VLScene: Vision-Language Guidance Distillation for Camera-Based 3D Semantic Scene Completion

Meng Wang, Huilong Pi*, Ruihui Li, Yunchuan Qin, Zhuo Tang*, Kenli Li

College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
{willem, liruihui, phl880217, qinyunchuan, ztang, lkl}@hnu.edu.cn

Abstract

Camera-based 3D semantic scene completion (SSC) provides dense geometric and semantic perception for autonomous driving. However, images provide limited information making the model susceptible to geometric ambiguity caused by occlusion and perspective distortion. Existing methods often lack explicit semantic modeling between objects, limiting their perception of 3D semantic context. To address these challenges, we propose a novel method VLScene: Vision-Language Guidance Distillation for Camera-based 3D Semantic Scene Completion. The key insight is to use the vision-language model to introduce high-level semantic priors to provide the object spatial context required for 3D scene understanding. Specifically, we design a vision-language guidance distillation process to enhance image features, which can effectively capture semantic knowledge from the surrounding environment and improve spatial context reasoning. In addition, we introduce a geometric-semantic sparse awareness mechanism to propagate geometric structures in the neighborhood and enhance semantic information through contextual sparse interactions. Experimental results demonstrate that VLScene achieves rank-1st performance on challenging benchmarks—SemanticKITTI and SSCBench-KITTI-360, yielding remarkably mIoU scores of 17.52 and 19.10, respectively.

Code — https://github.com/willemeng/VLScene

Introduction

The field of 3D perception faces new challenges with the emergence of autonomous driving. Autonomous vehicles need to predict the surrounding environment accurately to ensure safe navigation and obstacle avoidance. The Semantic Scene Completion (SSC) (Song et al. 2017; Roldao, de Charette, and Verroust-Blondet 2020; Yan et al. 2021) task aims to forecast the semantic occupancy of every voxel in the entire 3D scene based on limited observations.

Most existing SSC (Rist et al. 2021; Zhang et al. 2018; Guo and Tong 2018; Li et al. 2020; Roldao, de Charette, and Verroust-Blondet 2020; Yan et al. 2021) solutions rely on input RGB images and corresponding 3D data to predict volume occupancy and semantic labels. However, the reliance on 3D data often necessitates the use of specialized and expensive depth sensors, which limits the broader application of SSC algorithms. Recently, many researchers (Cao and de Charette 2022; Li et al. 2023c,a; Jiang et al. 2024; Wang and Tong 2024; Xue et al. 2024) have explored the use of camera-based approaches to recover dense 3D geometric structures and semantic information.

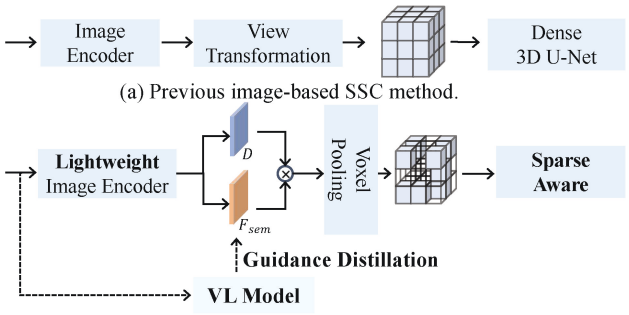


Figure 1: Our method uses vision-language model guidance distillation versus the previous image-based method.

In Figure 1, existing methods commonly depend on 2D-3D view transformations to construct 3D representations from image features, and utilise complex 3D models for geometric and semantic inference. Nevertheless, the limited information available from images renders these models susceptible to geometric ambiguities caused by occlusions and perspective distortions. Inferring geometry without sufficient visual input requires leveraging semantic knowledge of the surrounding environment and reasoning about the spatial context. Given these limitations, we consider the question: ***How can high-level semantic priors be utilized to improve semantic representation and spatial context?***

In this paper, we propose a novel camera-based SSC method: VLScene, Vision-Language Guidance Distillation for Camera-based 3D Semantic Scene Completion. VLScene leverages a vision-language model to extract high-level semantic priors, enhancing semantic representation and spatial context through distillation. As shown in the yellow box in Figure 2(a), semantic and spatial structure priors are obtained through vision-language models. The position and category of the person behind the car are ac-

*Corresponding authors.

curately inferred, and even densely packed cars in the image are effectively distinguished in the corresponding scene. Specifically, we introduce *vision-language guidance distillation* to improve image features, which effectively captures the semantic knowledge of the surrounding environment and enhances reasoning about the spatial context. We further design a *geometric-semantic sparse awareness* mechanism, consisting of two modules: neighborhood geometry propagation (NGP) and sparse semantic interaction (SSI), to sparsely perceive voxel information from both geometric and semantic perspectives. The NGP module alternates between large and small kernel convolutions to ensure comprehensive capture of objects of varying sizes in the adapted 3D scene. Meanwhile, SSI employs sparse convolutions to effectively utilize the geometric information in voxel features and enhances semantic information through contextual interactions. To evaluate the performance of VLScene, we conduct thorough experiments on SemanticKITTI (Behley et al. 2019) and SSCBench-KITTI360 (Liao, Xie, and Geiger 2022; Li et al. 2023b). As shown in Figure 2(b), our method achieves state-of-the-art performance with a mIoU of 17.52%, using only 47.4M parameters.

- We propose a novel method, VLScene, leveraging vision-language models to extract high-level semantic priors, thereby enhancing semantic representation and spatial context through a distillation process.
- We design vision-language guidance distillation that effectively captures semantic knowledge of the surrounding environment and improves spatial context reasoning.
- We introduce geometric-semantic sparse awareness to propagate geometric structure in the neighborhood and enhance semantic information through contextual sparse interactions.
- The proposed VLScene model achieves SOTA results on SemanticKITTI and SSCBench-KITTI-360 benchmarks, surpassing the latest methods.

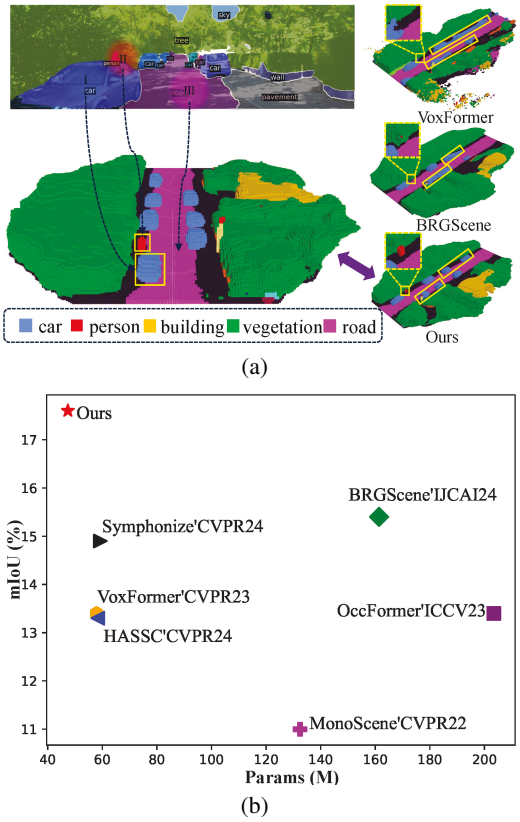


Figure 2: (a) VLScene uses 2D high-level semantic priors improve 3D scene. (b) Comparison of Params and mIoU.

Related Work

3D Semantic Scene Completion

The SSC task was designed to address the challenges of scene completion and semantic segmentation by predicting the occupancy and semantic categories of each voxel in a 3D scene. SSCNet (Song et al. 2017) was the first to define the semantic scene completion task, wherein both geometry and semantics were jointly inferred from incomplete visual observations. Subsequent research recognized the inherently 3D nature of this task, leading to numerous studies that directly employed 3D inputs (Rist et al. 2021; Zhang et al. 2018; Xia et al. 2023), such as depth information, occupancy meshes, and point clouds, to leverage their rich geometric cues. To incorporate additional texture information, many works (Cai et al. 2021; Li et al. 2019) explored the use of multi-modal inputs, combining RGB images with diverse geometric cues. As purely visual autonomous driving solutions became more cost-effective, MonoScene (Cao and de Charette 2022) was the first to infer dense geometry and semantics from a single monocular RGB image. TPVFormer (Huang et al. 2023) proposed a three-perspective view representation, which, along with BEV, introduced two additional vertical planes. OccFormer (Zhang, Zhu, and Du 2023) employed a dual-path transformer framework to encode 3D voxel features. VoxFormer (Li et al. 2023c) adopted a novel two-stage framework to elevate images into fully 3D voxelized semantic scenes. NDCScene (Yao et al. 2023) extended the 2D feature map to a normalized device coordinate space instead of the world space. MonoOcc (Zheng et al. 2024) further enhanced the 3D volume with an image-conditioned cross-attention module. H2GFormer (Wang and Tong 2024) effectively utilized 2D features through a progressive feature reconstruction process across various directions. Symphonize (Jiang et al. 2024) extracted high-level instance features from the image feature map, serving as the key and value for cross-attention. HASSC (Wang et al. 2024) introduced a self-distillation training strategy to improve the performance of VoxFormer. Finally, BRGScene (Li et al. 2023a) utilized binocular image inputs to implicitly generate stereo depth information and employed stereo matching to resolve geometric ambiguities.

Large Vision-Language Model

Large vision-language models, such as Contrastive Vision-Language Pre-training (CLIP) (Radford et al. 2021),

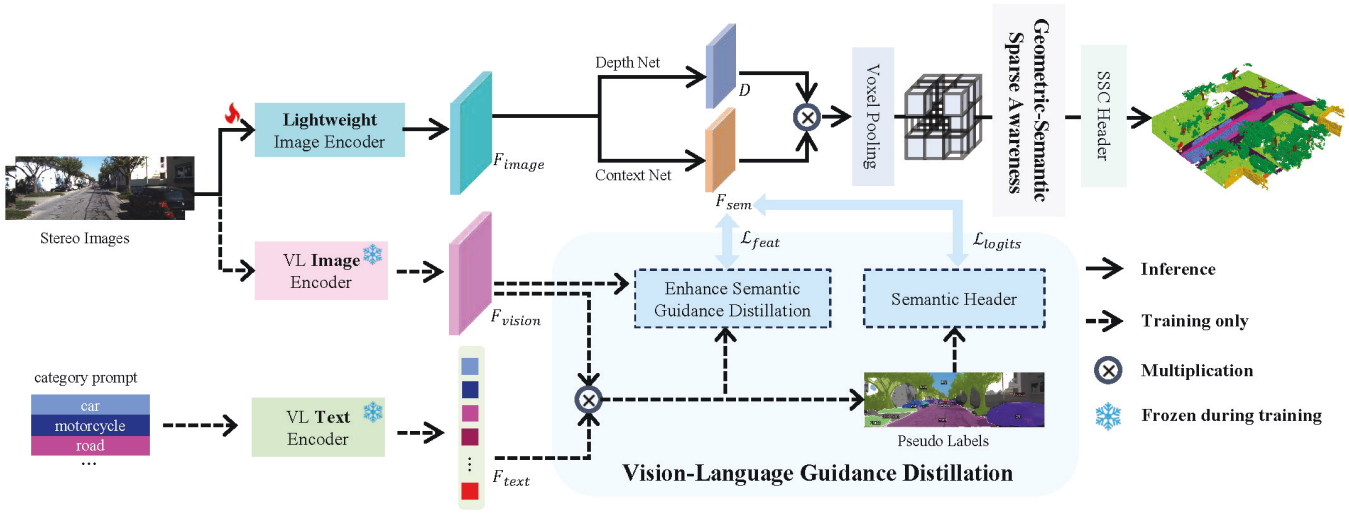


Figure 3: The VLScene framework is proposed for camera-based 3D semantic scene completion.

achieved significant advancements in open-set and zero-shot image classification tasks due to their strong alignment between visual and text embeddings. LSeg (Li et al. 2022) focused on aligning pixel-level image features with class embeddings generated by the CLIP text encoder. MaskCLIP (Dong et al. 2023) explored dense prediction problems using CLIP by making minor adjustments to the network structure. However, practical applications like autonomous driving and indoor navigation required a deeper understanding of 3D scenes. Consequently, recent research investigated the application of 2D vision-language pre-training for 3D perception tasks. For instance, CLIP2Scene (Chen et al. 2023) introduced a semantically driven cross-modal contrastive learning framework. OpenScene (Peng et al. 2023) used a pre-trained VL model (Ghiasi et al. 2022; Kuo et al. 2023) to extract per-pixel CLIP features and projected 3D points onto the image plane to derive dense 3D features. RegionPLC (Yang et al. 2023) leveraged regional visual cues to generate dense captions and performed point-discriminative contrastive learning.

Cross-modal Knowledge Distillation

The primary objective of cross-modal knowledge distillation was to transfer knowledge between different modalities. Knowledge distillation (Hinton, Vinyals, and Dean 2015; Wang et al. 2022; Mirzadeh et al. 2020) was initially introduced for model compression, focusing on transferring the learned knowledge from a teacher network to a student network. With advancements in multi-sensor technologies, 2D-to-3D distillation methods were developed to enable models to utilize data from various modalities, thereby enhancing their performance in 3D tasks. For example, PPKT (Liu et al. 2021) employed InfoNCE loss to assist 3D networks in extracting valuable knowledge from 2D image backbones. 2DPass (Yan et al. 2022) proposed an innovative approach to improve semantic information extraction from multimodal data by integrating auxiliary modality fusion and multi-scale fusion into a single distillation framework. CMKD (Hong, Dai, and Ding 2022) effectively transferred point cloud features and responses to images, significantly boosting performance. BEVDistill (Chen et al. 2022) unified the two modalities into BEV space for distillation, achieving more accurate object detection by using a grouping network to form a global descriptor. UniDistill (Zhou et al. 2023) focused on BEV object detection and leveraged knowledge distillation in features, relations, and responses.

Methodology

Problem setup. Given a set of stereo RGB images $I_{stereo} = \{I_l, I_r\}$. The goal is to infer the geometry and semantics of the 3D scene jointly. The scene is represented as a voxel grid $Y \in \mathbb{R}^{X \times Y \times Z \times (M+1)}$, where X, Y, Z represent height, width and depth in 3D space. For each voxel, it will be assigned to a unique semantic label belonging to $C \in \{C_0, C_1, ..., C_M\}$ that either occupies the empty space C_0 or falls in a specific semantic class $\{C_1, ..., C_M\}$. Here M represents the total number of semantic classes. We want to learn a transformation $Y = \theta(I_{stereo})$ as close to the ground truth $\hat{Y}$ as possible.

Overview. We illustrate our approach in Figure 3. We first use the lightweight image encoder RepViT (Wang et al. 2023) and FPN (Lin et al. 2017) to extract image features F_{image}, and use the vision-language model LSeg (Li et al. 2022) to extract visual features F_{vision} and text features F_{text}. F_{image} passes through the deep network and the context network to obtain discrete depth values D and semantic features F_{sem}. Subsequently, we use the VLGD module to improve the semantic features with the high-level semantic information captured by the VL model, and then obtain the voxel features V through the LSS view transformation. Then, V enters the GSSA module to sparsely perceive the voxel information from the geometric and semantic perspectives, and obtains the refined voxel features V_{fine}. Finally, V_{fine} outputs dense semantic voxels Y through upsampling and linear projection.

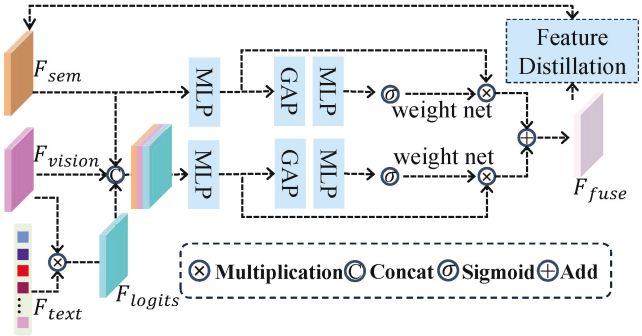


Figure 4: Illustration of the enhanced semantic feature distillation.

Vision-Language Guidance Distillation

Existing camera-based SSC methods often lack explicit semantic modeling and are susceptible to geometric ambiguities caused by occlusion and perspective distortion. To enrich semantic representations, we introduce a vision-language guidance distillation module in 2D space. This module leverages the VL model to extract high-level semantic priors, enhancing both semantic representations and spatial context through distillation. Details are shown in Figure 3.

Feature Distillation. Given an input image I_l, we extract image features F_{image} using a lightweight image encoder and visual features F_{vision} using a VL image encoder. Simultaneously, we input the category name as a prompt into the VL text encoder to obtain text features $F_{text} \in \mathbb{R}^{Q\times C}$, where Q represents the number of categories and C denotes the feature channels. We then compute the cosine similarity between the VL image and text features to generate a 2D semantic map:

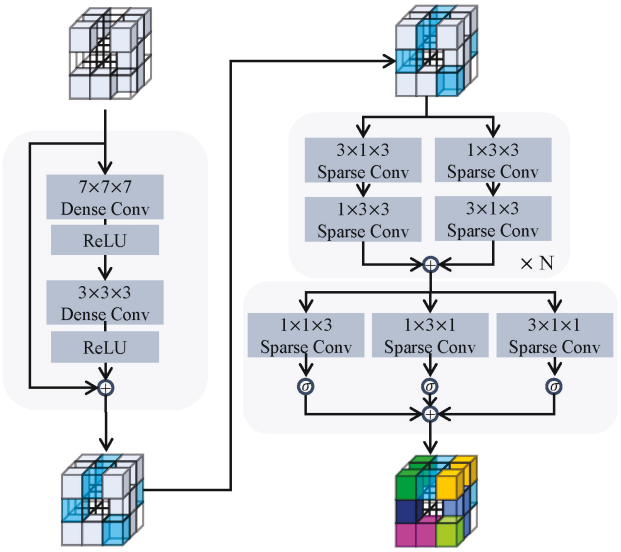
$$F_{logits} = \underset{q\in\{1,\dots,Q\}}{\text{softmax}} \frac{F_{vision} \otimes F_{text}^{\top}}{\|F_{vision}\| \, \|F_{text}\|}, \tag{1}$$

where $\otimes$ denotes matrix multiplication, and F_{logits} denotes the segmentation map by maximizing the similarity scores between VL image and text features along the category dimension.

To effectively select rich semantic cues from the VL model, we adaptively fuse the semantic information from both the image encoder and the VL model to obtain enhanced fused semantic features, which are then subjected to feature distillation. Figure 4 illustrates the detailed process of this enhanced semantic feature distillation. First, we concatenate the three features F_{sem}, F_{vision}, and F_{logits} and perform an initial fusion using two convolutional layers to obtain the feature $\hat{F}_{vision}$. Next, we calculate the channel attention for $\hat{F}_{vision}$ and F_{sem} to adaptively weight the feature channels,

$$\begin{aligned} \hat{F}_{vision}^{weight} &= \sigma[\text{GAP}(\text{MLP}(\hat{F}_{vision}))], \\ F_{sem}^{weight} &= \sigma[\text{GAP}(\text{MLP}(F_{sem}))], \end{aligned} \tag{2}$$

where σ denotes the sigmoid function, GAP represents the global average pooling operation. The weighted features are



(a) Neighborhood Geometry Propagation (b) Sparse Semantic Interaction

Figure 5: Illustration of the geometric-semantic sparse awareness.

then summed to obtain the fused feature F_{fuse},

$$\begin{aligned} F_{fuse} = \hat{F}_{vision}^{weight} * \text{MLP}(\hat{F}_{vision}) \\ + F_{sem}^{weight} * \text{MLP}(F_{sem}). \end{aligned} \tag{3}$$

For each pixel feature, we calculate the difference between its features on F_{sem} and F_{fuse} the feature distillation loss $\mathcal{L}_{kd\_feat}$:

$$\mathcal{L}_{kd\_feat} = \|F_{sem} - F_{fuse}\|_1. \tag{4}$$

Logits Distillation. We also perform explicit logits distillation on the semantic features. Specifically, we pass F_{sem} through the image semantic head composed of residual blocks to obtain the image segmentation result F_{pred} of category Q. Then calculate the loss with the image semantic pseudo label F_{logits} obtained by the vision-language model. The specific operation is as follows:

$$\mathcal{L}_{kd\_logits} = \text{CrossEntropyLoss}(F_{pred}, F_{logits}), \tag{5}$$

After our designed vision-language guidance distillation, we will get the enhanced semantic feature $\hat{F}_{sem}$. Then, we follow the view transformation module of BRGScene (Li et al. 2023a) and construct the voxel feature V using $\hat{F}_{sem}$.

Geometric-Semantic Sparse Awareness

It is observed that even after projecting the enhanced semantic features into voxel space, approximately half of the voxels remain empty (especially out-of-field, obstructed and distant areas). To fully utilize the available voxel features, we introduce the geometric-semantic sparse awareness module, which can perceive voxel information both geometrically and semantically. The detailed process of geometric-semantic sparse awareness is illustrated in Figure 5.

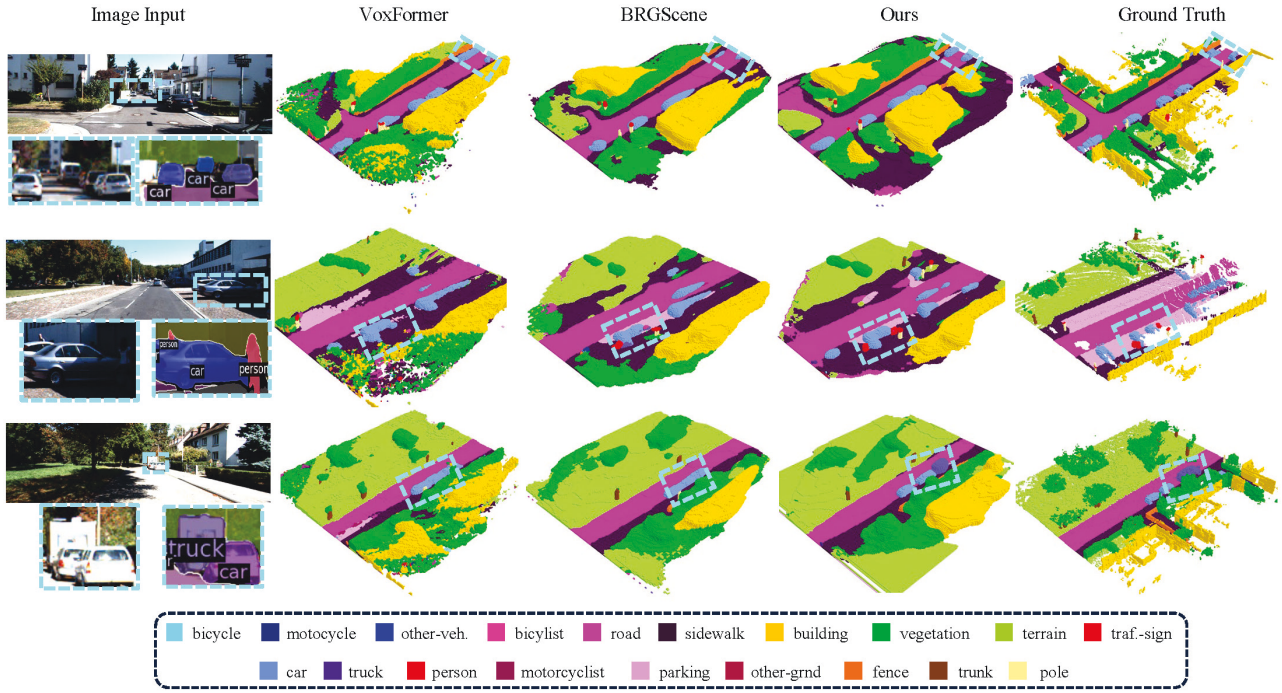


Figure 6: Qualitative results on the SemanticKITTI validation set.

Neighborhood Geometry Propagation. Geometric information is typically extracted only from existing non-empty voxels, leading to a lack of features in empty voxels. To address this, we design the Neighborhood Geometry Propagation (NGP) module, which diffuses voxel features into adjacent empty regions. The detailed process is shown in Figure 5(a).

Specifically, NGP consists of alternating $7 \times 7 \times 7$ large kernel convolution layers and $3 \times 3 \times 3$ small kernel convolution layers. The large kernel convolutions handle large objects (such as cars and trucks) and view boundaries, propagating features from non-empty voxels to surrounding empty ones, thereby capturing the complete geometric structure of larger objects. Conversely, small kernel convolutions target smaller objects, such as poles and pedestrians, providing finer granularity in feature propagation to accurately capture the complex details of these smaller objects. Additionally, inspired by SCPNet (Xia et al. 2023), we remove the convolution bias and batch normalization layers to reduce computational costs and maintain the efficiency of subsequent sparse convolutions. The specific calculation process is as follows:

$$V_{com} = V + \delta\left(\text{Conv}_{3\times3\times3}\left(\delta\left(\text{Conv}_{7\times7\times7}(V)\right)\right)\right), \quad (6)$$

where δ denotes the ReLU activation function. Through the above operation, we obtain the completion voxel feature.

Sparse Semantic Interaction. To enrich the semantic information of the scene, we introduce a Sparse Semantic Interaction (SSI) module. This module effectively leverages the geometric information within voxel features and enhances semantic content through contextual interactions.

Initially, the completion feature V_{com} is converted into a sparse representation by isolating non-empty voxels. The sparse tensor is stored in a commonly used coordinate format:

$$V_{com} = \{\mathrm{P} = [x, y, z], \mathrm{F} \in \mathbb{R}^{N\times C}\}, \quad (7)$$

where N represents the number of non-empty voxels, and P and F denote the coordinates and features of the voxels.

Subsequently, inspired by the observations and conclusions in Cylinder3D (Zhu et al. 2021), we decompose the traditional sparse convolution kernel into two orthogonal kernels oriented in vertical and horizontal directions. This decomposition aligns with the voxel distribution of objects in the driving scene, enabling the orthogonal kernels to capture specific semantic information along both directions. By configuring two parallel and asymmetric convolution kernels, we further enhance the semantic richness of the scene. Finally, following Cylinder3D, we apply three Rank-1 sparse kernels to extract low-rank features, which are then aggregated to produce refined voxel features, V_{fine}.

Training Loss

In the VLScene framework, we adopt the scene-class affinity loss from MonoScene (Cao and de Charette 2022) to optimize precision, recall, and specificity concurrently. The semantic scene completion loss is shown as follows:

$$\mathcal{L}_{ssc} = \mathcal{L}_{scal}^{sem} + \mathcal{L}_{scal}^{geo} + \mathcal{L}_{ce} + \mathcal{L}_{depth}, \quad (8)$$

The knowledge distillation loss is expressed as:

$$\mathcal{L}_{kd} = \mathcal{L}_{kd\_feat} + \mathcal{L}_{kd\_logits}, \quad (9)$$

The overall training loss function is formulated as follows:

$$\mathcal{L} = \lambda_{ssc}\mathcal{L}_{ssc} + \lambda_{kd}\mathcal{L}_{kd}, \quad (10)$$

Methods	IoU	road	sidewalk	parking	other-grnd	building	car	truck	bicycle	motocycle	other-vehicle	vegetation	trunk	terrain	person	bicylist	motorcyclist	fence	pole	traf.-sign	mIoU
MonoScene	34.16	54.7	27.1	24.8	5.7	14.4	18.8	3.3	0.5	0.7	4.4	14.9	2.4	19.5	1.0	1.4	0.4	11.1	3.3	2.1	11.08
TPVFormer	34.25	55.1	27.2	27.4	6.5	14.8	19.2	3.7	1.0	0.5	2.3	13.9	2.6	20.4	1.1	2.4	0.3	11.0	2.9	1.5	11.26
OccFormer	34.53	55.9	30.3	31.5	6.5	15.7	21.6	1.2	1.5	1.7	3.2	16.8	3.9	21.3	2.2	1.1	0.2	11.9	3.8	3.7	12.32
VoxFormer	43.21	54.1	26.9	25.1	7.3	23.5	21.7	3.6	1.9	1.6	4.1	24.4	8.1	24.2	1.6	1.1	0.0	13.1	6.6	5.7	13.41
MonoOcc	-	55.2	27.8	25.1	9.7	21.4	23.2	5.2	2.2	1.5	5.4	24.0	8.7	23.0	1.7	2.0	0.2	13.4	5.8	6.4	13.80
H2GFormer	44.20	56.4	28.6	26.5	4.9	22.8	23.4	4.8	0.8	0.9	4.1	24.6	9.1	23.8	1.2	2.5	0.1	13.3	6.4	6.3	13.72
HASSC	43.40	54.6	27.7	23.8	6.2	21.1	22.8	4.7	1.6	1.0	3.9	23.8	8.5	23.3	1.6	4.0	0.3	13.1	5.8	5.5	13.34
Symphonize	42.19	58.4	29.3	26.9	11.7	24.7	23.6	3.2	3.6	2.6	5.6	24.2	**10.0**	23.1	**3.2**	1.9	**2.0**	16.1	7.7	8.0	15.04
BRGScene	43.34	61.9	31.2	30.7	10.7	24.2	22.8	2.8	3.4	2.4	6.1	23.8	8.4	27.0	2.9	2.2	0.5	16.5	7.0	7.2	15.36
Ours	**45.14**	**64.7**	**34.7**	**32.4**	**13.1**	**27.3**	**26.1**	**6.5**	**4.2**	**3.8**	**8.3**	**26.4**	**10.0**	**29.4**	2.8	**5.1**	0.9	**20.0**	**8.9**	**8.4**	**17.52**

Table 1: Quantitative results on the SemanticKITTI hidden test set. **Bold** denotes the best performance.

Methods	IoU	car	bicycle	motocycle	truck	other-vehicle	person	road	parking	sidewalk	other-grnd	building	fence	vegetation	terrain	pole	traf.-sign	other-struct.	other-obj.	mIoU
MonoScene	37.87	19.3	0.4	0.6	8.0	2.0	0.9	48.4	11.4	28.1	3.3	32.9	3.5	26.2	16.8	6.9	5.7	4.2	3.1	12.31
VoxFormer	38.76	17.8	1.2	0.9	4.6	2.1	1.6	47.0	9.7	27.2	2.9	31.2	5.0	29.0	14.7	6.5	6.9	3.8	2.4	11.91
TPVFormer	40.22	21.6	1.1	1.4	8.1	2.6	2.4	53.0	12.0	31.1	3.8	34.8	4.8	30.1	17.5	7.5	5.9	5.5	2.7	13.64
OccFormer	40.27	22.6	0.7	0.3	9.9	3.8	2.8	54.3	13.4	31.5	3.6	36.4	4.8	31.0	19.5	7.8	8.5	7.0	4.6	13.81
Symphonies	44.12	**30.0**	1.9	5.9	**25.1**	**12.1**	**8.2**	54.9	13.8	32.8	**6.9**	35.1	8.6	**38.3**	11.5	14.0	9.6	**14.4**	**11.3**	18.58
Ours	**46.08**	29.0	**4.7**	**7.7**	18.3	7.6	7.4	**60.1**	**17.4**	**39.0**	6.0	**42.1**	**9.6**	36.5	**24.8**	**17.0**	**18.8**	10.5	6.5	**19.10**

Table 2: Quantitative results on the SSCBench-KITTI360 test set. **Blod** denotes the best performance.

where several λ are balancing coefficients.

Experiments

To assess the effectiveness of our VLScene, we conducted thorough experiments using the large outdoor datasets SemanticKITTI (Behley et al. 2019) and SSCBench-KITTI-360 (Li et al. 2023b).

Qualitative Results

To intuitively demonstrate VLScene performance, Figure 6 presents the qualitative results of VoxFormer, BRGScene, and our method on the SemanticKITTI validation set. The enlarged area in the first column shows the semantic prior of the scene object structure. Compared to VoxFormer and BRGScene, our method more completely and accurately reconstructs distant objects, such as the car in the blue box in the first row. Additionally, our VLScene effectively captures the position and details of small objects, like the person in the blue box in the second row. It demonstrates significant advantages in delineating the complete outlines and boundaries of large objects, such as the car and truck in the third row.

Quantitative Results

Table 1 presents a comparison of our VLScene with other state-of-the-art camera-based SSC methods on the SemanticKITTI test sets. Our VLScene outperforms existing methods, achieving state-of-the-art results. Compared to BRGScene, VLScene demonstrates an improvement of 2.16% in mIoU and 1.8% in IoU, with significant gains across all semantic categories. Additionally, compared to Symphonize, which extracts and fuses high-level instance features, VLScene achieves a lead of 2.95% in IoU and 2.48% in mIoU. These results validate the effectiveness of our approach in both geometric and semantic aspects, and VLScene achieves the highest mIoU in nearly all categories.

As shown in Table 2, VLScene also exhibits a significant advantage in semantic and geometric analysis over current camera-based approaches on the rich data samples SSCBench-KITTI-360 benchmark, surpassing all published methods in both IoU and mIoU metrics.

Furthermore, Table 3 shows that we provide different ranges of results on the SemanticKITTI validation set. It can be seen that our method significantly surpasses the existing methods at all three distances.

In addition, as shown in Table 4, we compare the inference time and number of parameters with other SOTA methods on the SemanticKITTI validation set. Our method achieves SOTA performance with 17.83% mIoU using only 47.4M parameters. VLScene also demonstrates superior inference time while being more lightweight.

Methods	Venues	mIoU		
		12.8m	25.6m	51.2m
SSCNet	CVPR'17	16.32	14.55	10.27
LMSCNet	3DV'20	15.69	14.13	9.94
MonoScene	CVPR'22	12.25	12.22	11.30
VoxFormer	CVPR'23	17.66	16.48	12.35
OccFormer	ICCV'23	20.91	17.90	13.46
HASSC	CVPR'24	18.98	17.95	13.48
H2GFormer	AAAI'24	20.49	18.39	13.73
BRGScene	IJCAI'24	23.27	21.15	15.24
Ours		**26.51**	**24.37**	**17.83**

Table 3: Comparison of different ranges on SemanticKITTI val set.

Method	mIoU (%)↑	Times (s)↓	Params (M)↓
MonoScene	11.08	0.274	132.4
OccFormer	13.46	0.338	203.4
VoxFormer	13.35	0.256	57.9
Symphonize	14.89	0.319	59.3
BRGScene	15.43	0.285	161.4
Ours	**17.83**	**0.233**	**47.4**

Table 4: Comparison of inference time and number of parameters.

Ablation Studies

Ablation on the Architectural Components. The Table 5 shows a breakdown analysis of various architectural components in the VLScene, including the lightweight image encoder RepViT, vision-language guidance distillation (VLGD), neighborhood geometric propagation (NGP), and sparse semantic interaction (SSI). We will analyze the contents of the Table 5 line by line. First, we adopted BRGScene (Li et al. 2023a) as our baseline. (1) When we replaced the baseline image encoder with RepViT, the model parameters were reduced by 41.4M. (2) Significantly improved mIoU by 0.99% when equipped with our VLGD, proving that VLGD injects rich semantic information through knowledge distillation. (3) The introduction of NGP can effectively improve its geometric completion performance, and the IoU has achieved a 0.56% improvement, while only 35.6M parameters are required. (4) The SSI module enhanced semantic information through contextual interaction, and the mIoU increased by 0.63%. (5) The combination of NGP and SSI improves both geometric and semantic performance. (6) The final full model uses only 47.4M parameters, achieving 1.15% IoU and 2.4% mIoU improvements over baseline, proving that all of these components contribute to the best results.

Ablation Study for VLGD. To further explore the role of VLGD, we examine the individual effects of feature distillation and logical distillation, as shown in Table 6. Our findings indicate that visual language knowledge distillation significantly enhances the model's semantic representation learning. Both distillation methods result in improvements in IoU and mIoU, with VLGD contributing to a 1.15% increase in IoU and a 2.4% rise in mIoU compared to the baseline (Li et al. 2023a). Notably, our method introduces no additional time or computational overhead during inference.

Method	RepViT	VLGD	NGP	SSI	IoU	mIoU	Params
Baseline					43.54	15.43	161.4M
1	✓				43.13	15.59	120.0M
2	✓	✓			44.02	16.58	120.0M
3	✓	✓	✓		44.58	16.43	**35.6M**
4	✓	✓		✓	44.03	17.21	45.9M
5	✓		✓	✓	44.15	16.66	47.4M
6	✓	✓	✓	✓	**44.69**	**17.83**	47.4M

Table 5: Ablation study for Architecture.

Setting	IoU	mIoU
Baseline	43.54	15.43
Logits Distillation	43.75	16.32
Feature Distillation	43.88	16.54
VLGD (Ours)	**44.69**	**17.83**

Table 6: Ablation study for Vision-Language Guidance Distillation.

Setting	IoU	mIoU	Params
3D Swin	44.18	15.87	82.3M
3D ResNet	44.03	15.66	37.4M
GSSA (Ours)	**44.69**	**17.83**	**13.3M**

Table 7: Ablation study for Geometric-Semantic Sparse Aware.

Ablation Study for GSSA. As shown in Table 7, We compare GSSA with other baseline methods. The fixed size of the convolutional kernel in 3D ResNet (He et al. 2016) limits its ability to capture information over a broader region. Similarly, the restricted range of local windows in 3D Swin (Liu et al. 2022) challenges its capacity to effectively capture telematics through self-attention, with sliding window operations only impacting a small neighborhood around each window. Moreover, these methods require more parameters and computational resources. Our GSSA achieves significant improvements while utilizing only 13.3M parameters.

Conclusion

In this paper, we propose a novel method, VLScene: Vision-Language Guidance Distillation for Camera-based 3D Semantic Scene Completion. Specifically, we design a vision-language guidance distillation process that effectively captures semantic knowledge from the surrounding environment and improves spatial context reasoning. In addition, we introduce a geometric-semantic sparse awareness mechanism to propagate geometric structures in the neighborhood and enhance semantic information through contextual sparse interactions. Experimental results demonstrate that VLScene achieves SOTA performance on the SemanticKITTI and SSCBench-KITTI-360 datasets.

Acknowledgments

The work is supported by the Science and Technology Innovation 2030 - New Generation Artificial Intelligence Major Project (2021ZD40300), the National Natural Science Foundation of China (Grant Nos.62225205, 62202151, 52203288, 62473137, 62321003), the China Association for Science and Technology Young Talent Support Project Doctoral Special Program, and the Postgraduate Scientific Research Innovation Project of Hunan Province (CX20240427).

References

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, 9297–9307.

Cai, Y.; Chen, X.; Zhang, C.; Lin, K.-Y.; Wang, X.; and Li, H. 2021. Semantic scene completion via integrating instances and scene in-the-loop. In *CVPR*, 324–333.

Cao, A.-Q.; and de Charette, R. 2022. MonoScene: Monocular 3D Semantic Scene Completion. In *CVPR*.

Chen, R.; Liu, Y.; Kong, L.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; Qiao, Y.; and Wang, W. 2023. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7020–7030.

Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; and Zhao, F. 2022. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. *arXiv preprint arXiv:2211.09386*.

Dong, X.; Bao, J.; Zheng, Y.; Zhang, T.; Chen, D.; Yang, H.; Zeng, M.; Zhang, W.; Yuan, L.; Chen, D.; et al. 2023. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10995–11005.

Ghiasi, G.; Gu, X.; Cui, Y.; and Lin, T.-Y. 2022. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, 540–557. Springer.

Guo, Y.; and Tong, X. 2018. View-volume network for semantic scene completion from a single depth image. In *IJCAI*, 726–732.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hong, Y.; Dai, H.; and Ding, Y. 2022. Cross-modality knowledge distillation network for monocular 3d object detection. In *European Conference on Computer Vision*, 87–104. Springer.

Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2023. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, 9223–9232.

Jiang, H.; Cheng, T.; Gao, N.; Zhang, H.; Lin, T.; Liu, W.; and Wang, X. 2024. Symphonize 3d semantic scene completion with contextual instance queries. In *CVPR*, 20258–20267.

Kuo, W.; Cui, Y.; Gu, X.; Piergiovanni, A.; and Angelova, A. 2023. Open-Vocabulary Object Detection upon Frozen Vision and Language Models. In *The Eleventh International Conference on Learning Representations*.

Li, B.; Sun, Y.; Jin, X.; Zeng, W.; Zhu, Z.; Wang, X.; Zhang, Y.; Okae, J.; Xiao, H.; and Du, D. 2023a. Stereoscene: Bev-assisted stereo matching empowers 3d semantic scene completion. *arXiv preprint arXiv:2303.13959*.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*.

Li, J.; Han, K.; Wang, P.; Liu, Y.; and Yuan, X. 2020. Anisotropic convolutional networks for 3d semantic scene completion. In *CVPR*, 3351–3359.

Li, J.; Liu, Y.; Gong, D.; Shi, Q.; Yuan, X.; Zhao, C.; and Reid, I. 2019. Rgbd based dimensional decomposition residual network for 3d semantic scene completion. In *CVPR*, 7693–7702.

Li, Y.; Li, S.; Liu, X.; Gong, M.; Li, K.; Chen, N.; Wang, Z.; Li, Z.; Jiang, T.; Yu, F.; Wang, Y.; Zhao, H.; Yu, Z.; and Feng, C. 2023b. SSCBench: Monocular 3D Semantic Scene Completion Benchmark in Street Views. *arXiv preprint arXiv:2306.09001*.

Li, Y.; Yu, Z.; Choy, C.; Xiao, C.; Alvarez, J. M.; Fidler, S.; Feng, C.; and Anandkumar, A. 2023c. VoxFormer: Sparse Voxel Transformer for Camera-based 3D Semantic Scene Completion. In *CVPR*.

Liao, Y.; Xie, J.; and Geiger, A. 2022. KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. *TPAMI*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *CVPR*, 2117–2125.

Liu, Y.-C.; Huang, Y.-K.; Chiang, H.-Y.; Su, H.-T.; Liu, Z.-Y.; Chen, C.-T.; Tseng, C.-Y.; and Hsu, W. H. 2021. Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining. *arXiv preprint arXiv:2104.04687*.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022. Video swin transformer. In *CVPR*, 3202–3211.

Mirzadeh, S. I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; and Ghasemzadeh, H. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 5191–5198.

Peng, S.; Genova, K.; Jiang, C. M.; Tagliasacchi, A.; Pollefeys, M.; and Funkhouser, T. 2023. OpenScene: 3D Scene Understanding with Open Vocabularies.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rist, C. B.; Emmerichs, D.; Enzweiler, M.; and Gavrila, D. M. 2021. Semantic scene completion using local deep implicit functions on lidar data. *TPAMI*, 44(10): 7205–7218.

Roldao, L.; de Charette, R.; and Verroust-Blondet, A. 2020. Lmscnet: Lightweight multiscale 3d semantic completion. In *3DV*, 111–119. IEEE.

Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; and Funkhouser, T. 2017. Semantic scene completion from a single depth image. In *CVPR*, 1746–1754.

Wang, A.; Chen, H.; Lin, Z.; Han, J.; and Ding, G. 2023. RepViT: Revisiting Mobile CNN From ViT Perspective. arXiv:2307.09283.

Wang, L.; Li, X.; Liao, Y.; Jiang, Z.; Wu, J.; Wang, F.; Qian, C.; and Liu, S. 2022. Head: Hetero-assists distillation for heterogeneous object detectors. In *European Conference on Computer Vision*, 314–331. Springer.

Wang, S.; Yu, J.; Li, W.; Liu, W.; Liu, X.; Chen, J.; and Zhu, J. 2024. Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation. In *CVPR*, 14792–14801.

Wang, Y.; and Tong, C. 2024. H2gformer: Horizontal-to-global voxel transformer for 3d semantic scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5722–5730.

Xia, Z.; Liu, Y.; Li, X.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; and Qiao, Y. 2023. SCPNet: Semantic Scene Completion on Point Cloud. In *CVPR*, 17642–17651.

Xue, Y.; Li, R.; Wu, F.; Tang, Z.; Li, K.; and Duan, M. 2024. Bi-SSC: Geometric-Semantic Bidirectional Fusion for Camera-based 3D Semantic Scene Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20124–20134.

Yan, X.; Gao, J.; Li, J.; Zhang, R.; Li, Z.; Huang, R.; and Cui, S. 2021. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *AAAI*, volume 35, 3101–3109.

Yan, X.; Gao, J.; Zheng, C.; Zheng, C.; Zhang, R.; Cui, S.; and Li, Z. 2022. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *European Conference on Computer Vision*, 677–695. Springer.

Yang, J.; Ding, R.; Deng, W.; Wang, Z.; and Qi, X. 2023. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. *arXiv preprint arXiv:2304.00962*.

Yao, J.; Li, C.; Sun, K.; Cai, Y.; Li, H.; Ouyang, W.; and Li, H. 2023. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *ICCV*, 9455–9465.

Zhang, J.; Zhao, H.; Yao, A.; Chen, Y.; Zhang, L.; and Liao, H. 2018. Efficient semantic scene completion network with spatial group convolution. In *ECCV*, 733–749.

Zhang, Y.; Zhu, Z.; and Du, D. 2023. OccFormer: Dual-path Transformer for Vision-based 3D Semantic Occupancy Prediction. *arXiv preprint arXiv:2304.05316*.

Zheng, Y.; Li, X.; Li, P.; Zheng, Y.; Jin, B.; Zhong, C.; Long, X.; Zhao, H.; and Zhang, Q. 2024. Monoocc: Digging into monocular semantic occupancy prediction. *arXiv preprint arXiv:2403.08766*.

Zhou, S.; Liu, W.; Hu, C.; Zhou, S.; and Ma, C. 2023. UniDistill: A Universal Cross-Modality Knowledge Distillation Framework for 3D Object Detection in Bird's-Eye View. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5116–5125.

Zhu, X.; Zhou, H.; Wang, T.; Hong, F.; Ma, Y.; Li, W.; Li, H.; and Lin, D. 2021. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9939–9948.