

VisHall3D: Monocular Semantic Scene Completion from Reconstructing the Visible Regions to Hallucinating the Invisible Regions

Haoang Lu Yuanqi Su* Xiaoning Zhang Longjun Gao Yu Xue Le Wang
Xi'an Jiaotong University, China

<https://github.com/luang321/vishall3d>

Abstract

This paper introduces VisHall3D, a novel two-stage framework for monocular semantic scene completion that aims to address the issues of feature entanglement and geometric inconsistency prevalent in existing methods. VisHall3D decomposes the scene completion task into two stages: reconstructing the visible regions (vision) and inferring the invisible regions (hallucination). In the first stage, VisFrontierNet, a visibility-aware projection module, is introduced to accurately trace the visual frontier while preserving fine-grained details. In the second stage, OcclusionMAE, a hallucination network, is employed to generate plausible geometries for the invisible regions using a noise injection mechanism. By decoupling scene completion into these two distinct stages, VisHall3D effectively mitigates feature entanglement and geometric inconsistency, leading to significantly improved reconstruction quality.

The effectiveness of VisHall3D is validated through extensive experiments on two challenging benchmarks: SemanticKITTI and SSCBench-KITTI-360. VisHall3D achieves state-of-the-art performance, outperforming previous methods by a significant margin and paves the way for more accurate and reliable scene understanding in autonomous driving and other applications.

1. Introduction

Monocular Semantic Scene Completion (Monocular SSC) [3, 20, 29, 42, 43] has emerged as a promising solution for enabling autonomous vehicles to perceive and understand their surroundings, as it can reconstruct complete 3D scenes using only single RGB images. This cost-effective and flexible approach has the potential to revolutionize 3D perception in autonomous driving. However, existing Monocular SSC methods still suffer from two major challenges: feature entanglement and geometric inconsistency.

These issues stem from the inherent distinction between visible and invisible regions in the 3D scene. The visible regions, or the visual frontier, correspond to the surface points of the 3D scene captured in the input image. When estimating 3D occupancy grids, the ground truth is typically obtained by accumulating data from multiple adjacent Lidar frames, inevitably including some voxels that are not visible in the current image. As a result, the occupancy estimation task faces two distinct challenges: tracing the visual frontier and generating the unseen voxels, as shown in Fig.1.

Existing methods, such as MonoScene [3], Ocdepth [29], and NDC-Scene [42], among others [12, 20, 37, 43, 47], have made significant progress in Monocular SSC. However, by treating SSC as a single-stage task, they fail to recognize the inherent distinction between visible and invisible regions, leading to feature entanglement and geometric inconsistency. Feature entanglement occurs when the features learned for visible and invisible regions are mixed together, while geometric inconsistency arises when the reconstructed visible and invisible regions are misaligned or have conflicting structures.

To address these challenges, we propose VisHall3D, a novel two-stage framework that explicitly separates the tasks of reconstructing visible regions (vision) and inferring invisible regions (hallucination). In the first stage, VisHall3D uses VisFrontierNet, a visibility-aware projection module, to accurately trace the visual frontier while preserving fine-grained details. By explicitly modeling the boundary between visible and invisible regions, VisFrontierNet helps to mitigate feature entanglement. In the second stage, VisHall3D employs OcclusionMAE, a hallucination network that generates plausible geometries for invisible regions using a noise injection mechanism. By decomposing SSC into these two distinct stages, VisHall3D effectively addresses feature entanglement and geometric inconsistency, leading to significantly improved reconstruction quality, as shown in Fig.1.

We validate the effectiveness of VisHall3D through extensive experiments on SemanticKITTI [1] and SSCBench-KITTI-360 [21, 25], achieving state-of-the-art performance.

*Corresponding Author. Email: yuanqisu@mail.xjtu.edu.cn

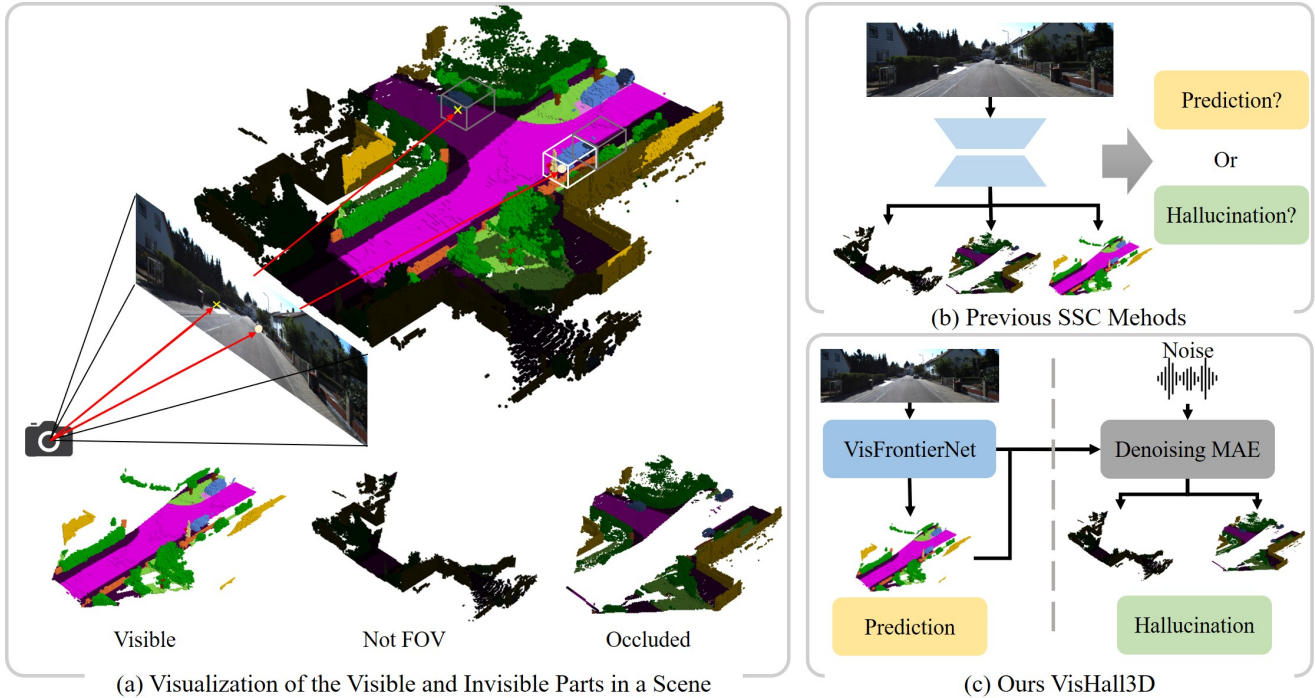


Figure 1. Visualization of the decoupling of prediction for visible regions and hallucination for invisible regions: (a) Division of visible and invisible areas: OOV (Out-of-view) voxels are in black, and occluded FOV (Field-of-View) voxels are in dark; (b) Pipeline of current mainstream methods; (c) Pipeline of our approach and its decoupling strategy for prediction and hallucination.

Our main contributions are:

- A novel two-stage framework, VisHall3D, for Monocular SSC that explicitly separates vision and hallucination processes to mitigate feature entanglement and geometric inconsistency.
- VisFrontierNet, a visibility-aware projection module that accurately traces the visual frontier by modeling the boundary between visible and invisible regions.
- OcclusionMAE, a hallucination network that generates plausible geometries for invisible regions using a noise injection mechanism.
- VisHall3D sets a new standard for Monocular SSC, paving the way for more accurate and reliable scene understanding in various applications.

2. Related Work

2.1. Semantic Scene Completion

3D Semantic Scene Completion aims to generate dense 3D semantic voxel grids from incomplete observations, as first defined in SSCNet [34]. Early SSC works typically relied on geometric inputs such as LiDAR [6, 31, 40] or depth maps [18, 19, 34]. However, recent studies have explored reconstructing entire SSC scenes using visual-only inputs, starting from multi-view methods [11, 22], then progressing to stereo-based approaches [14], and finally to monocular

frameworks [3, 42].

Among multi-view methods, BEVFormer [22] and TPVFormer [11] leverage diverse 3D feature representations to reduce computational overhead and enhance network capability. FB-OCC [24] and OccTransformer [26], on the other hand, focus on improving the transition from 2D to 3D.

Monoscene [3] pioneered single-image SSC, while subsequent works introduced NDC coordinates [42], depth prediction [20], horizontal direction emphasis [37], and voxel-to-graph structures [41] to optimize 2D-to-3D transformation. The latest advancements, Symphonies [12] and CGFormer [43], refine the approach with object-aware and context-aware optimizations. However, these methods still suffer from feature entanglement and geometric inconsistency due to the inherent ambiguity in monocular 3D reconstruction.

2.2. 3D From a Single Image

3D reconstruction from a single image is an ill-posed problem due to the lack of explicit depth information. Early monocular 3D tasks focused on reconstructing coarse information within the visible region by extending 2D methods [28, 48], exploiting geometric knowledge of objects [13, 23, 44, 45], or treating it as a Perspective-n-Points (PnP) problem [5, 27].

The emergence of monocular semantic scene comple-

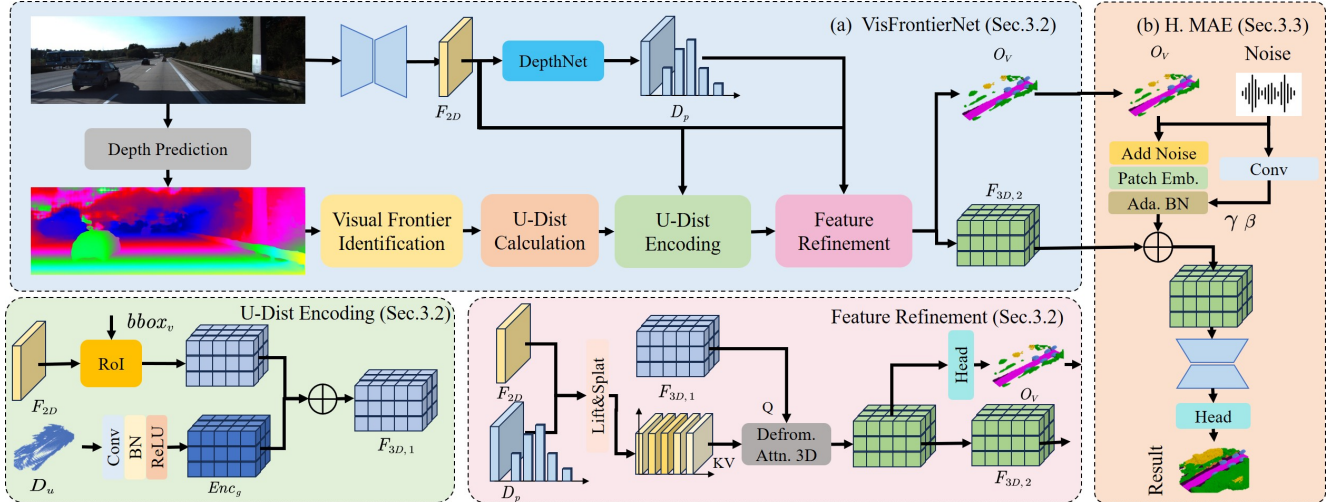


Figure 2. Overview of our method and the structure of each module. (a) The pipeline of our proposed VisFrontierNet, including unsigned distance (uDistance) Encoding and Feature Refinement. (b) The detailed architecture of the OcclusionMAE. The unsigned distance is calculated for the identified visual frontier which is then encoded and combined with image feature for the 3D representation.

tion (SSC) [3, 36, 47] has expanded the scope of monocular 3D tasks to reconstructing both visible and invisible regions with detailed shapes and semantics. This requires models to possess not only visual perception capabilities but also the ability to hallucinate missing information. However, existing methods often struggle to effectively handle the distinction between visible and invisible regions, leading to sub-optimal results.

2.3. Multi-stage Refinement

Multi-stage refinement strategy has been widely adopted in computer vision, particularly in object detection [2, 4, 35, 49], to refine proposals and enhance accuracy. Inspired by this, we propose to decouple the tasks of visual perception and hallucination in monocular SSC. We first generate a coarse prediction of the visible regions, and then leverage Hallucinating MAE [9] to hallucinate and optimize the entire scene.

3. VisHall3D: VisFrontierNet and Occlusion-MAE

We propose VisHall3D, a two-stage framework for generating a 3D occupancy grid from a single image: (1) refinement of the visual frontier by *VisFrontierNet* and (2) completion of the invisible voxels through *OcclusionMAE*, as shown in Fig.2. The visual frontier refers to the set of voxels directly visible from the camera’s viewpoint, corresponding to the surface points of the 3D scene captured in the input image, as shown in Fig.3.

In Sec.3.1, we provide an overview of the proposed model, including its overall architecture and key compo-

nents. We then delve into the details of *VisFrontierNet* in Sec.3.2, explaining its role in refining the visual frontier. Subsequently, in Sec.3.3, we present *OcclusionMAE* and its approach to completing the invisible voxels. Each subsection offers a comprehensive explanation of the respective modules and their roles within the framework.

3.1. Method Overview

In our approach, we divide the 3D occupancy grid voxels into visible, occluded, and out-of-view categories based on visibility (Fig.1). Visible voxels are directly observed; occluded voxels are hidden by obstructions; out-of-view voxels lie beyond the camera’s field of view.

Our proposed method seamlessly integrates feature lifting and 3D generation to reconstruct accurate 3D occupancy from a single input image.

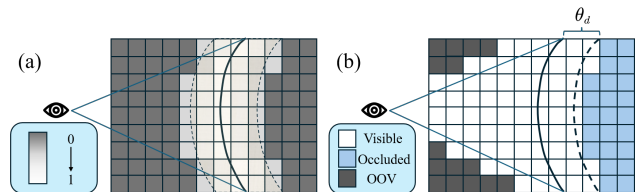


Figure 3. Illustration of the visual frontier from side view. The solid line gives the estimated depth.

Estimating Visible Voxel Occupancy: For visible voxels, we leverage the direct correspondence between the 2D image pixels and their corresponding 3D coordinates in the occupancy grid. Given the estimated depth map D , we determine the visible frontier in the 3D space and identify the voxels in front of the frontier as visible voxels.

To capture the frontier’s geometry, we compute a truncated unsigned distance map D_u representing voxel distances to the frontier. This map is encoded using 3D convolutions and fused with image features projected into 3D space using voxel-projected bounding boxes $bbox_v$ and RoI pooling, injecting visual information into the 3D volume.

We then refine the visual frontier by lifting 2D features into 3D using a depth probability map D_p and scattering them onto a discretized view frustum. Finally, 3D deformable attention refines frontier features, focusing the network on informative regions for better reconstruction.

Inferring Occluded and Out-of-View Voxel Occupancy: For occluded and out-of-view voxels, the direct visual evidence is not available, requiring a more sophisticated generation process. We propose OcclusionMAE, a denoising Masked Autoencoder that generates the occupancy for the entire set of voxels by taking the noisy prediction on the visible voxels from VisFrontierNet as input.

The denoising procedure in OcclusionMAE involves adding noise to the visible voxel predictions and injecting the noise level into the network via adaptive batch normalization. OcclusionMAE employs a 3D U-Net architecture to process the modulated features and generate the final occupancy grid, effectively aligning the results with the image semantics while accounting for the noisy nature of the visible voxel predictions.

3.2. Visual Frontier Propagation via VisFrontierNet

VisFrontierNet is designed to estimate the occupancy of visible voxels by leveraging the correspondence between 2D image pixels and their 3D coordinates in the occupancy grid. Given the 3D grids G , our goal is to predict the label of each voxel $v \in G$ from an input image. The main steps of VisFrontierNet include: (1) identifying the visual frontier, (2) representing and encoding it, and (3) refining its feature.

We first transform the coordinates of each voxel into the camera frame; then project them onto the image plane, obtaining the corresponding coordinates (x_v, y_v, d_v) for each voxel v . Let π denote the projection process, it maps each voxel v to a corresponding 2D coordinate (x_v, y_v) on the image plane.

$$(x_v, y_v, d_v) = \pi(v) \quad (1)$$

Here, we also obtain the depth value d_v for each voxel.

Visual Frontier Identification. The estimated depth map D gives a visual frontier that determines the visible and occluded voxels. The voxels in front of the visual frontier are considered visible, as shown in Fig. 3.

$$V := \{v | v \in G, d_v < D(x_v, y_v) + \theta_d\} \quad (2)$$

By comparing the projected depth d_v with the estimated depth of the projected point $D(x_v, y_v)$, we determine

whether the voxel lies in front of or behind the visual frontier. A relaxation factor θ_d is added to accommodate potential errors in depth estimation. The filtered visible voxels V allow our network to focus on estimating the correct lifting from the 2D image plane into the 3D space.

Unsigned Distance Function and Encoding. To capture the geometric characteristics of the visual frontier, we introduce the truncated unsigned distance map (D_u). It measures the unsigned distance $dist_v = |d_v - D(x_v, y_v)|$ from each voxel to the visual frontier, and is defined as:

$$D_u = \begin{cases} 2 - 2\sigma(\gamma * dist_v), & dist_v < \theta \\ 0, & \text{Otherwise} \end{cases} \quad (3)$$

where σ is the sigmoid function, γ is a controlling factor set to 10 for sharp decaying, and θ controls the truncated region (set to 1), as shown in Fig.3 (a).

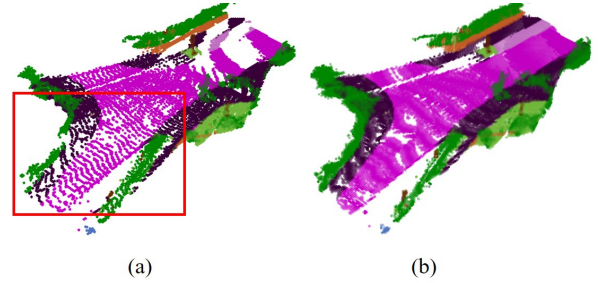


Figure 4. Visual comparison of Hard Lifting and our Unsigned Distance approach. (a) Illustration of Hard Lifting. (b) Geometric information captured by our Unsigned Distance approach.

Through this approach, we address the challenge of sparse voxels in distant regions, a limitation inherent in the hard lifting of geometric information adopted by previous methods [12, 20], where each pixel corresponds to a single voxel, as illustrated in the Fig. 4.

The unsigned distance map D_u is then fed into a 3D convolutional module for geometric encoding Enc_g , which captures the geometric properties of the visual frontier. To inject the image features into the 3D space, we combine the geometric encoding with the image feature for each voxel using RoI pooling [7]:

$$F_{3D,1}(v) = Enc_g(v) \oplus RoI(F_{2D}, bbox_v) \quad (4)$$

where F_{2D} denotes the image feature map, and $bbox_v$ is the projected bounding box for voxel v . In practice, We use ResNet-50 [8] as the backbone for extracting multi-scale features from the input image. Similar to Symphonies [12], we utilize MaskDINO [16]’s neck to combine multi-scale image features into a single one F_{2D} .

Feature Refinement for Visual Frontier. After encoding both the geometric and visual features for the frontier,

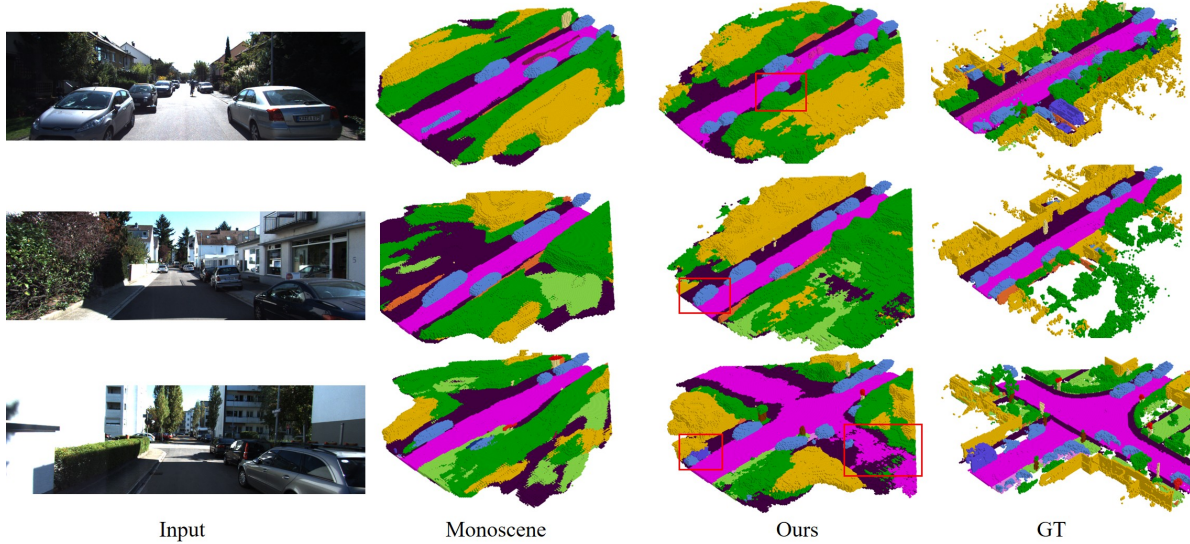


Figure 5. Qualitative visual comparison with Monoscene[3] and Ground Truth (GT) on the SemanticKITTI[1] dataset.

a refinement procedure is utilized to modify the representation for visual frontier according to image features. We lift the 2D features F_{2D} into the 3D space using a depth probability map D_p , which represents the likelihood of each pixel belonging to different depth ranges. The lifted features are then scattered onto a discretized grid on the view frustum [30], as shown in Fig.2.

To refine the features for the visual frontier, we employ 3D deformable attention [17, 49]. It learns to dynamically adjust the receptive field and sampling locations based on the input features, enabling the network to focus on the most informative regions and adapt to the specific characteristics of the visual frontier, producing refined 3D features $F_{3D,2}$.

$$F_{3D,2} \leftarrow \text{Deform3D}(F_{3D,1}, F_{2D}, D_p) \quad (5)$$

3.3. Generation via Denoising MAE

OcclusionMAE is a denoising Masked Autoencoder designed to generate the occupancy for the entire set of voxels, taking the noisy prediction O_V on the visible voxels from VisFrontierNet as input.

Noise Adding to Visible Voxel Predictions. Considering the noisy nature of the predicted occupancy on visible voxels, we introduce a denoising procedure. For each visible voxel in V , we randomly assign it a value from its neighboring voxels within a horizontal range R_h and a depth range R_d . Here, considering the spatial resolution of the scene (256 voxels in the horizontal and depth directions, but only 32 voxels in the vertical direction), we focus on adding noise in the horizontal and depth directions while leaving the vertical direction unchanged. The noise adding procedure is described as follows.

$$\tilde{O}_V = \text{AddNoise}(O_V, t(R_h, R_d)) \quad (6)$$

Where, t is the noise level. It is important to note that the added noise does not alter the original semantics (e.g., changing a vehicle to a person) but only perturbs geometric information (e.g., adjusting voxel positions). This is because we do not want the network to generate non-existent classes.

Noise Level Injecting via Adaptive Batch Normalization. The noise level t is injected into the network using adaptive batch normalization [10]. Let F denote the features extracted from \tilde{O}_V through patch embedding and 3D convolution. The adaptive batch normalization works as follows:

$$\text{ada.BN}(F(v)) = \gamma \cdot \frac{F(v) - \mu}{\sqrt{\delta^2 + \epsilon}} + \beta \quad (7)$$

where μ and δ^2 are the mean and variance of F , respectively. The features from the noise level t are used to generate the parameters γ and β . This allows the network to adapt to different noise levels and generate more accurate occupancy predictions.

Generating Occupancy Grid with 3D U-Net. The resulting 3D feature is concatenated with refined context feature derived from VisFrontierNet $F_{3D,2}(v)$ and fed into a 3D U-Net to produce the final occupancy prediction. The 3D U-Net processes the modulated features and generates the occupancy grid for the entire set of voxels, effectively aligning the results with the image semantics while accounting for the noisy nature of the visible voxel predictions.

The denoising procedure can be summarized as follows.

$$O = \text{OcclusionMAE}(\tilde{O}_V, t) \quad (8)$$

By introducing a denoising procedure and injecting the

Method	Date	IoU \uparrow	mIoU \uparrow	road	sidewalk	parking	other-grnd.	building	car	truck	bicycle	motorcycle	other-veh.	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traf.-sign
				█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█
Stereo camera-based methods																						
StereoScene[14]	IJCAI2024	43.34	15.36	61.90	31.20	30.70	10.70	24.20	22.80	2.80	3.40	2.40	6.10	23.80	8.40	27.00	2.90	2.20	0.50	16.50	7.00	7.20
Monocular temporal methods																						
VoxFormer-T[20]	CVPR2023	43.21	13.41	54.10	26.90	25.10	7.30	23.50	21.70	3.60	1.90	1.60	4.10	24.40	8.10	24.20	1.60	1.10	0.00	13.10	6.60	5.70
HASSC-T[36]	CVPR2024	42.87	14.38	55.30	29.60	25.90	11.30	23.10	23.00	2.90	1.90	1.50	4.90	24.80	9.80	26.50	1.40	3.00	0.00	14.30	7.00	7.10
HTCL[15]	ECCV2024	44.23	17.09	64.40	34.80	33.80	12.40	25.90	27.30	5.70	1.80	2.20	5.40	25.30	10.80	31.20	1.10	3.10	0.90	21.10	9.00	8.30
H2GFormer-T[37]	AAAI2024	43.52	14.60	57.90	30.40	30.00	6.90	24.00	23.70	5.20	0.60	1.20	5.00	25.20	10.70	25.80	1.10	0.10	0.00	14.60	7.50	9.30
Monocular single-frame methods																						
MonoScene[3]	CVPR2023	34.16	11.08	54.70	27.10	24.80	5.70	14.40	18.80	3.30	0.50	0.70	4.40	14.90	2.40	19.50	1.00	1.40	0.40	11.10	3.30	2.10
VoxFormer-S[20]	CVPR2023	42.95	12.20	53.90	25.30	21.10	5.60	19.80	20.80	3.50	2.60	0.70	3.70	22.40	7.50	21.30	1.40	2.60	0.20	11.10	5.10	4.90
TPVFormer[11]	CVPR2023	34.25	11.26	55.10	27.20	27.40	6.50	14.80	19.20	3.70	1.00	0.50	2.30	13.90	2.60	20.40	1.10	2.40	0.30	11.00	2.90	1.50
SurroundOcc[38]	ICCV2023	34.72	11.86	56.90	28.30	30.20	6.80	15.20	20.60	1.40	1.60	1.20	4.40	14.90	3.40	19.30	1.40	2.00	0.10	11.30	3.90	2.40
OccFormer[46]	ICCV2023	34.53	12.32	55.90	30.30	31.50	6.50	15.70	21.60	1.20	1.50	1.70	3.20	16.80	3.90	21.30	2.20	1.10	0.20	11.90	3.80	3.70
IAMSSC[39]	T-ITS2024	43.74	12.37	54.00	25.50	24.70	6.90	19.20	21.30	3.80	1.10	0.60	3.90	22.70	5.80	19.40	1.50	2.90	0.50	11.90	5.30	4.10
DepthSSC[41]	arXiv2024	44.58	13.11	55.64	27.25	25.72	5.78	20.46	21.94	3.74	1.35	0.98	4.17	23.37	7.64	21.56	1.34	2.79	0.28	12.94	5.87	6.23
HASSC-S[36]	CVPR2024	43.40	13.34	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Symphonize[12]	CVPR2024	42.19	15.04	58.40	29.30	26.90	11.70	24.70	23.60	3.20	3.60	<u>2.60</u>	5.60	24.20	10.00	23.10	3.20	1.90	2.00	16.10	7.70	8.00
H2GFormer-S[37]	AAAI2024	44.20	13.72	56.40	28.60	26.50	4.90	22.80	23.40	4.80	0.80	0.90	4.10	24.60	9.10	23.80	1.20	2.50	0.10	13.30	6.40	6.30
MonoOcc-L[47]	ICRA2024	-	15.63	59.10	30.90	27.10	9.80	22.90	23.90	<u>7.20</u>	4.50	2.40	7.70	<u>25.00</u>	9.80	26.10	<u>2.80</u>	<u>4.70</u>	0.60	16.90	7.30	8.40
CGFormer[43]	NIPS2024	44.41	16.63	<u>64.30</u>	34.20	34.10	<u>12.10</u>	<u>25.80</u>	<u>26.10</u>	4.30	<u>3.70</u>	1.30	2.70	24.50	<u>11.20</u>	29.30	1.70	3.60	0.40	<u>18.70</u>	<u>8.70</u>	9.30
Ours	ICCV2025	46.50	17.46	64.60	<u>34.10</u>	<u>32.00</u>	12.50	26.90	26.70	7.50	2.90	3.30	<u>6.20</u>	27.30	12.50	<u>28.00</u>	2.30	5.10	<u>1.90</u>	19.50	9.20	<u>9.20</u>

Table 1. Quantitative results on the hidden test set of SemanticKITTI [1], where the highest and second-highest scores for each metric are highlighted in **bold** and underline, respectively.

noise level into the network, *OcclusionMAE* learns to generate accurate occupancy predictions.

3.4. Training Losses

To effectively train VisFrontierNet and OcclusionMAE, we employ a combination of several loss functions that capture different aspects of the 3D reconstruction problem. The primary loss is a category frequency-weighted cross-entropy loss \mathcal{L}_{ce} , which guides the learning of both networks. Additionally, we introduce the Scene-Class Affinity Loss, inspired by MonoScene [3], to separately constrain the recall, precision, and specificity of the scene’s geometry and semantics, denoted as \mathcal{L}_{geo} and \mathcal{L}_{sem} , respectively. Finally, a depth loss \mathcal{L}_d is applied to ensure the coincidence of the probabilistic depth distribution D_p and the depth map D .

4. Experimental Evaluations

In this section, we evaluate the performance of VisHall3D on two mainstream outdoor SSC (Semantic Scene Completion) datasets: SemanticKITTI [1] and SSCBench-KITTI-360 [21, 25]. We compare VisHall3D with state-of-the-art (SOTA) methods, including those utilizing stereo vision or temporal information. Furthermore, we conduct abla-

tion studies to investigate the impact of each module on the model’s performance.

4.1. Dataset and Evaluation Metrics

SemanticKITTI [1] and SSCBench-KITTI-360 [21, 25] are two widely-used outdoor SSC datasets. Both datasets adopt a scene range of $51.2\text{m} \times 51.2\text{m} \times 6.4\text{m}$, with each voxel having an edge length of 0.2m, resulting in a scene resolution of $256 \times 256 \times 32$. SemanticKITTI consists of 10 training sequences (3,834 samples), 1 validation sequence (815 samples), and 11 test sequences (3,992 samples), with 20 valid classes and 1 invalid class. SSCBench-KITTI-360 provides 7 training sequences (8,487 samples), 1 validation sequence (1,812 samples), and 1 test sequence (2,566 samples), with 19 valid classes.

Following previous methods[12, 20, 36, 37, 43, 47], we use MobileStereoNet[33] for depth prediction.

4.2. Comparisons

We present the comparison of VisHall3D with SOTA methods on the SemanticKITTI [1] and SSCBench-KITTI-360 [21, 25] datasets in Tab.1 and Tab.2, respectively.

On the SemanticKITTI dataset, VisHall3D achieves an mIoU of 17.46%, outperforming all existing monocular

Method	Date	IoU		Classes																	
		IoU \uparrow	mIoU \uparrow	car	bicycle	motorcycle	truck	other-veh.	person	road	parking	sidewalk	other-grnd.	building	fence	vegetation	terrain	pole	traf.-sign	other-struct	other-obj.
LiDAR-based methods																					
SSCNet [34]	CVPR2017	53.58	16.95	31.95	0.00	0.17	10.29	0.00	0.07	65.70	17.33	41.24	3.22	44.41	6.77	43.72	28.87	0.78	0.75	8.69	0.67
LMSCNet [32]	3DV 2020	47.35	13.65	20.91	0.00	0.00	0.26	0.58	0.00	62.95	13.51	33.51	0.20	43.67	0.33	40.01	26.80	0.00	0.00	3.63	0.00
Monocular camera-based methods																					
MonoScene [3]	CVPR2023	37.87	12.31	19.34	0.43	0.58	8.02	2.03	0.86	48.35	11.38	28.13	3.32	32.89	3.53	26.15	16.75	6.92	5.67	4.20	3.09
TPVFormer [11]	CVPR2023	40.22	13.64	21.56	1.09	1.37	8.06	2.57	2.38	52.99	11.99	31.07	3.78	34.83	4.80	30.08	17.52	7.46	5.86	5.48	2.70
OccFormer [46]	ICCV2023	40.27	13.81	22.58	0.66	0.26	9.89	3.82	2.77	54.30	13.44	31.53	3.55	36.42	4.80	31.00	19.51	7.77	8.51	6.95	4.60
VoxFormer [20]	CVPR2023	38.76	11.91	17.84	1.16	0.89	4.56	2.06	1.63	47.01	9.67	27.21	2.89	31.38	4.97	28.99	14.69	6.51	6.92	3.79	2.43
IAMSSC [39]	T-ITS2024	41.80	12.97	18.53	<u>2.45</u>	1.76	5.12	3.92	3.09	47.55	10.56	28.35	4.12	31.53	6.28	29.17	15.24	8.29	7.01	6.35	4.19
DepthSSC [41]	arXiv2024	40.85	14.28	21.90	2.36	4.30	11.51	4.56	2.92	50.88	12.89	30.27	2.49	37.33	5.22	29.61	21.59	5.97	7.71	5.24	3.51
Symphonies [12]	CVPR2024	44.12	18.58	<u>30.02</u>	1.85	<u>5.90</u>	25.07	12.06	8.20	54.94	13.83	32.76	6.93	35.11	8.58	38.33	11.52	14.01	9.57	14.44	11.28
CGFormer[43]	NIPS2024	48.07	<u>20.05</u>	29.85	3.42	3.96	17.59	6.70	6.63	63.85	17.15	40.72	5.53	42.73	8.22	38.80	24.04	16.24	17.45	10.18	6.77
Ours	ICCV2025	49.12	20.95	30.77	1.91	6.60	<u>17.99</u>	<u>8.72</u>	8.67	64.35	18.83	41.53	4.48	43.87	9.07	39.75	24.94	16.52	20.66	<u>10.30</u>	<u>7.99</u>

Table 2. Quantitative results on the test set of SSCBench-KITTI-360 [21, 25], where the highest and second-highest scores for each metric are highlighted in **bold** and underline, respectively.

Method	Params(M) \downarrow	Memory(M) \downarrow	Times(s) \downarrow
Monoscene[3]	149.5	19,041	0.49
CGFormer[43]	122.4	19,330	0.41
Ours	127.8	22,597	0.34

Table 3. Computational Complexity and Memory Usage

methods. Notably, VisHall3D surpasses the latest stereo method StereoScene [14], as well as the top-performing temporal method HTCL [15] in terms of both IoU and mIoU. This demonstrates the effectiveness of our decoupled two-stage framework in capturing both the visible and invisible regions of the scene.

On the SSCBench-KITTI-360 dataset, VisHall3D sets a new state-of-the-art with an mIoU of 20.95%, outperforming the previous best method CGFormer [43] by 0.90%. Remarkably, VisHall3D even surpasses early LiDAR-based methods such as SSCNet [34] and LMSCNet [32], showcasing the potential of monocular methods in capturing complex 3D scenes.

Fig. 5 presents a qualitative comparison of VisHall3D with MonoScene [3] and the ground truth on the SemanticKITTI dataset. VisHall3D generates more accurate and complete scene reconstructions, especially in the invisible regions occluded by foreground objects. This visual comparison further validates the superiority of our method in capturing the global scene context and hallucinating plausible geometries.

We also compare VisHall3D against SOTA methods in terms of computational complexity in Tab. 3. Despite our two-stage framework, VisHall3D maintains parameters comparable to CGFormer, while outperforming it in mIoU

and IoU. The slight memory increase is justified by performance gains. Under identical conditions, VisHall3D runs faster than CGFormer and MonoScene while delivering superior reconstruction quality.

4.3. Ablation Studies

In line with other works [3, 12, 20, 43], our ablation studies are primarily conducted on the SemanticKITTI [1] validation set. These studies encompass three key aspects: (1) the overall architectural components, (2) the division of visible and invisible regions, and (3) the noise incorporated in the OcclusionMAE.

Overall Architectural Components. Tab. 4 presents our component-wise analysis of the network architecture. The baseline model retains the backbone and neck but adopts a direct hard-assign lifting approach from [12, 20], where each pixel corresponds to a single voxel. Its architecture can be considered as a variation of Symphonies [12] where the Symphonies Decoder is replaced with a 3D UNet.

The introduction of unsigned distance function for the visual frontier not only enhances the mIoU but also significantly improves the IoU by 4.33%, likely due to its superior preservation of geometric information for the visual frontier in the predicted depth map. The unsigned distance function provides a more informative representation of the visual frontier compared to binary masks, enabling the network to better capture the visibility relationships between voxels. Subsequently, by incorporating the OcclusionMAE, we divided the network into two parts: the first part maintains the original structure but predicts voxels for visible regions only, while the second part, the OcclusionMAE, hal-

Method	IoU \uparrow	mIoU \uparrow	Params (M)	Memory (M)
Baseline	42.11	14.56	57.2	15260
VisFrontierNet w/o Feature Refinement	46.44 (+4.33)	15.11 (+0.55)	74.3	17349
+ OcclusionMAE w/o Denoising	46.38 (-0.06)	15.99 (+0.88)	86.7	18746
+ Feature Refinement	45.88 (-0.50)	16.59 (+0.60)	125.5	21246
+ Denoising	46.14 (+0.26)	17.06 (+0.47)	127.8	22597

Table 4. Ablation study of the architectural components on the validation set of SemanticKITTI [1], where the model initially used the hard assignment method from [12, 20] for lifting 2D features to 3D.

illuminates the entire scene, including invisible regions. Although this does not improve the IoU, it substantially boosts the mIoU by 0.88%.

Further, we integrated the feature refinement based on 3D deformable attention to enhance the model’s perception of visible regions. While this results in a slight decrease in IoU by 0.50%, the mIoU continues to improve by 0.60%, reaching 16.59. Finally, to enhance the capability of the OcclusionMAE, we introduced a denoising strategy, which further improves the mIoU by 0.47%.

Interestingly, VisFrontierNet without feature refinement achieves the highest IoU, showing that its unsigned distance function alone offers a strong geometric prior for accurate visible region prediction.

Invisibility	Threshold θ_d (m)	IoU \uparrow	mIoU \uparrow
OOV	-	43.98	15.89
OOV + Occ.	1.5	45.87	16.65
OOV + Occ.	2.5	45.83	16.92
OOV + Occ.	3.5	46.14	17.06
OOV + Occ.	4.5	45.94	16.95

Table 5. Ablation study on visible and invisible region division on SemanticKITTI [1]’s validation set, where OOV and Occ. refer to using the out of view and occlusion respectively.

Visibility Identification. Decoupling visible and invisible scenes is one of our most critical components, and the key lies in how to delineate what constitutes the visible and invisible regions. Tab. 5 presents our ablation study on the division of visible and invisible regions.

The baseline approach solely relies on the out of view checking to distinguish visible from invisible areas, considering only whether voxels project onto the image plane while entirely ignoring occlusion issues. As shown in Tab. 5, this naive implementation performs significantly worse than methods that account for occlusion. In occlusion-aware methods, we fine-tuned the threshold and found that 3.5m gives the best results, suggesting moderate visibility relaxation helps handle occlusion and depth uncertainty. However, larger thresholds (e.g., 4.5m) introduce

noise and harm performance.

R_h (Voxel)	R_d (Voxel)	IoU \uparrow	mIoU \uparrow
0	0	46.03	16.32
0	2	45.92	16.79
0	3	45.83	16.92
0	4	45.73	16.66
1	3	46.19	16.52
1	4	46.12	16.69

Table 6. Ablation study on the noise level in OcclusionMAE on SemanticKITTI [1], where R_h and R_d refer to the noise range in the horizontal and depth directions, respectively (conducted with the threshold θ_d of 2.5m).

Noise in the OcclusionMAE. To achieve effective hallucination, we introduce a certain level of noise into the OcclusionMAE, which also enhances the network’s robustness. Tab. 6 presents our study on the noise parameters in the Hallucinating MAE. These parameters consist of two components: the maximum sampling distance in the horizontal direction and the depth direction. When no noise is applied (both parameters are set to 0), the model performs worst on mIoU. In contrast, the best mIoU is achieved when the horizontal noise is set to 0 and the depth noise to 3.

Interestingly, the impact of increasing horizontal noise on the network is significantly greater than that of depth noise, likely due to the inherent inaccuracies in depth estimation. This finding suggests that the model is more sensitive to the perturbation in the horizontal direction, and a careful balance between the two noise components is crucial for optimal performance.

5. Conclusion

In this paper, we propose VisHall3D, a two-stage framework for monocular semantic scene completion that tackles feature entanglement and geometric inconsistency by explicitly separating visible-region reconstruction and invisible-region hallucination, and demonstrate its effectiveness with extensive experiments.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (No.62473306, N0.U24B20181, No.62441616) and Science and Technology Research and Development Plan of China State Railway Group Co., Ltd. (No. RITS2023KF03).

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 1, 5, 6, 7, 8
- [2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1483–1498, 2019. 3
- [3] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 1, 2, 3, 5, 6, 7
- [4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3
- [5] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2781–2790, 2022. 2
- [6] 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. 2
- [7] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4
- [9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3
- [10] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 5
- [11] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023. 2, 6, 7
- [12] Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20258–20267, 2024. 1, 2, 4, 6, 7, 8
- [13] Mykola Lavreniuk. Spidepth: Strengthened pose information for self-supervised monocular depth estimation. *arXiv preprint arXiv:2404.12501*, 2024. 2
- [14] Bohan Li, Yasheng Sun, Zhujin Liang, Dalong Du, Zhuanghui Zhang, Xiaofeng Wang, Yunnan Wang, Xin Jin, and Wenjun Zeng. Bridging stereo geometry and bev representation with reliable mutual interaction for semantic scene completion. *arXiv preprint arXiv:2303.13959*, 2023. 2, 6, 7
- [15] Bohan Li, Jiajun Deng, Wenyao Zhang, Zhujin Liang, Dalong Du, Xin Jin, and Wenjun Zeng. Hierarchical temporal context learning for camera-based semantic scene completion. *arXiv preprint arXiv:2407.02077*, 2024. 6, 7
- [16] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3041–3050, 2023. 4
- [17] Hongyang Li, Hao Zhang, Zhaoyang Zeng, Shilong Liu, Feng Li, Tianhe Ren, and Lei Zhang. Dfa3d: 3d deformable attention for 2d-to-3d feature lifting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6684–6693, 2023. 5
- [18] Jie Li, Yu Liu, Xia Yuan, Chunxia Zhao, Roland Siegwart, Ian Reid, and Cesar Cadena. Depth based semantic scene completion with position importance aware loss. *IEEE Robotics and Automation Letters*, 5(1):219–226, 2019. 2
- [19] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3351–3359, 2020. 2
- [20] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9098, 2023. 1, 2, 4, 6, 7, 8
- [21] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, Yue Wang, Hang Zhao, Zhiding Yu, and Chen Feng. Ss-cbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024. 1, 6, 7
- [22] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. 2

- [23] Zhuoling Li, Zhan Qu, Yang Zhou, Jianzhuang Liu, Haoqian Wang, and Lihui Jiang. Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2791–2800, 2022. 2
- [24] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 2
- [25] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022. 1, 6, 7
- [26] Jian Liu, Sipeng Zhang, Chuixin Kong, Wenyuan Zhang, Yuhang Wu, Yikang Ding, Borun Xu, Ruibo Ming, Donglai Wei, and Xianming Liu. Occtransformer: Improving bev-former for 3d camera-only occupancy prediction. *arXiv preprint arXiv:2402.18140*, 2024. 2
- [27] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. Autosshape: Real-time shape-aware monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15641–15650, 2021. 2
- [28] Shujie Luo, Hang Dai, Ling Shao, and Yong Ding. M3dssd: Monocular 3d single stage object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6145–6154, 2021. 2
- [29] Ruihang Miao, Weizhou Liu, Mingrui Chen, Zheng Gong, Weixin Xu, Chen Hu, and Shuchang Zhou. Occdepth: A depth-aware method for 3d semantic scene completion. *arXiv preprint arXiv:2302.13540*, 2023. 1
- [30] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Proceedings of the European Conference on Computer Vision*, 2020. 5
- [31] 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. 2
- [32] Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 111–119. IEEE, 2020. 7
- [33] Faranak Shamsafar, Samuel Woerz, Rafia Rahim, and Andreas Zell. Mobilestereonet: Towards lightweight deep networks for stereo matching. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2417–2426, 2022. 6
- [34] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. 2, 7
- [35] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021. 3
- [36] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14792–14801, 2024. 3, 6
- [37] Yu Wang and Chao Tong. H2gformer: Horizontal-to-global voxel transformer for 3d semantic scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5722–5730, 2024. 1, 2, 6
- [38] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023. 6
- [39] Haihong Xiao, Hongbin Xu, Wenxiong Kang, and Yuqiong Li. Instance-aware monocular 3d semantic scene completion. *IEEE Transactions on Intelligent Transportation Systems*, 25(7):6543–6554, 2024. 6, 7
- [40] 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*, pages 3101–3109, 2021. 2
- [41] Jiawei Yao and Jusheng Zhang. Depthssc: Depth-spatial alignment and dynamic voxel resolution for monocular 3d semantic scene completion. *arXiv preprint arXiv:2311.17084*, 2023. 2, 6, 7
- [42] 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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9421–9431. IEEE Computer Society, 2023. 1, 2
- [43] Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-liang Shen. Context and geometry aware voxel transformer for semantic scene completion. In *Advances in Neural Information Processing Systems*, pages 1531–1555, 2024. 1, 2, 6, 7
- [44] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9155–9166, 2023. 2
- [45] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3298, 2021. 2
- [46] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023. 6, 7
- [47] Yupeng Zheng, Xiang Li, Pengfei Li, Yuhang Zheng, Bu Jin, Chengliang Zhong, Xiaoxiao Long, Hao Zhao, and Qichao

- Zhang. Monoocc: Digging into monocular semantic occupancy prediction. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 18398–18405. IEEE, 2024. [1](#), [3](#), [6](#)
- [48] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. [2](#)
- [49] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. [3](#), [5](#)