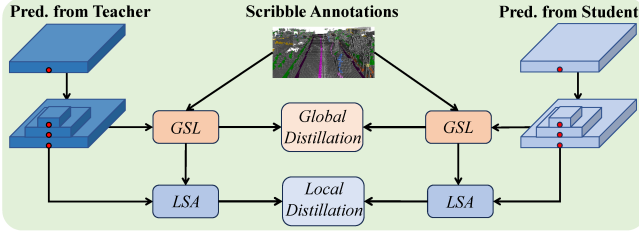
Figure 5: Illustration of **range-guided offline-to-online distillation**. The red dot denotes the location of the ego-vehicle. The global and local distillation with different ranges are performed, respectively.

rate pseudo-labels on training.

Range-Guided Offline-to-Online Distillation. When there is a large difference in the input and network performance of the teacher and student models, directly minimizing the Kullback-Leibler (KL) divergence or other metrics to align the outputs, *i.e.* $\mathcal{O}^{(S)} \in \mathbb{R}^{X \times Y \times Z \times C}$ from the student and $\mathcal{O}^{(T)} \in \mathbb{R}^{X \times Y \times Z \times C}$ from the teacher, often does not work well [Huang *et al.*, 2022]. Inspired by CleanerS [Wang *et al.*, 2023a], we adopt global semantic logit combined with local semantic affinity rather than the original predictions to perform distillation between the teacher and student at the logit level. Considering that the outdoor driving scene involves a wider range, only using the semantic logit and affinity of the whole scene cannot reflect the distribution of each class well. Moreover, the semantic logit and affinity of the area closer to the ego-vehicle are inherently more amenable to learning processes and endow more significance in ensuring vehicular safety assurances. Therefore, we consider introducing range information as guidance to model the distribution of each semantic class.

Specifically, we take the ego-vehicle as the center and divide the whole volume in the predictions of teacher $\mathcal{O}^{(T)}$ and student $\mathcal{O}^{(S)}$ into three ranges (including near, middle and far) according to the distance as shown in Fig. 5. Then the global semantic logits (*GSL*) of teacher $\{o^{(T,r)}\}_{r=1}^{3}$ and student $\{o^{(S,r)}\}_{r=1}^{3}$ at different ranges are calculated with scribble annotations $\mathcal{S}$. To adequately measure the difference of global semantic logits between student and teacher, we introduce the inter-relation and intra-relation loss to perform global distillation

$$\mathcal{L}_{\text{global},r} = \frac{1}{C}[\alpha \sum_{i=1}^{C} d_{\text{p}}(o_{i,:}^{(S,r)}, o_{i,:}^{(T,r)}) + \beta \sum_{j=1}^{C} d_{\text{p}}(o_{:,j}^{(S,r)}, o_{:,j}^{(T,r)})], \quad (2)$$

where d_{p} is Pearson's distance, and $r = 1, 2, 3$ denote different ranges. α and β are the balanced weights of the inter- and intra-relation loss, respectively. Moreover, we compute the local semantic affinity (*LSA*) with the global semantic logit and the prediction at each range for teacher $\{\mathcal{A}^{(T,r)}\}_{r=1}^{3}$ and student $\{\mathcal{A}^{(S,r)}\}_{r=1}^{3}$. Then local distillation loss is computed below

$$\mathcal{L}_{\text{local},r} = \text{MSE}(\mathcal{A}^{(S,r)}, \mathcal{A}^{(T,r)}), \quad (3)$$

where $\text{MSE}(\cdot, \cdot)$ denotes the mean square error function.

Additionally, we adopt $\text{MSE}(\cdot, \cdot)$ as the feature-level distillation loss $\mathcal{L}_{\text{feat}}$. We choose the 2D features from the image encoder as the targets to align. Overall, the range-guided

offline-to-online distillation is composed of the above items

$$\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{feat}} + \sum_{r=1}^{3} w_r \cdot (\mathcal{L}_{\text{global,r}} + \mathcal{L}_{\text{local},r}), \quad (4)$$

where w_r denotes the range loss coefficients, and $r = 1, 2, 3$. $\mathcal{L}_{\text{global,r}}$ and $\mathcal{L}_{\text{local,r}}$ represent the global and local distillation losses at near, middle, and far range, respectively.

4.4 Training and Inference

Overall Loss. The total loss $\mathcal{L}_{\text{total}}$ for the online model training consists of semantic loss $\mathcal{L}_{\text{sem}}$, geometric loss $\mathcal{L}_{\text{geo}}$ and distillation loss $\mathcal{L}_{\text{distill}}$ as below

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{distill}}, \quad (5)$$

where $\mathcal{L}_{\text{sem}}$ is the commonly used weighted cross-entropy loss. We employ the pseudo labels $\mathcal{P}^{(D)}$ from Dean-Labelers as the full semantic supervision. $\mathcal{L}_{\text{geo}}$ is the geometric scene-class affinity loss proposed in MonoScene [Cao and de Charette, 2022].

Inference. At the inference stage, we only need to preserve the student branch, which can obtain similar performance on accuracy while retaining efficient inference as the fully-supervised model.

5 Experiments

5.1 Experimental Setup

Dataset. Our models are trained on ScribbleSC. Unless specified, the performance is mainly evaluated on the *validation set* of the fully-annotated SemanticKITTI [Behley *et al.*, 2019], which is a highly challenging benchmark. All input images come from the KITTI Odometry Benchmark [Geiger *et al.*, 2012] consisting of 22 sequences. Following the official setting, we use the sequences 00-10 except 08 for training with ScribbleSC while sequence 08 is preserved as the *validation set*. We submit the predictions of sequences 11-21 to the online evaluation website and obtain the scores on the *hidden test set*. Additionally, we have conducted extra experiments on the SemanticPOSS [Pan *et al.*, 2020], which is another challenging dataset collected in a campus-based environment. Since the scribble-based annotations on point clouds of SemanticPOSS are unavailable, we randomly sample 10% of its full annotations to obtain similar sparse labels as scribbles. Then we construct semantic scene completion labels, including sparse semantic labels and dense geometric labels as described in Sec. 3. Adhering to the original configuration, the sequences (00-01, 03-05) / 02 are split as *training* and *validation set*, respectively.

Evaluation Protocol. We follow the official evaluation benchmark and employ intersection over union (IoU) to evaluate the scene completion performance, which only measures the class-agnostic geometric quality. The standard mean intersection over union metric (mIoU) of 19 semantic classes is reported for semantic scene completion. We choose the class-wise mIoU as the dominant evaluation metric. To comprehensively compare with the fully-supervised methods, we provide evaluation scores from three different ranges on *validation set* including $12.8 \times 12.8 \times 6.4 \text{m}^3$, $25.6 \times 25.6 \times 6.4 \text{m}^3$, and $51.2 \times 51.2 \times 6.4 \text{m}^3$.

Methods	Scribble2Scene (Ours)			VoxFormer† [CVPR'23]			VoxFormer [CVPR'23]			TPVFormer [CVPR'23]			MonoScene* [CVPR'22]		
Supervision	Scribble			Scribble			Fully			Fully			Fully		
Range	12.8m	25.6m	51.2m	12.8m	25.6m	51.2m	12.8m	25.6m	51.2m	12.8m	25.6m	51.2m	12.8m	25.6m	51.2m
car (3.92%)	40.06	33.15	24.16	31.02	24.99	17.77	44.90	37.46	26.54	34.81	31.72	23.79	24.34	24.64	23.29
bicycle (0.03%)	0.12	0.57	0.30	0.78	0.59	0.33	5.22	2.87	1.28	0.33	0.69	0.35	0.07	0.23	0.28
motorcycle (0.03%)	3.62	1.45	1.01	0.05	0.02	0.03	2.98	1.24	0.56	0.16	0.08	0.05	0.05	0.20	0.59
truck (0.16%)	14.32	17.06	17.32	7.39	5.18	4.35	9.80	10.38	7.26	17.77	13.15	6.92	15.44	13.84	9.29
other-veh. (0.20%)	11.59	5.76	3.69	4.02	1.51	0.87	17.21	10.61	7.81	10.06	7.47	4.29	1.18	2.13	2.63
person (0.07%)	5.01	3.53	1.98	4.02	3.81	2.27	4.44	3.50	1.93	1.56	1.06	0.52	0.90	1.37	2.00
bicyclist (0.07%)	1.33	1.08	0.47	3.59	4.85	2.67	2.65	3.92	1.97	2.57	1.93	0.91	0.54	1.00	1.07
motorcyclist (0.05%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
road (15.30%)	68.05	60.75	49.90	61.11	54.56	45.23	75.45	66.15	53.57	75.91	69.42	56.47	57.37	57.11	55.89
parking (1.12%)	20.88	22.47	20.12	14.45	16.52	16.19	21.01	23.96	19.69	29.88	26.22	20.59	20.04	18.60	14.75
sidewalk (11.13%)	44.43	35.71	26.93	37.88	30.27	20.67	45.39	34.53	26.52	47.05	36.69	25.83	27.81	27.58	26.50
other-grnd (0.56%)	0.00	0.40	0.87	0.00	0.12	0.57	0.00	0.76	0.42	0.00	1.36	0.94	1.73	2.00	1.63
building (14.10%)	25.76	30.62	20.14	21.41	25.86	16.52	25.13	29.45	19.54	11.37	18.23	13.89	16.67	15.97	13.55
fence (3.90%)	12.37	8.68	6.12	10.90	6.59	4.11	16.17	11.15	7.31	9.81	7.98	5.99	7.57	7.37	6.60
vegetation (39.3%)	44.00	38.35	25.99	39.30	33.87	21.56	43.55	38.07	26.10	24.90	24.32	16.93	19.52	19.68	17.98
trunk (0.51%)	21.23	14.00	8.03	17.10	11.35	5.86	21.39	12.75	6.10	8.91	4.53	2.25	2.02	2.57	2.44
terrain (9.17%)	41.83	38.86	32.39	38.89	35.96	30.36	42.82	39.61	33.06	41.12	38.02	30.35	31.72	31.59	29.84
pole (0.29%)	11.84	10.43	7.52	9.98	7.41	4.55	20.66	15.56	9.15	7.30	4.99	3.13	3.10	3.79	3.91
traf.-sign (0.08%)	7.39	6.94	5.25	6.04	5.55	4.06	10.63	8.09	4.94	2.35	2.31	1.52	3.69	2.54	2.43
IoU (%)	65.02	57.45	43.80	56.80	50.44	37.76	65.38	57.69	44.15	54.75	46.03	35.62	38.42	38.55	36.80
mIoU (%)	19.68	17.36	13.27	16.21	14.16	10.42	21.55	18.42	13.35	17.15	15.27	11.30	12.25	12.22	11.30
SS/FS (%)	91.32	94.25	99.40	75.22	76.87	78.05	-	-	-	-	-	-	-	-	-

Table 1: Quantitative comparisons against the fully-supervised camera-based SSC methods on the *validation set* of SemanticKITTI. † represents the results that are directly retrained with ScribbleSC. ∗ denotes the results are reported from VoxFormer. SS/FS measures the relative performance ratio of the scribble-supervised (SS) model over the fully-supervised (FS) one. The best results in scribble- and fully-supervised models are marked in **blue** and **bold**, respectively.

Methods	Supervision	IoU (%)	car	bicycle	motorcycle	truck	other-veh.	person	bicyclist	motorcyclist	road	parking	sidewalk	other-grnd	building	fence	vegetation	trunk	terrain	pole	traf.-sign	mIoU (%)
MonoScene [CVPR'22]	Fully	34.16	18.80	0.50	0.70	3.30	4.40	1.00	1.40	0.40	54.70	24.80	27.10	5.70	14.40	11.10	14.90	2.40	19.50	3.30	2.10	11.08
TPVFormer [CVPR'23]	Fully	34.25	19.20	1.00	0.50	3.70	2.30	1.10	2.40	0.30	55.10	27.40	27.20	6.50	14.80	11.00	13.90	2.60	20.40	2.90	1.50	11.26
OccFormer [ICCV'23]	Fully	34.53	21.60	1.50	1.70	1.20	3.20	2.20	1.10	0.20	55.90	31.50	30.30	6.50	15.70	11.90	16.80	3.90	21.30	3.80	3.70	12.32
NDC-Scene [ICCV'23]	Fully	36.19	19.13	1.93	2.07	4.77	6.69	3.44	2.77	1.64	58.12	25.31	28.05	6.53	14.90	12.85	17.94	3.49	25.01	4.43	2.96	12.58
VoxFormer [CVPR'23]	Fully	43.21	21.70	1.90	1.60	3.60	4.10	1.60	1.10	0.00	54.10	25.10	26.90	7.30	23.50	13.10	24.40	8.10	24.20	6.60	5.70	13.41
VoxFormer† [CVPR'23]	Scribble	34.43	17.50	1.90	1.00	1.20	5.60	4.00	1.40	0.00	44.30	18.90	23.10	10.10	17.00	8.00	18.20	7.10	23.10	3.40	4.80	10.87
Scribble2Scene (Ours)	Scribble	42.60	20.10	2.70	1.60	5.60	4.50	1.60	1.80	0.00	50.30	20.60	27.30	11.30	23.70	13.30	23.50	9.60	23.80	5.60	6.50	13.33

Table 2: Quantitative comparisons between Scribble2Scene and the state-of-the-art camera-based methods on the *hidden test set* of SemanticKITTI. † represents the results that are retrained with ScribbleSC.

Implementation Details. For Dean-Labeler, we adopt Cylinder3D [Zhu *et al.*, 2021a] as the *SCN* backbone and use a single GPU to train the network with a batch size of 4. For Teacher-Labeler and student model, we use the same backbone of VoxFormer-T [Li *et al.*, 2023c], which takes the current and previous 4 images as input. All models based on VoxFormer are trained on 4 GPUs with 20 epochs, a batch size of 1 (containing 5 images) per GPU. Our baseline model is directly trained with partial cross-entropy loss and geometric loss under the available scribble- and geometry-based supervisions. For our proposed range-guided distillation scheme, we choose three different ranges, which are the same as the evaluation part to perform the global and local distillation, *i.e.* 12.8m, 25.6m, and 51.2m. More implementation details and model complexity analyses are provided in the Supplementary Material.

5.2 Main Results

We firstly compare Scribble2Scene with state-of-the-art fully-supervised camera-based methods on the *validation set* of SemanticKITTI, including VoxFormer [Li *et al.*, 2023c], TPVFormer [Huang *et al.*, 2023], and MonoScene [Cao and de Charette, 2022]. As shown in Tab. 1, Scribble2Scene obtains 99% performance (13.27% mIoU *v.s.* 13.35% mIoU) of fully-supervised VoxFormer at full-range 51.2m, which only uses 13.5% of the labeled voxels. The competitive accuracy is also achieved against other camera-based models at different ranges. Compared with VoxFormer trained directly using scribble annotations, our method has a significant improvement in the most of categories. To further examine the effectiveness of our method, we submit results on the extremely challenging *test set* of SemanticKITTI without extra tricks. As illustrated in Tab. 2, our scribble-based method achieves 13.33% mIoU and 42.60% IoU, which outperforms most of

Methods	Input	Supervision	IoU (%)	mIoU (%)
SCPNet [CVPR'23]	L	Fully	49.90	**37.20**
S3CNet [CoRL'20]	L	Fully	57.12	33.08
SSC-RS [IROS'23]	L	Fully	58.62	24.75
JS3C-Net [AAAI'21]	L	Fully	53.09	22.67
VoxFormer [CVPR'23]	C	Fully	44.15	**13.35**
TPVFormer [CVPR'23]	C	Fully	35.62	11.30
MonoScene [CVPR'22]	C	Fully	36.80	11.30
Dean-Labeler	G	Scribble	100.00	**42.28**
Teacher-Labeler	C & G	Scribble	82.80	21.70
Scribble2Scene (Ours)	C	Scribble	43.80	13.27

Table 3: Performance of geometric-aware auto-labelers (Stage-I) against the state-of-the-art semantic scene completion models. "L", "C", and "G" denote the LiDAR, camera, and complete geometry as input, respectively.

Baseline	Scribble2Scene			IoU (%)	mIoU (%)
	DL	TL	RGO^{2}D		
✓				37.76	10.42
✓		✓		36.51	10.37
✓		✓	✓	36.86	10.79
	✓			44.19	10.56
	✓	✓		44.51	11.27
	✓	✓	✓	43.80	**13.27**

Table 4: Impact of the each module in overall Scribble2Scene framework.

Methods	IoU (%)	mIoU (%)
Baseline (w/o KD)	44.19	10.56
Vanilla KD	44.51	11.27
MGD [ECCV'22]	43.29	11.26
DIST [NeurIPS'22]	43.11	11.40
CleanerS [CVPR'23]	44.79	11.86
RGO^{2}D (Ours)	43.80	**13.27**

Table 5: Performance comparisons with other knowledge distillation schemes.

Methods	IoU (%)	mIoU (%)
RGO^{2}D *w/o.* **global-distill.**	43.55	11.62
RGO^{2}D *w/o.* **local-distill.**	43.46	12.86
RGO^{2}D *w/o.* **range-info.**	43.56	12.78
RGO^{2}D *w/o.* **feature-distill.**	44.77	12.67
RGO^{2}D	43.80	**13.27**

Table 6: Ablation study for each item in our RGO^{2}D module.

the fully-supervised models and demonstrates generalization capability in more scenarios.

Additional experiments are conducted on SemanticPOSS with our Scribble2Scene framework. Due to limited space, the corresponding experimental table and implementation details are provided in the Supplementary Material.

5.3 Ablation Studies

Effectiveness of Auto-Labelers. Firstly, we verify the effectiveness of scribbles as annotations to train the Dean-Labeler and Teacher-Labeler. As shown in Tab. 3, the promising performance is achieved with our training pipeline compared with SOTA fully-supervised completion methods including LiDAR-based models. Our Dean-Labeler obtains 100% IoU with the highest score of 42.28% mIoU on ScribbleSC, which ensures that the quality of our pseudo-labels is sufficient to provide reliable supervision for the student model. The Teacher-Labeler also achieves a comparable performance with LiDAR-based methods and outperforms the camera-based models by a large margin. This observation reveals that the main bottleneck of current camera-based methods lies in the estimation of geometry.

Impact of Each Module. Secondly, we study the impact of each module in the whole framework as shown in Tab. 4. The Scribble2Scene with Dean-Labeler (DL) means that we replace the scribble annotations with the pseudo labels provided by DL. It can be seen that the models trained with DL obtain obviously higher geometric performance than those

trained with scribbles directly. Moreover, we directly add the Teacher-Labeler (TL) with the Vanilla KD [Hinton *et al.*, 2015], and a slight precision improvement is achieved. Furthermore, we perform range-guided offline-to-online distillation (RGO^{2}D) with TL and the best performance is obtained. **Effect of Offline-to-Online Distillation.** Finally, we compare our RGO^{2}D with other knowledge distillation methods including MGD [Yang *et al.*, 2022], DIST [Huang *et al.*, 2022] for common KD and CleanerS [Wang *et al.*, 2023a] designed for SSC. As shown in Tab. 5, our proposed offline-to-online distillation scheme outperforms all other methods. We further conduct the ablation analysis on each item of our distillation module as described in Tab. 6. Compared to local distillation, the global one generates a larger performance impact. The range-guided information further enhances the performance by computing semantic logit and affinity at different ranges. With feature distillation, the online model obtains the best semantic scene completion accuracy.

6 Conclusion

In this work, we have presented a scribble-based label-efficient benchmark ScribbleSC for semantic scene completion in autonomous driving. To enhance the performance in this setting, an effective scribble-supervised approach Scribble2Scene has been developed. The offline Dean-Labeler provides dense semantic supervision and Teacher-Labeler guides the online model to learn structured occupancy information with new range-guided offline-to-online distillation. Extensive experiments demonstrate that our Scribble2Scene closes the gap between the sparse scribble-based approach and densely annotated methods, which shows competitive performance against the fully-supervised counterparts.

Acknowledgments

ported by Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

References

[Bearman *et al.*, 2016] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016.

[Behley *et al.*, 2019] 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 *ICCV*, 2019.

[Cai *et al.*, 2021] Yingjie Cai, Xuesong Chen, Chao Zhang, Kwan-Yee Lin, Xiaogang Wang, and Hongsheng Li. Semantic scene completion via integrating instances and scene in-the-loop. In *CVPR*, 2021.

[Cao and de Charette, 2022] Anh-Quan Cao and Raoul de Charette. MonoScene: Monocular 3D semantic scene completion. In *CVPR*, 2022.

[Cheng *et al.*, 2021] Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Liu Bingbing. S3CNet: A sparse semantic scene completion network for LiDAR point cloud. In *CoRL*, 2021.

[Chibane *et al.*, 2022] Julian Chibane, Francis Engelmann, Tuan Anh Tran, and Gerard Pons-Moll. Box2Mask: Weakly supervised 3D semantic instance segmentation using bounding boxes. In *ECCV*, 2022.

[Cho and Hariharan, 2019] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *ICCV*, 2019.

[Fan *et al.*, 2022] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Pointly-supervised panoptic segmentation. In *ECCV*, 2022.

[Furlanello *et al.*, 2018] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, 2018.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 2012.

[Heo *et al.*, 2019a] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, 2019.

[Heo *et al.*, 2019b] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *AAAI*, 2019.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Huang *et al.*, 2022] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *NeurIPS*, 2022.

[Huang *et al.*, 2023] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3D semantic occupancy prediction. In *CVPR*, 2023.

[Li *et al.*, 2020] Siqi Li, Changqing Zou, Yipeng Li, Xibin Zhao, and Yue Gao. Attention-based multi-modal fusion network for semantic scene completion. In *AAAI*, 2020.

[Li *et al.*, 2022] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Xian-Sheng Hua, and Lei Zhang. Box-supervised instance segmentation with level set evolution. In *ECCV*, 2022.

[Li *et al.*, 2023a] Wentong Li, Yuqian Yuan, Song Wang, Jianke Zhu, Jianshu Li, Jian Liu, and Lei Zhang. Point2mask: Point-supervised panoptic segmentation via optimal transport. In *ICCV*, 2023.

[Li *et al.*, 2023b] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, et al. SSCBench: A large-scale 3d semantic scene completion benchmark for autonomous driving. *arXiv preprint arXiv:2306.09001*, 2023.

[Li *et al.*, 2023c] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. VoxFormer: Sparse voxel transformer for camera-based 3D semantic scene completion. In *CVPR*, 2023.

[Li *et al.*, 2024a] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Risheng Yu Xiansheng Hua, and Lei Zhang. Box2mask: Box-supervised instance segmentation via level-set evolution. *T-PAMI*, 2024.

[Li *et al.*, 2024b] Wentong Li, Yuqian Yuan, Song Wang, Wenyu Liu, Dongqi Tang, Jianke Zhu, Lei Zhang, et al. Label-efficient segmentation via affinity propagation. In *NeurIPS*, 2024.

[Liang *et al.*, 2022] Zhiyuan Liang, Tiancai Wang, Xiangyu Zhang, Jian Sun, and Jianbing Shen. Tree energy loss: Towards sparsely annotated semantic segmentation. In *CVPR*, 2022.

[Lin *et al.*, 2016] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 2016.

[Liu *et al.*, 2018] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. In *NeurIPS*, 2018.

[Liu *et al.*, 2022] Yufan Liu, Jiajiong Cao, Bing Li, Weiming Hu, and Stephen Maybank. Learning to explore distillability and sparsability: A joint framework for model compression. *T-PAMI*, 2022.

[Liu *et al.*, 2024] Xiaolu Liu, Song Wang, Wentong Li, Ruizi Yang, Junbo Chen, and Jianke Zhu. Mgmap: Mask-guided learning for online vectorized hd map construction. In *CVPR*, 2024.

[Mei *et al.*, 2023] Jianbiao Mei, Yu Yang, Mengmeng Wang, Tianxin Huang, Xuemeng Yang, and Yong Liu. SSC-RS:

Elevate LiDAR semantic scene completion with representation separation and BEV fusion. In *IROS*, 2023.

[Pan *et al.*, 2020] 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 *IV*, 2020.

[Roldao *et al.*, 2020] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. LMSCNet: Lightweight multi-scale 3D semantic completion. In *3DV*, 2020.

[Romero *et al.*, 2015] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *ICLR*, 2015.

[Shi *et al.*, 2024] Hao Shi, Song Wang, Jiaming Zhang, Xiaoting Yin, Zhongdao Wang, Zhijian Zhao, Guangming Wang, Jianke Zhu, Kailun Yang, and Kaiwei Wang. Occfiner: Offboard occupancy refinement with hybrid propagation. *arXiv preprint arXiv:2403.08504*, 2024.

[Sima *et al.*, 2023] Chonghao Sima, Wenwen Tong, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, and Hongyang Li. Scene as occupancy. In *ICCV*, 2023.

[Song *et al.*, 2017] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2017.

[Tian *et al.*, 2021] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. BoxInst: High-performance instance segmentation with box annotations. In *CVPR*, 2021.

[Tian *et al.*, 2023] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yue Wang, Yilun Wang, and Hang Zhao. Occ3D: A large-scale 3D occupancy prediction benchmark for autonomous driving. In *NeurIPS*, 2023.

[Unal *et al.*, 2022] Ozan Unal, Dengxin Dai, and Luc Van Gool. Scribble-supervised LiDAR semantic segmentation. In *CVPR*, 2022.

[Wang *et al.*, 2022] Song Wang, Jianke Zhu, and Ruixiang Zhang. Meta-rangeseg: Lidar sequence semantic segmentation using multiple feature aggregation. *RA-L*, 2022.

[Wang *et al.*, 2023a] Fengyun Wang, Dong Zhang, Hanwang Zhang, Jinhui Tang, and Qianru Sun. Semantic scene completion with cleaner self. In *CVPR*, 2023.

[Wang *et al.*, 2023b] Song Wang, Wentong Li, Wenyu Liu, Xiaolu Liu, and Jianke Zhu. Lidar2map: In defense of lidar-based semantic map construction using online camera distillation. In *CVPR*, 2023.

[Wang *et al.*, 2023c] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. OpenOccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *ICCV*, 2023.

[Wang *et al.*, 2024] Song Wang, Jiawei Yu, Wentong Li, Wenyu Liu, Xiaolu Liu, Junbo Chen, and Jianke Zhu. Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation. In *CVPR*, 2024.

[Wei *et al.*, 2023] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. SurroundOcc: Multi-camera 3D occupancy prediction for autonomous driving. In *ICCV*, 2023.

[Wilson *et al.*, 2022] 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. *RA-L*, 2022.

[Xia *et al.*, 2023] Zhaoyang Xia, Youquan Liu, Xin Li, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, and Yu Qiao. SCPNet: Semantic scene completion on point cloud. In *CVPR*, 2023.

[Yan *et al.*, 2021] 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 *AAAI*, 2021.

[Yang *et al.*, 2022] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In *ECCV*, 2022.

[Yao *et al.*, 2023] Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. NDC-Scene: Boost monocular 3D semantic scene completion in normalized device coordinates space. In *ICCV*, 2023.

[Ye and Bors, 2022] Fei Ye and Adrian G Bors. Dynamic self-supervised teacher-student network learning. *T-PAMI*, 2022.

[Ye and Bors, 2023] Fei Ye and Adrian G Bors. Continual variational autoencoder via continual generative knowledge distillation. In *AAAI*, 2023.

[Zhang *et al.*, 2018] Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient semantic scene completion network with spatial group convolution. In *ECCV*, 2018.

[Zhang *et al.*, 2023] Yunpeng Zhang, Zheng Zhu, and Dalong Du. OccFormer: Dual-path transformer for vision-based 3D semantic occupancy prediction. In *ICCV*, 2023.

[Zhao *et al.*, 2022] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, 2022.

[Zhu *et al.*, 2021a] 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 *CVPR*, 2021.

[Zhu *et al.*, 2021b] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2021.