

OneOcc: Semantic Occupancy Prediction for Legged Robots with a Single Panoramic Camera

Hao Shi^{1,2,*} Ze Wang^{1,3,*} Shangwei Guo^{1,*} Mengfei Duan⁴ Song Wang¹ Teng Chen⁵
 Kailun Yang^{4,†} Lin Wang^{2,†} Kaiwei Wang^{1,†}
¹ZJU ²NTU ³MirrorMe Technology ⁴HNU ⁵Xiaomi Corporation

Abstract

Robust 3D semantic occupancy is crucial for legged/humanoid robots, yet most semantic scene completion (SSC) systems target wheeled platforms with forward-facing sensors. We present **OneOcc**, a vision-only panoramic SSC framework designed for gait-introduced body jitter and 360° continuity. **OneOcc** combines: (i) *Dual-Projection fusion (DP-ER)* to exploit the annular panorama and its equirectangular unfolding, preserving 360° continuity and grid alignment; (ii) *Bi-Grid Voxelization (BGV)* to reason in Cartesian and cylindrical-polar spaces, reducing discretization bias and sharpening free/occupied boundaries; (iii) a lightweight decoder with *Hierarchical AMoE-3D* for dynamic multi-scale fusion and better long-range/occlusion reasoning; and (iv) *plug-and-play Gait Displacement Compensation (GDC)* learning feature-level motion correction without extra sensors. We also release two panoramic occupancy benchmarks: **QuadOcc** (real quadruped, first-person 360°) and **Human360Occ (H3O)** (CARLA human-ego 360° with RGB, Depth, semantic occupancy; standardized within-/cross-city splits). **OneOcc** sets a new state of the art on **QuadOcc**, outperforming strong vision baselines and remaining competitive with classical LiDAR baselines; on **H3O** it gains +3.83 mIoU (within-city) and +8.08 (cross-city). Modules are lightweight, enabling deployable full-surround perception for legged/humanoid robots. Datasets and code will be publicly available at [OneOcc](#).

1. Introduction

A holistic 3D understanding of nearby surroundings is fundamental for robotic autonomy in navigation [1–6], locomotion [7–10], manipulation [11–14], and human-robot interaction [15–17]. Among 3D perception tasks, *Semantic Scene Completion (SSC)*—predicting complete volumetric

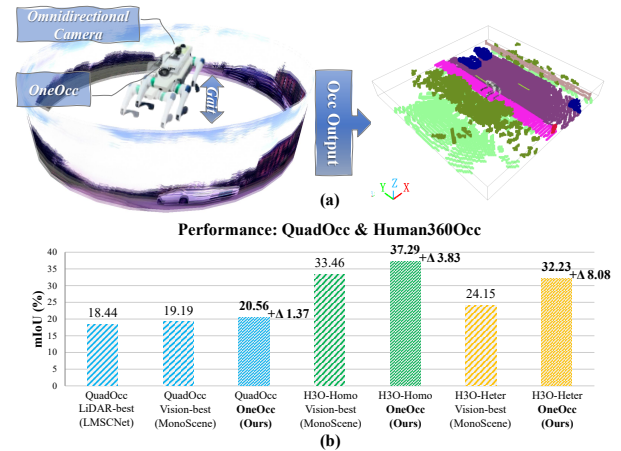


Figure 1. **Teaser.** (a) Single-sensor panoramic semantic scene completion (SSC) on a quadruped: an omnidirectional camera observes full 360° with gait-induced motion; **OneOcc** outputs voxelized semantic occupancy. (b) Summary on **QuadOcc** and **Human360Occ (H3O)**: bars show mIoU(%). On **QuadOcc**, **OneOcc** reaches 20.56 mIoU, surpassing the best LiDAR baseline (LMSCNet, 18.44) and vision baseline (MonoScene, 19.19). On **H3O**, **OneOcc** attains 37.29/32.23 mIoU under within-/cross-city, outperforming the best vision baselines (33.46/24.15).

geometry and semantics from partial observations—has become central [18–21], with most progress driven by autonomous driving where forward-facing sensors on stable wheeled platforms yield benign motion and sensing.

Why legged and humanoid platforms are different. Directly transferring SSC is non-trivial: agile gaits cause *body jitter* that corrupts evidence and breaks temporal coherence; these platforms require full 360° situational awareness in cluttered, uneven terrain; and tight payload/power budgets favor lightweight, single-sensor, low-latency solutions. Hence, we need a panoramic SSC pipeline explicitly robust to gait-induced motion.

Panoramic SSC with a single sensor. A single panoramic camera provides compact, true 360° coverage but introduces annular distortions and seam artifacts after unwrapping [36]—issues poorly handled by perspective-camera SSC pipelines. This motivates a design that preserves raw panoramic cues while exploiting grid-friendly projections.

*Equal contribution.

†Corresponding authors (e-mail: kailun.yang@hnu.edu.cn, linwang@ntu.edu.sg, wangkaiwei@zju.edu.cn).

Datasets	Data			Domain				Modality	Volume Size	Scenes	Frames	Anno.
	Camera	Sur.	Plat.	Move.	Env.	Weath.	Lighting					
NYUv2 [22]	Pinhole	✗			Indoor	✗	✗	C&D	(144×240×240)	1.4K	1.4K	
ScanNet [23]	Pinhole	✗			Indoor	✗	✗	C&D	(31×62×62)	1.5K	1.5K	
SUNCG [24]	Pinhole	✗			Indoor	✗	✗	C&D	(144×240×240)	46K	140K	
SemanticPOSS [25]	<i>n.a.</i>	✗			Outdoor	✗		L	(32×256×256)	6	3K	
SemanticKITTI [26]	Pinhole	✗			Outdoor	✗		C&L	(32×256×256)	22	9K	
nuScenes [27]	Pinhole+Fisheye	✓			Outdoor	✓		C&L	(40×512×512)	850	200K	
Occ3D-Waymo [28]	Pinhole+Fisheye	✓			Outdoor	✓		C&L	(32×200×200)	1K	200K	
SSCBench-KITTI-360 [29]	Fisheye	✓			Outdoor	✗		C&L	(32×200×200)	11	91K	
ScribbleSC [30]	Pinhole	✗			Outdoor	✗		C&L	(32×256×256)	11	9K	
OmniHD-Scenes [31]	Pinhole	✓			Outdoor	✗		C&L&R	—	200	60K	
WildOcc [32]	Pinhole	✗			Outdoor	✗		C&L	(40×100×100)	5	10K	
DSEC-SSC [33]	Pinhole	✗			Outdoor	✗		C&L&E	(16×128×128)	6	3K	
ORAD-3D [34]	Pinhole	✓			Outdoor	✓		C&L	—	145	58K	
Co3SOP [35]	Pinhole	✗			Outdoor	✗	✗	C&L	(70×1000×1000)	—	—	
Human360Occ (ours)	Panoramic	✓			Outdoor	✓		C&D	(16×128×128)	160	8K	
QuadOcc (ours)	Panoramic	✓			Outdoor	✗		C&L	(8×64×64)	10	24K	

Table 1. Typical datasets for Semantic Scene Completion (SSC). Abbreviations: (Autonomous Car), (Quadruped Robot), (Vision Foundation Model), (Human), (Synthetic), (Internet images/videos), (Wheels), (Gait), (Stationary), (Day), (Dusk), (Night), Sur. (Surround), Plat. (Platform), Move. (Movement), Env. (Environment), Weath. (Weather diversity), Anno. (Annotation), C (Camera), D (Depth), L (LiDAR), R (Radar).

Method: OneOcc. We propose OneOcc, a vision-only panoramic SSC framework for severe body jitter and 360° continuity, integrating: (i) *Dual-Projection fusion (DP-ER)* to jointly process the raw annular panorama and its equirectangular unfolding; (ii) *Bi-Grid Voxelization (BGV)* that reasons in Cartesian and polar/cylindrical voxel spaces to reduce discretization bias and better match panoramic geometry; (iii) a lightweight decoder with *Hierarchical AMoE-3D* for dynamic multi-scale 3D fusion; and (iv) a plug-and-play *Gait Displacement Compensation (GDC)* that learns feature-level jitter correction without extra sensors.

Benchmarks: QuadOcc and Human360Occ (H3O). To advance panoramic SSC on legged/humanoid robots, we release two complementary benchmarks. *QuadOcc* is a real first-person 360° dataset on a quadruped within a campus domain, with standardized day/dusk/night coverage and semi-automatic ground truth (multi-frame LiDAR aggregation, Grounded-SAM [37] initialization, targeted manual fixes). It contains 10 scenes and 24K frames, uses 6 semantic categories for training. The occupancy uses 0.4m voxels on a 64×64×8 grid—reflecting embodied compute and lower speed than cars. *Human360Occ (H3O)* is CARLA-based human-ego 360° with simulated gait, 160 sequences / 8K frames over 16 maps and multiple weathers/times; each frame provides RGB, metric depth, semantic occupancy at 64×64×8 and 128×128×16, and pose. It supports within-city (per-map 8:2 train/val) and cross-city (default 12/4 maps) splits. See details on the construction pipeline and dataset statistics in the supplementary material.

Results in brief. On *QuadOcc*, OneOcc reaches 20.56 mIoU, surpassing the best LiDAR baseline (LMSCNet, 18.44; +2.12 mIoU, +11.5%) and the best vision base-

line (MonoScene, 19.19; +1.37 mIoU). On *H3O*, it attains 37.29/32.23 mIoU under within-/cross-city, improving over the best vision baselines (33.46/24.15) by +3.83 and +8.08 mIoU (up to +33.5% relative); see Fig. 1(b).

Contributions.

- **Method.** *OneOcc* integrates *DP-ER*, *BGV*, *Hierarchical AMoE-3D*, and *GDC* to deliver jitter-robust, full-surround SSC with lightweight computation.
- **Datasets.** Two panoramic occupancy benchmarks for legged robots: *QuadOcc* (real quadruped, 360°, 10 scenes/24K frames, 6 categories, 0.4m voxels on 64×64×8) and *H3O* (CARLA human-ego 360°, 160 seq/8K frames on 16 maps, per-frame RGB/Depth/voxels/pose, two voxel resolutions).
- **Evaluation.** Extensive benchmarking shows vision-only *OneOcc* rivals or surpasses classic LiDAR-based methods under realistic legged/humanoid settings.

2. Related Work

Semantic scene completion. Semantic Scene Completion (SSC) predicts a complete 3D voxel scene with semantics from partial observations. *LiDAR-based SSC.* From SSC-Net [24] to LMSCNet [38], LiDAR pipelines evolved from dense 3D CNNs to efficient BEV-like decoders; later works (e.g., JS3C-Net [39], SCPNet [40]) enhance long-range context and 3D aggregation. Non-Cartesian discretizations mitigate quantization bias (PointOcc [41], inspired by Cylinder3D [42]); recent efforts also improve supervision efficiency and robustness [43, 44]. Yet LiDAR stacks are often heavy and power-hungry for constrained platforms [45–49]. *Vision-based SSC.* The ill-posed 2D→3D lifting has progressed rapidly: MonoScene [50] set a strong baseline;

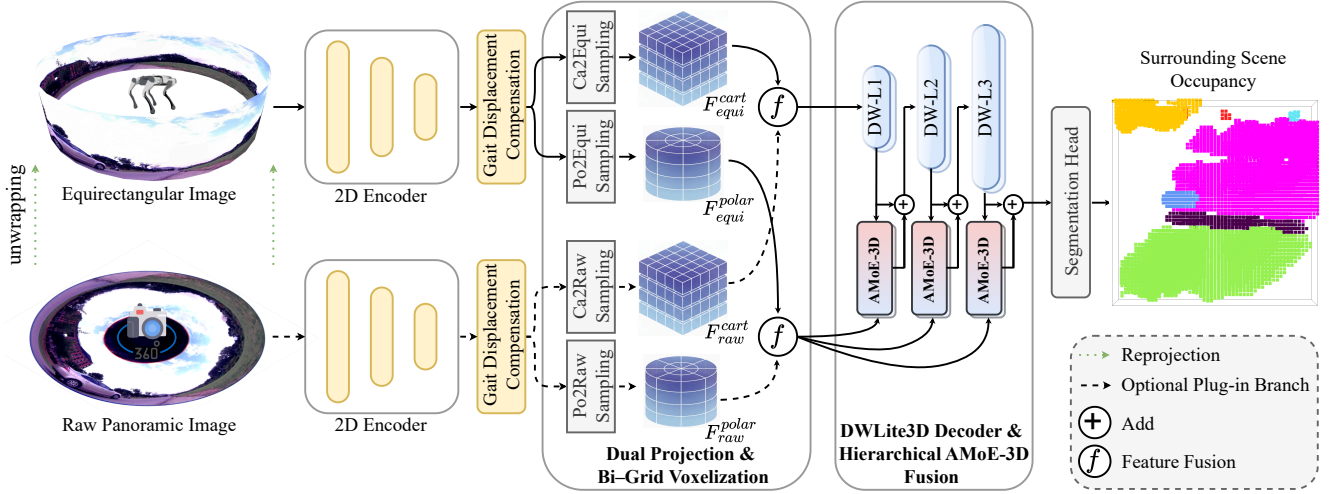


Figure 2. **OneOcc pipeline.** Calibrated *unwrapping* maps the raw panorama to equirectangular. Two 2D encoders (*DP-ER*) feed *GDC* for feature-level gait compensation (optional on H3O). At three scales, *BGV* builds Cartesian and polar/cylindrical voxel grids; voxel-centroid features are sampled to each view (Ca/Po $\rightarrow$ Equi/Raw) and fused. A lightweight 3D decoder with *Hierarchical AMoE-3D* aggregates multi-scale context, followed by a head predicting full-surround occupancy. *Abbrev.*: Ca/Po = Cartesian/Polar; Equi = equirectangular.

transformer models (VoxFormer [51], OccFormer [52]) improve global aggregation. Follow-ups refine supervision/projection (RenderOcc [53], InverseMatrixVT3D [54], SfMOcc [55]) and pursue efficiency via sparsity or streamlined designs [56–58]. Other lines study context disentangling/causality [59, 60] and scene-adaptive decoders or occlusion-aware projections [61, 62]; for deployment, OccFiner [63] and TALOS [64] target offboard refinement and test-time adaptation. *Gaussian/weak-supervision SSC.* GaussianOcc-style methods [65, 66] reduce dense 3D labeling via Gaussian/flow priors, and GaussianFormer variants [67, 68] model scenes as Gaussian fields for efficient camera-based occupancy. *Panoramic SSC and multi-modality.* While most camera-based SSC methods assume forward pinhole views, emerging works leverage omnidirectional sensing: panoramic depth for occupancy [69], real-time fisheye occupancy without semantics [70], cylindrical fusion [71], event cameras for HDR/blur [33], and radar-camera fusion for joint tasks [72]. Multimodal pipelines [45, 73–75] highlight complementary cues but increase system complexity. In contrast, we target a *single panoramic camera* with *full-360° continuity* under *gait-induced jitter*, a setting under-explored by prior literature.

Semantic scene understanding for legged robots. Legged autonomy motivates perception tailored to agile motion and strict payload budgets. Prior efforts include vision-based terrain reconstruction [76], large-scale datasets [77], and traversability learning [78–80]; complementary modules cover robust tracking [81, 82] and instruction grounding [83, 84]. With increasing evidence that limited FoV hinders agility [85–87], panoramic sensors are adopted for omnidirectional tracking [88], multimodal odometry [89], and even satellite communication on the move [90]. Some

works study probabilistic semantic mapping [91], and humanoid systems explore panoramic/occupancy perception [21, 92–94]. However, *dense 3D semantic occupancy* from a *single panoramic camera* on dynamic legged platforms is still largely open. We address this by releasing *QuadOcc* and *Human360Occ* and proposing *OneOcc*, a lightweight panoramic occupancy framework built around dual-projection fusion, bi-grid voxelization, hierarchical 3D MoE fusion, and gait displacement compensation.

3. Method

3.1. Problem Formulation

Given a panoramic RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ and calibrated intrinsics/extrinsics, we aim to predict a dense 3D semantic occupancy $\mathbf{S} \in \{0, \dots, C\}^{X \times Y \times Z}$. We produce per-voxel logits $\mathbf{Z} \in \mathbb{R}^{X \times Y \times Z \times C}$, and the final label is $\hat{s}_{x,y,z} = \arg \max_c \mathbf{Z}_{x,y,z,c}$.

3.2. Overview

Motivation. (1) A 360° panorama exhibits *ring continuity* and strong projection distortion: a single projection cannot balance resolution and receptive fields near the poles vs. the equator; (2) Legged locomotion induces *gait jitters* (impulsive foot strikes, small roll/pitch) that hurt the feature $\rightarrow$ voxel lifting step more than the image $\rightarrow$ feature step; (3) Surrounding SSC must reconcile *near-field contact geometry* and *far-field ring context*. **Principles.** We therefore combine *dual-projection encoders (DP-ER)*, *bi-grid voxelization (BGV)*, a lightweight *Gait Displacement Compensation (GDC)* before lifting, and a *Hierarchical AMoE-3D* decoder for scale-aware, anisotropic fusion.

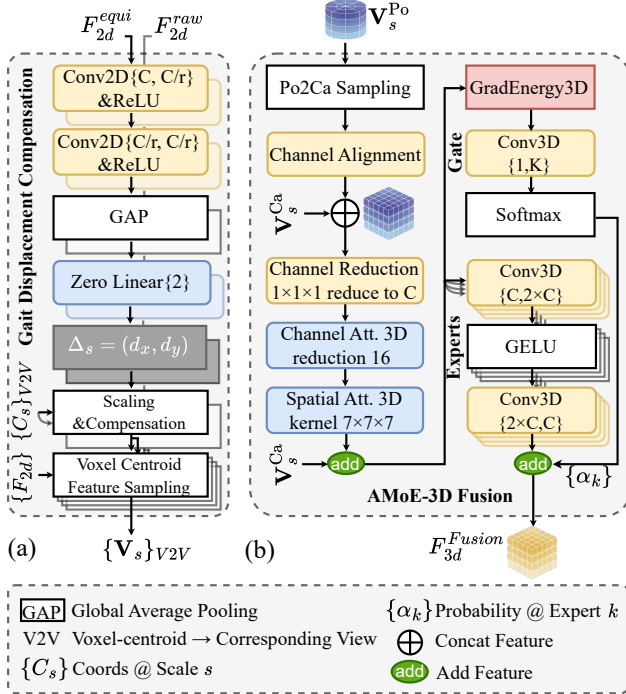


Figure 3. **GDC and AMoE-3D Fusion Module.** (a) GDC regresses $\Delta = (dx, dy)$ from 2D features via GAP+Linear (zero-init), applies scale/shift to multi-scale view coordinates $\{C_s\}$ for voxel-centroid sampling, and yields $\{F_{3d}\}_{V2V}$. (b) AMoE-3D fuses Po2Ca-sampled polar features with Cartesian ones via channel alignment, concatenation, and reduction, applies 3D channel- and spatial attention, and performs MoE-Fuse3D with K experts and GradEnergy3D-Softmax gating: $y = \sum_k P_k \odot E_k(x)$. The fused F_{3d}^{Fusion} feeds the 3D decoder.

3.3. Unwrapping and Dual-Projection Encoders

Motivation. Equirectangular views preserve *azimuthal continuity* that is convolution-friendly; the raw panoramic annulus retains *native geometry and fine textures*. Running both in parallel compensates the equator-pole trade-off.

Method. We calibrate the omnidirectional panoramic annular lens (PAL) camera with the Taylor polynomial model [95]. The unwrapping samples the raw annulus at coordinates given by the calibrated inverse mapping:

$$\mathbf{I}^{\text{equi}}(u, v) = \mathbf{I}^{\text{raw}}(\mathcal{U}^{-1}(u, v; \kappa)), \quad (1)$$

where $\kappa = \{a_i, u_0, v_0, \mathbf{A}\}$ are the Taylor coefficients and principal point plus a 2×2 affine matrix (scale/skew), and $(u, v) \in [0, W] \times [0, H]$. For the equirectangular ray with spherical angles (ϕ, θ) :

$$\phi = \frac{2\pi}{W} u - \pi, \quad \theta = \frac{\pi}{2} - \frac{\pi}{H} v, \quad (2)$$

$$r(\theta) = \sum_{i=0}^N a_i \theta^i, \quad \begin{bmatrix} u_{\text{raw}} \\ v_{\text{raw}} \end{bmatrix} = \begin{bmatrix} u_0 \\ v_0 \end{bmatrix} + \mathbf{A} r(\theta) \begin{bmatrix} \cos \phi \\ \sin \phi \end{bmatrix}. \quad (3)$$

Two encoders then produce multi-scale features:

$$\{\mathbf{F}_{1/s}^{\text{equi}}\}_{s \in \{1, 4, 8, 16\}}, \quad \{\mathbf{F}_{1/s}^{\text{raw}}\}_{s \in \{1, 4, 8, 16\}}, \quad (4)$$

with GroupNorm and 1×1 bottleneck alignment to standardize channels.

Why it helps. For legged robots, the equatorial ring dominates motion priors while polar areas are highly distorted. Using the Taylor unwrapping respects PAL optics, preserving *ring continuity* in ER space and *local textures* in the raw annulus, providing more stable cues for lifting.

3.4. Bi-Grid Voxelization and View2View Sampling

Motivation. Cartesian voxels are accurate for *near-field contact geometry* (footholds, obstacles), whereas polar/cylindrical voxels preserve *azimuthal continuity* that matches panoramic imaging. In a 360° camera, the equirectangular horizontal axis is linearly related to azimuth ϕ ; cylindrical voxels (r, φ, z) therefore form a *natural, camera-aligned* parameterization where $\varphi \approx \phi$ and r correlates with depth, yielding uniform sampling along the ring and reduced far-field aliasing. Combining Cartesian and cylindrical grids reduces quantization bias and balances near/far evidence under legged motion.

Method. We define voxel centroids

$$\begin{cases} \mathbf{c}_{ijk}^{\text{Ca}} = (x_i, y_j, z_k), \\ \mathbf{c}_{pqk}^{\text{Po}} = (r_p \cos \varphi_q, r_p \sin \varphi_q, z_k). \end{cases} \quad (5)$$

For view $v \in \{\text{equi}, \text{raw}\}$ and scale s , we lift features by bilinear sampling at the projected pixel $\pi_v(\cdot)$:

$$\begin{cases} \mathbf{V}_s^{\text{Ca}}(\mathbf{c}) = \text{bilinear}(\mathbf{F}_{1/s}^v, \pi_v(\mathbf{c}; \kappa)), \\ \mathbf{V}_s^{\text{Po}}(\mathbf{c}) = \text{bilinear}(\mathbf{F}_{1/s}^v, \pi_v(\mathbf{c}; \kappa)). \end{cases} \quad (6)$$

We fuse scales with per-voxel convex weights:

$$\begin{cases} \mathbf{V}^{\text{Ca}} = \sum_s \alpha_s^{\text{Ca}} \odot \mathbf{V}_s^{\text{Ca}}, & \sum_s \alpha_s^{\text{Ca}} = 1, \\ \mathbf{V}^{\text{Po}} = \sum_s \alpha_s^{\text{Po}} \odot \mathbf{V}_s^{\text{Po}}, & \sum_s \alpha_s^{\text{Po}} = 1. \end{cases} \quad (7)$$

To inject polar context into the Cartesian stream, we precompute cross-grid indices $\{\mathcal{J}_\ell\}_{\ell \in \{1, 2, 4\}}$ and form

$$\tilde{\mathbf{V}}_\ell^{\text{Ca}} = \text{Align}_{1 \times 1 \times 1}(\mathbf{V}_\ell^{\text{Po}}[\mathcal{J}_\ell]) \parallel \mathbf{V}_\ell^{\text{Ca}}. \quad (8)$$

Why it helps. On legged robots, near-field foothold safety leverages Cartesian precision while far-field loop/context exploits cylindrical efficiency. The camera-aligned polar stream improves *azimuthal coherence* and reduces spherical-distortion artifacts; the fused bi-grid volume balances near/far evidence under gait jitters.

Table 2. **Semantic scene completion results on the QuadOcc validation set.** All models are (re)trained by us on QuadOcc. **Bold** = best, underline = second best (within the LiDAR-based block; Vision-based block marked separately.)

Method	Input	vehicle (0.90%)	pedestrian (2.09%)	road (52.34%)	building (14.00%)	vegetation (22.82%)	terrain (7.85%)	Precision $\uparrow$	Recall $\uparrow$	IoU $\uparrow$	mIoU $\uparrow$
<i>LiDAR-based</i>											
SSCNet [24]	L	0.00	0.04	44.34	<u>16.06</u>	22.41	4.77	63.42	65.82	47.71	14.60
SSCNet-full [24]	L	0.71	0.04	<u>55.14</u>	14.93	19.66	7.76	<u>78.81</u>	60.03	<u>51.69</u>	16.38
LMSCNet [38]	L	<u>0.88</u>	<u>0.20</u>	57.02	16.45	<u>24.39</u>	11.70	79.70	<u>64.49</u>	55.40	18.44
OccRWKV [46]	L	2.05	0.35	54.17	10.29	24.92	<u>8.38</u>	71.98	62.80	50.47	<u>16.69</u>
<i>Vision-based</i>											
VoxFormer-S [51]	P+D	0.31	1.67	36.83	6.48	7.09	3.81	45.96	69.95	38.38	9.36
SGN [96]	P+D	7.39	1.88	53.35	9.06	21.95	9.39	59.23	67.41	46.05	17.17
VoxFormer-S [†] [51]	P+L	0.24	1.66	38.51	10.90	11.88	4.45	51.29	75.84	44.08	11.27
VoxFormer-T [†] [51]	P+L	0.30	1.65	40.21	14.48	15.86	4.64	56.94	77.10	48.70	12.86
SGN [†] [96]	P+L	<u>11.62</u>	<u>2.47</u>	53.06	<u>15.60</u>	<u>25.67</u>	9.91	73.70	59.62	49.16	<u>19.72</u>
OccFormer [52]	P	0.29	0.37	49.46	10.36	15.00	2.64	45.79	<u>76.99</u>	40.28	13.02
MonoScene [50]	P	8.15	1.59	55.66	12.88	26.10	10.78	62.10	69.28	48.69	19.19
OneOcc (Ours)	P	12.16	2.86	<u>54.41</u>	16.03	24.91	13.01	<u>66.69</u>	64.74	<u>48.92</u>	20.56

[†] Uses a different sensor modality from the original paper (here: *adding LiDAR*).

Input legend: L = LiDAR; P = *Panoramic* camera; D = monocular Depth (pred.).

3.5. Gait Displacement Compensation

Motivation. Gait shocks induce phase errors; correcting them *after* lifting acts on already voxel-quantized evidence. Compensating *before* lifting—at sampling coordinates—avoids quantization and is cheaper.

Method. For each scale and projection, we regress a 2D displacement $\Delta_s = (d_x, d_y)$ with a zero-initialized head:

$$\Delta_s = \text{Linear}_0(\text{GAP}(\mathbf{F}_{1/s}^v)), \quad (9)$$

where Linear_0 has all weights and bias set to zero so the initial warp is identity (cf. zero-conv [97]). Before lifting we correct and resample:

$$\hat{\mathbf{p}} = \pi_v(\mathbf{c}; \boldsymbol{\kappa}) + \Delta_s, \quad (10)$$

$$\mathbf{V}_s^{(\cdot)}(\mathbf{c}) = \text{bilinear}(\mathbf{F}_{1/s}^v, \hat{\mathbf{p}}). \quad (11)$$

In practice, this also upgrades the lifting operator from integer-index gather [50] to bilinear sampling, which reduces projection aliasing and further stabilizes GDC. Here, $v \in \{\text{unwrapped, spherical}\}$ indexes the projection path; $\pi_v(\mathbf{c}; \boldsymbol{\kappa})$ projects a 3D voxel centroid $\mathbf{c}$ to image pixels using the calibrated Taylor camera with parameters $\boldsymbol{\kappa}$. $\hat{\mathbf{p}}$ is the corrected sampling pixel; $\text{bilinear}(\cdot)$ samples the 2D feature map $\mathbf{F}_{1/s}^v$, and $\mathbf{V}_s^{(\cdot)}$ denotes the lifted 3D feature.

Why it helps. GDC routes the *phase error* back to 2D, preventing mis-attributed voxels. The zero-init head stabilizes early training and leaves features unperturbed when motion is tiny ($\Delta_s \approx 0$). We additionally provide qualitative analysis of GDC in the supplementary material.

3.6. Hierarchical AMoE-3D

Motivation. Panoramic scenes are *anisotropic*: strong azimuthal variation vs. vertical structure, and large near/far scale disparity. We thus (i) inject polar cues hierarchically (coarse-to-fine), and (ii) apply *voxel-wise expert selection* to prevent over-smoothing in flat regions while enhancing edges/contacts critical for locomotion.

Method. The 3D decoder is a depthwise-separable UNet with trilinear upsampling at three levels (L1/L2/L3). At each level we fuse $\tilde{\mathbf{V}}^{\text{Ca}}$ using a *dual-path volumetric saliency* and a *Mixture-of-Experts* (MoE) fuse; we term this module *AMoE-3D*, where *A* denotes attention implemented via the dual-path (channel & spatial) saliency gates.

Dual-path volumetric saliency (channel & spatial gates):

$$\begin{cases} \mathbf{A}_c = \sigma(\text{MLP}(\text{GAP}(\mathbf{X})) + \text{MLP}(\text{GMP}(\mathbf{X}))), \\ \mathbf{A}_s = \sigma(g^{7 \times 7 \times 7}([\text{Avg}(\mathbf{X}); \text{Max}(\mathbf{X})])), \\ \mathbf{Y} = \mathbf{X} \odot \mathbf{A}_c \odot \mathbf{A}_s. \end{cases} \quad (12)$$

Table 3. **Semantic scene completion using Panoramas on H3O (Human360Occ)**. Two vertically-stacked panels report results on **HOMO (within-city)** and **HETER (cross-city)** splits. Header frequencies are computed on *non-empty* voxels. **Input legend:** P = Panoramic camera; $D_{pred.}$ = predicted monocular depth; D_{gt} = ground-truth depth.

HOMO (within-city)															
Method	Input	road (34.04%)	sidewalk (26.24%)	building (10.00%)	vegetation (24.02%)	car (2.92%)	truck (0.99%)	bus (0.25%)	two-wheeler (0.07%)	person (0.97%)	pole (0.50%)	Precision	Recall	IoU	mIoU
VoxFormer-S [51]	$P+D_{pred.}$	22.27	17.44	2.86	5.65	0.87	0.14	0.00	0.00	61.46	0.18	40.21	59.49	31.57	11.09
VoxFormer-S [51]	$P+D_{gt}$	35.84	25.86	12.32	9.90	3.73	0.70	0.19	0.17	64.79	0.93	59.91	71.12	48.19	15.44
SGN-T [96]	$P+D_{pred.}$	54.55	44.61	15.51	35.16	23.91	7.68	4.52	15.77	65.24	19.46	61.82	64.36	46.05	28.64
SGN-S [96]	$P+D_{pred.}$	54.81	47.44	15.64	33.22	24.23	8.53	2.62	13.51	65.88	19.72	62.28	63.72	45.98	28.56
OccFormer [52]	P	54.11	46.41	15.14	32.84	17.57	4.30	0.00	1.65	65.08	11.72	57.30	69.94	45.98	24.88
MonoScene [50]	P	62.45	57.18	19.97	38.91	29.76	13.77	5.03	16.76	70.15	20.67	68.78	71.66	54.08	33.46
OneOcc (Ours)	P	63.82	62.92	22.86	41.03	33.74	17.79	6.47	26.58	71.13	26.55	69.14	74.18	55.73	37.29

HETER (cross-city)															
Method	Input	road (34.04%)	sidewalk (26.24%)	building (10.00%)	vegetation (24.02%)	car (2.92%)	truck (0.99%)	bus (0.25%)	two-wheeler (0.07%)	person (0.97%)	pole (0.50%)	Precision	Recall	IoU	mIoU
VoxFormer-S [51]	$P+D_{pred.}$	19.02	15.82	2.84	5.86	0.95	0.10	0.00	0.01	61.62	0.10	36.68	52.00	27.40	10.63
VoxFormer-S [51]	$P+D_{gt}$	32.55	25.95	8.98	9.93	4.68	0.44	0.00	0.00	65.45	0.26	59.08	61.21	42.99	14.82
SGN-T [96]	$P+D_{pred.}$	42.57	23.94	9.42	15.84	18.22	10.21	1.55	3.11	63.44	11.93	45.84	47.11	30.26	20.02
SGN-S [96]	$P+D_{pred.}$	43.47	24.70	9.19	20.27	17.21	8.21	1.85	5.08	61.10	14.89	47.82	50.98	32.76	20.60
OccFormer [52]	P	47.00	39.01	10.14	24.85	11.84	1.52	0.05	0.77	66.60	6.93	53.12	63.57	40.73	20.87
MonoScene [50]	P	49.68	41.41	11.04	20.57	20.34	10.36	1.98	2.79	69.68	13.69	67.39	55.00	43.44	24.15
OneOcc (Ours)	P	58.60	48.28	16.34	25.98	30.68	19.35	12.87	14.16	72.89	23.11	67.89	64.77	49.58	32.23

MoE-Fuse3D with gradient-energy gating:

$$\begin{cases} \text{GradEnergy3D}(\mathbf{Y}) = \sum_{a \in \{x, y, z\}} \|\nabla_a \mathbf{Y}\|_2^2, \\ \alpha = \text{softmax}(W_g * \text{GradEnergy3D}(\mathbf{Y})), \\ \tilde{\mathbf{Y}} = \mathbf{Y} + \sum_{k=1}^K \alpha_k E_k(\mathbf{Y}), \end{cases} \quad (13)$$

where E_k are $1 \times 1 \times 1$ Conv-GELU-Conv experts. Here, $\mathbf{X} \in \mathbb{R}^{C \times D \times H \times W}$ is the level- ℓ Cartesian feature and $\tilde{\mathbf{V}}_{Ca}$ is the polar volume resampled onto the same Cartesian grid; $\mathbf{A}_c \in \mathbb{R}^{C \times 1 \times 1 \times 1}$ and $\mathbf{A}_s \in \mathbb{R}^{1 \times D \times H \times W}$ are the channel and spatial gates. ∇_a denotes the discrete 3D gradient along axis $a \in \{x, y, z\}$, $*$ is 3D convolution, and $[\cdot]$ concatenates along channels.

Why it helps. Hierarchical polar injections aggregate *far-field azimuthal cues* at coarse levels—e.g., ring-consistent *road/building/vegetation* context—and refine *near-field contact geometry* at fine levels. Gradient-energy gating em-

phasizes high-contrast structures at category transitions (vehicles, poles) while suppressing overfitting on large ground flats (road), improving stability of foothold decisions.

3.7. Segmentation Head and Loss

Method. A $1 \times 1 \times 1$ head outputs per-voxel logits $\mathbf{Z}$. We supervise only *valid* voxels (from lifting visibility) with deep supervision at strides $\{1, 2, 4\}$.

Loss. We follow MonoScene [50] for most losses: (i) standard cross-entropy on valid voxels, (ii) the scene-class affinity (SCAL) terms for semantics and geometry, and (iii) the frustums proportion (FP) loss. We do *not* use the relation loss, as we observed it encourages co-occurrence priors that over-smooth azimuthal boundaries and suppress small near-field classes under panoramic legged setting, yielding no gain and occasional degradation on our data. The final objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{SCAL}}^{\text{sem}} + \mathcal{L}_{\text{SCAL}}^{\text{geo}} + \mathcal{L}_{\text{FP}}. \quad (14)$$

Class re-weighting for $\mathcal{L}_{\text{CE}}$ follows standard SSC practice.

Table 4. **Ablations on QuadOcc (Panorama only)**. Core columns show if a component is enabled (✓) or disabled (✗).

Variant	Core Designs					Per-class IoU (QuadOcc)						Overall			
	GDC Grid Disp. Compensation	DP-ER Dual Projection (Egnt. + Row)	BGV Bi-Grid Voxelization	AMoE-3D Attention-MoE 3D Fusion		vehicle (0.90%)	pedestrian (2.09%)	road (52.34%)	building (14.00%)	vegetation (22.82%)	terrain (7.85%)	SC Precision	SC Recall	SC IoU	SSC mIoU
(Q0) baseline	✗	✗	✗	✗		8.15	1.59	55.66	12.88	26.10	10.78	62.10	69.28	48.69	19.19
(Q1) + GDC	✓	✗	✗	✗		11.76	2.48	53.73	15.07	22.95	11.51	64.05	66.08	48.78	19.58
(Q2) + DP-ER	✓	✓	✗	✗		11.82	2.51	54.16	15.39	23.40	12.08	65.42	65.01	48.38	19.89
(Q3) + BGV	✓	✓	✓	✗		<u>12.07</u>	<u>2.77</u>	54.29	<u>15.76</u>	24.19	<u>12.74</u>	<u>66.56</u>	64.70	48.68	<u>20.30</u>
(Q4) + AMoE-3D	✓	✓	✓	✓		12.16	2.86	<u>54.41</u>	16.03	<u>24.91</u>	13.01	66.69	64.74	48.92	20.56

Table 5. **Performance comparison on the QuadOcc validation set across different lighting conditions**. Results are shown for day, dusk, and night scenarios. P (Precision), R (Recall), IoU and mIoU are reported. Methods marked with † use a sensor modality different from the original paper. The best performance for each input modality is highlighted in **bold**.

Method	Day ☀				Dusk 🌄				Night 🌃			
	P	R	IoU	mIoU	P	R	IoU	mIoU	P	R	IoU	mIoU
<i>LiDAR-based</i>												
SSCNet [24]	67.12	62.95	48.12	14.03	52.97	76.34	45.50	14.88	57.81	75.30	48.60	9.51
SSCNet-full [24]	80.56	58.52	51.28	16.04	73.79	64.34	52.37	17.14	73.64	67.31	54.24	11.56
LMSCNet [38]	81.18	62.50	54.59	17.33	75.34	70.92	57.56	18.91	75.68	72.67	58.91	13.40
OccRWKV [46]	72.84	63.83	51.56	16.25	66.43	59.30	45.63	15.42	74.10	58.31	48.44	12.61
<i>Vision-based</i>												
OccFormer [52]	46.23	77.08	40.64	12.82	41.79	79.45	37.72	14.46	49.49	72.18	41.56	10.21
VoxFormer-S† [51]	55.80	73.96	46.64	11.44	37.42	82.81	34.73	8.95	47.10	82.41	42.80	11.84
VoxFormer-T† [51]	61.77	75.02	51.24	12.91	42.25	84.68	39.25	11.06	51.93	84.62	47.45	12.88
SGN [96]	61.54	68.71	48.06	17.94	54.07	69.06	43.53	15.19	46.83	52.29	32.81	10.58
MonoScene [50]	65.47	69.94	51.09	18.58	47.62	77.10	41.72	15.14	51.35	66.47	40.78	14.20
OneOcc (Ours)	69.10	64.66	50.15	21.15	57.46	71.98	46.96	19.86	63.56	53.75	41.09	13.50

4. Experiments

4.1. Setup

Datasets. We evaluate on two panoramic occupancy benchmarks for legged/humanoid platforms. **QuadOcc**: real first-person 360° data on a quadruped within a campus; day/dusk/night, sequence-heterogeneous split; 24K *consecutive* frames with *stride-5* training; 6 classes on a 64×64×8 grid (0.4 m). **Human360Occ (H3O)**: CARLA-based human-ego 360° with simulated gait over 16 maps and diverse weather/lighting; RGB, depth, occupancy (two resolutions), and poses. We report *within-city* (H3O-Homo) and *cross-city* (H3O-Heter). See supplementary for the data collection pipelines and statistics of QuadOcc and H3O.

Metrics. Following semantic scene completion (SSC) practice, we report per-class IoU and mean IoU (mIoU) on non-empty voxels, plus overall precision (P) and recall (R).

Baselines. We re-train popular *LiDAR* SSC systems (SSCNet [24], LMSCNet [38], OccRWKV [46]) and adapt representative *vision* SSC methods (OccFormer [52], VoxFormer [51], SGN [96], MonoScene [50]) to panoramas via calibrated unwrapping to equirectangular. For *P+L* entries in Table 2, we add LiDAR to the corresponding vision model per the table footnote. Details in the supplementary.

Implementation. Unless noted, training follows official configs with minimal dataset-specific changes; schedules/image sizes/voxelization match for fairness. On H3O we disable the raw-projection branch (native equirectangu-

lar), keeping components as in ablations. Training uses a single RTX-4090 GPU. Details in the supplementary.

4.2. Quantitatively Comparisons

QuadOcc. Table 2 reports *QuadOcc (val)*. OneOcc attains 20.56 mIoU, surpassing the best LiDAR baseline LMSCNet (18.44) and the strongest vision baseline MonoScene (19.19). SGN† (P+L) achieves the highest geometry IoU (49.16) among vision entries by leveraging LiDAR, yet camera-only OneOcc delivers the best *mIoU* (20.56) among vision methods. Task-aligned panoramic fusion (DP-ER, BGV, AMoE-3D, with GDC on legged data) enables a camera-only pipeline to rival and even surpass popular LiDAR stacks at this range/resolution.

Human360Occ (H3O). Table 3 summarizes generalization: On H3O-Homo OneOcc reaches 37.29 mIoU (+3.83 vs. MonoScene), while on H3O-Heter it attains 32.23 mIoU (+8.08). The larger margin under *heterogeneous* cross-city shifts highlights OneOcc’s strong distribution robustness: DP-ER and AMoE-3D transfer across maps and lighting/weather, and their distortion-aware priors mitigate panoramic aliasing, yielding state-of-the-art camera-only mIoU under distribution shift.

Robustness to Lighting Conditions. Table 5 reports day/dusk/night. OneOcc leads the best vision baseline in day (21.15 vs. 18.58) and dusk (19.86 vs. 15.14); at night its mIoU (13.50) trails MonoScene (14.20) but shows *higher precision*, likely from frustum-artifact suppression (cf. Fig. 4). *LiDAR trend.* Dusk often exceeds day due to weaker solar background in near-IR, while night drops with altered return statistics and sparser echoes on low-albedo/specular surfaces, hurting long-range completion.

4.3. Ablations

We ablate GDC, DP-ER, BGV, AMoE-3D (Table 4). Starting from Q0 (ER-only, no GDC, single-grid, no AMoE-3D; 19.19 mIoU), we add: +GDC → 19.58 (+0.39), +DP-ER → 19.89 (+0.31), +BGV → 20.30 (+0.41), +AMoE-3D (full) → 20.56 (+0.26). Gains are additive: GDC stabilizes sampling; DP-ER provides complementary cues; BGV reduces discretization bias; AMoE-3D sharpens edges/contacts without over-smoothing flats. We further provide more detailed per-module ablations, calibration-robustness analysis, temporal aggregation baselines, and resolution scaling results in the supplementary material.

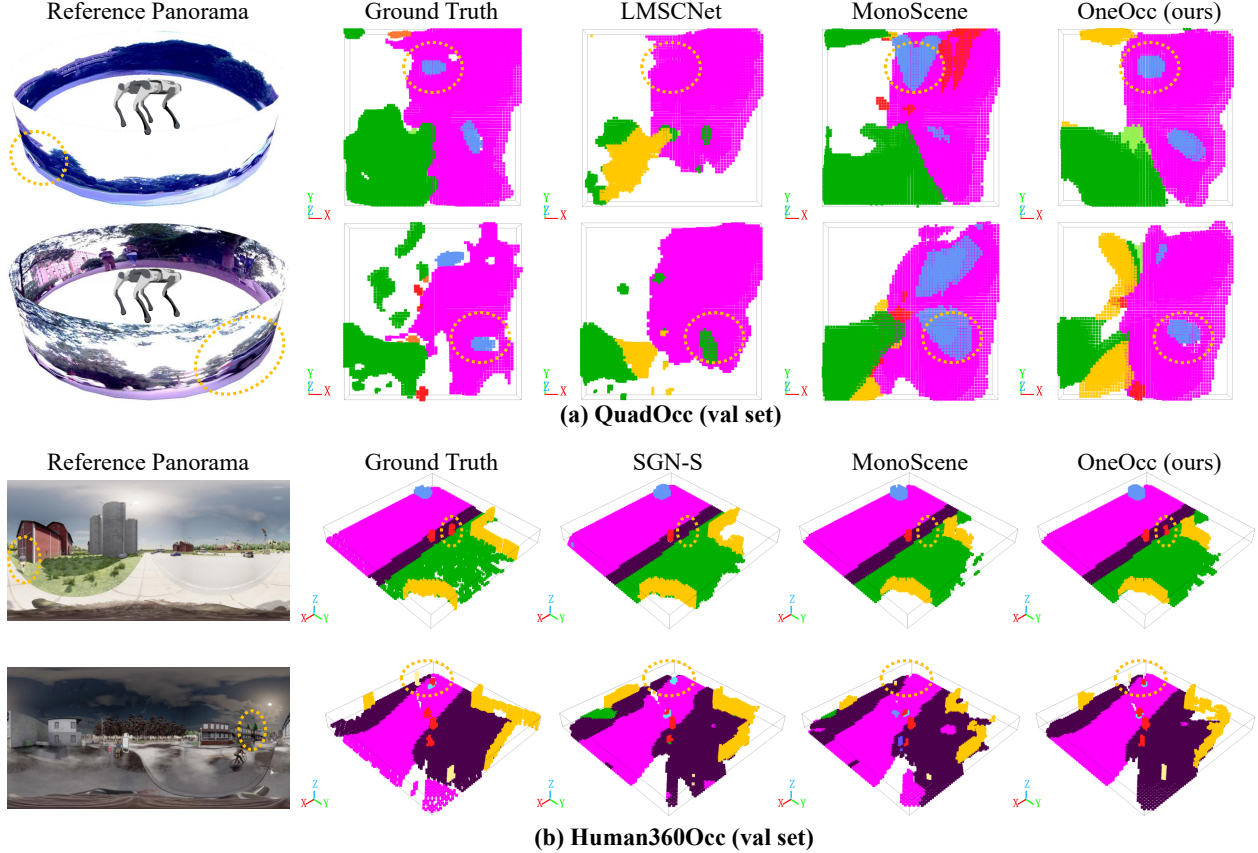


Figure 4. **Qualitative comparisons on (a) *QuadOcc* and (b) *Human360Occ* (val sets).** Left to right: Reference panorama, ground-truth occupancy, baselines, and OneOcc (ours). Across both real (*QuadOcc*) and simulated (*H3O*) panoramic settings, OneOcc better preserves the global layout and 360° continuity, produces cleaner class boundaries, and recovers occluded structures. Yellow dotted circles highlight typical improvements over LMSCNet/SGN-S and MonoScene (e.g., foreground-background disambiguation and far-range layout).

4.4. Qualitative Comparisons

Fig. 4 shows typical cases. On *QuadOcc*, MonoScene exhibits frustum-direction artifacts (ghost elongation), while OneOcc suppresses them via *DP-ER* (cross-view corroboration), *GDC* (phase-stable sampling), and *BGV* (near/far balance). On *Human360Occ*, OneOcc recovers pedestrians close in color to background that competing vision methods miss. Overall, OneOcc preserves global layout and 360° continuity and better handles occlusions.

4.5. Efficiency

Inference on an RTX 4090 at 608×1216 , batch size 1. OneOcc averages 69.93 ms (FPS ≈ 14.30). The model has 101.76M parameters and peaks at 1.82 GB CUDA allocation (2.58 GB reserved) in FP32. With mixed precision (FP32+FP16), latency drops to 52.84 ms (FPS ≈ 18.92), and peak allocation is 1.49 GB. The footprint remains moderate, making OneOcc suitable for onboard legged perception on a single GPU. We additionally report measured on-device runtime and a module-wise latency breakdown on NVIDIA Jetson AGX Orin in the supplementary material.

5. Conclusion

We introduced *OneOcc*, a panoramic *camera-only* framework for semantic occupancy on legged and humanoid platforms. OneOcc stabilizes feature phases under periodic motion via *Gait Displacement Compensation* (GDC), fuses raw annular and equirectangular streams with *Dual-Projection* (DP-ER), balances near-far geometry through *Bi-Grid Voxelization* (BGV), and aggregates multi-scale evidence using a hierarchical *AMoE-3D* decoder. Together, these components turn monocular panoramas into a geometry-aware 3D grid. We also contribute *QuadOcc* (real 360° quadruped) and report cross-domain evaluation on *Human360Occ*, facilitating systematic study across robot morphologies and environments.

Limitations & Outlook. OneOcc assumes accurate calibration with bounded drift. A possible remedy is online extrinsic self-calibration, optionally regularized by lightweight odometry priors. We propose *robotic occupancy* as an intermediate layer for world models and vision-language-action models: tokens for language control, and occupancy-sequence pretraining for cross-robot transfer.

Acknowledgment

This research was funded by the Natural Science Foundation of Zhejiang Province (Grant No. LZ24F050003), the National Natural Science Foundation of China (Grant Nos. 12174341 and 62473139), the Hunan Provincial Research and Development Project (Grant No. 2025QK3019), the MOE AcRF Tier 1 SSHR-TG Incubator Grant FY24 (Grant No. RSTG7/24), and the opening project of the State Key Laboratory of Autonomous Intelligent Unmanned Systems (Grant No. ZZKF2025-2-10). Hao Shi was also supported by the China Scholarship Council (CSC) for joint Ph.D. training at Nanyang Technological University (NTU).

References

- [1] Junming Wang, Xiuxian Guan, Zekai Sun, Tianxiang Shen, Dong Huang, Fangming Liu, and Heming Cui. OMEGA: Efficient occlusion-aware navigation for air-ground robots in dynamic environments via state space model. *IEEE Robotics and Automation Letters*, 2025. 1
- [2] David Hoeller, Nikita Rudin, Christopher Choy, Animashree Anandkumar, and Marco Hutter. Neural scene representation for locomotion on structured terrain. *IEEE Robotics and Automation Letters*, 2022.
- [3] Junming Wang, Zekai Sun, Xiuxian Guan, Tianxiang Shen, Dong Huang, Zongyuan Zhang, Tianyang Duan, Fangming Liu, and Heming Cui. HE-Nav: A high-performance and efficient navigation system for aerial-ground robots in cluttered environments. *IEEE Robotics and Automation Letters*, 2024.
- [4] Junming Wang, Zekai Sun, Xiuxian Guan, Tianxiang Shen, Zongyuan Zhang, Tianyang Duan, Dong Huang, Shixiong Zhao, and Heming Cui. AGRNav: Efficient and energy-saving autonomous navigation for air-ground robots in occlusion-prone environments. In *ICRA*, 2024.
- [5] Akash Patel, Mario Alberto Valdes Saucedo, Nikolaos Stathouloupoulos, Viswa Narayanan Sankaranarayanan, Ilias Tevetzidis, Christoforos Kanellakis, and George Nikolakopoulos. A hierarchical graph-based terrain-aware autonomous navigation approach for complementary multimodal ground-aerial exploration. *arXiv preprint arXiv:2505.14859*, 2025.
- [6] Hongyu Li, Taşkın Padır, and Huaizu Jiang. StereoNavNet: Learning to navigate using stereo cameras with auxiliary occupancy voxels. In *IROS*, 2024. 1
- [7] Shun Nijima, Ryoichi Tsuzaki, Noriaki Takasugi, and Masaya Kinoshita. Real-time multi-plane segmentation based on GPU accelerated high-resolution 3D voxel mapping for legged robot locomotion. *arXiv preprint arXiv:2510.01592*, 2025. 1
- [8] Yinzhaodong, Ji Ma, Liu Zhao, Wanyue Li, and Peng Lu. MARG: Mastering risky gap terrains for legged robots with elevation mapping. *IEEE Transactions on Robotics*, 2025.
- [9] Alejandro Escontrela, Justin Kerr, Arthur Allshire, Jonas Frey, Rocky Duan, Carmelo Sferrazza, and Pieter Abbeel. GaussGym: An open-source real-to-sim framework for learning locomotion from pixels. *arXiv preprint arXiv:2510.15352*, 2025.
- [10] Yeke Chen, Ji Ma, Zeren Luo, Yimin Han, Yinzhaodong, Bowen Xu, and Peng Lu. Learning autonomous and safe quadruped traversal of complex terrains using multi-layer elevation maps. *IEEE Robotics and Automation Letters*, 2025. 1
- [11] Rainer Kartmann, Danqing Liu, and Tamim Asfour. Semantic scene manipulation based on 3D spatial object relations and language instructions. In *Humanoids*, 2021. 1
- [12] Nils Dengler, Jesper Mücke, Rohit Menon, and Maren Bennewitz. Efficient manipulation-enhanced semantic mapping with uncertainty-informed action selection. In *Humanoids*, 2025.
- [13] Ying Li, Xiaobao Wei, Xiaowei Chi, Yuming Li, Zhongyu Zhao, Hao Wang, Ningning Ma, Ming Lu, and Shanghang Zhang. ManipDreamer3D: Synthesizing plausible robotic manipulation video with occupancy-aware 3D trajectory. *arXiv preprint arXiv:2509.05314*, 2025.
- [14] Kaijun Wang, Liqin Lu, Mingyu Liu, Jianuo Jiang, Zeju Li, Bolin Zhang, Wancai Zheng, Xinyi Yu, Hao Chen, and Chunhua Shen. ODYSSEY: Open-world quadrupeds exploration and manipulation for long-horizon tasks. *arXiv preprint arXiv:2508.08240*, 2025. 1
- [15] Yuqi Wu, Wenzhao Zheng, Sicheng Zuo, Yuanhui Huang, Jie Zhou, and Jiwen Lu. EmbodiedOcc: Embodied 3D occupancy prediction for vision-based online scene understanding. In *ICCV*, 2025. 1
- [16] Hao Wang, Xiaobao Wei, Xiaohan Zhang, Jianing Li, Chengyu Bai, Ying Li, Ming Lu, Wenzhao Zheng, and Shanghang Zhang. EmbodiedOcc++: Boosting embodied 3D occupancy prediction with plane regularization and uncertainty sampler. In *MM*, 2025.
- [17] Xinpeng Liu, Haowen Hou, Yanchao Yang, Yong-Lu Li, and Cewu Lu. Revisit human-scene interaction via space occupancy. In *ECCV*, 2024. 1
- [18] Quan Lai, Haifeng Zheng, Xinxin Feng, Minghui Zheng, Huacong Chen, and Wenqiang Chen. RTONet: Real-time occupancy network for semantic scene completion. *IEEE Robotics and Automation Letters*, 2024. 1
- [19] Helin Cao and Sven Behnke. SLCF-Net: Sequential LiDAR-camera fusion for semantic scene completion using a 3D recurrent U-Net. In *ICRA*, 2024.
- [20] Yuxiang Yan, Boda Liu, Jianfei Ai, Qinbu Li, Ru Wan, and Jian Pu. PointSSC: A cooperative vehicle-infrastructure point cloud benchmark for semantic scene completion. In *ICRA*, 2024.
- [21] Zhang Zhang, Qiang Zhang, Wei Cui, Shuai Shi, Yijie Guo, Gang Han, Wen Zhao, Hengle Ren, Renjing Xu, and Jian Tang. RoboOcc: Enhancing the geometric and semantic scene understanding for robots. *arXiv preprint arXiv:2504.14604*, 2025. 1, 3
- [22] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. 2
- [23] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scan-

- Net: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017. 2
- [24] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas A. Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2017. 2, 5, 7
- [25] Yancheng Pan, Biao Gao, Jilin Mei, Sibao Geng, Chengkun Li, and Huijing Zhao. SemanticPOSS: A point cloud dataset with large quantity of dynamic instances. In *IV*, 2020. 2
- [26] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jürgen Gall. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In *ICCV*, 2019. 2
- [27] 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. 2
- [28] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3D: A large-scale 3D occupancy prediction benchmark for autonomous driving. In *NeurIPS*, 2023. 2
- [29] 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. SSCBench: A large-scale 3D semantic scene completion benchmark for autonomous driving. In *IROS*, 2024. 2
- [30] Song Wang, Jiawei Yu, Wentong Li, Hao Shi, Kailun Yang, Junbo Chen, and Jianke Zhu. Label-efficient semantic scene completion with scribble annotations. In *IJCAI*, 2024. 2
- [31] Lianqing Zheng, Long Yang, Qunshu Lin, Wenjin Ai, Minghao Liu, Shouyi Lu, Jianan Liu, Hongze Ren, Jingyue Mo, Xiaokai Bai, Jie Bai, Zhixiong Ma, and Xichan Zhu. OmniHD-Scenes: A next-generation multimodal dataset for autonomous driving. *arXiv preprint arXiv:2412.10734*, 2024. 2
- [32] Heng Zhai, Jilin Mei, Chen Min, Liang Chen, Fangzhou Zhao, and Yu Hu. WildOcc: A benchmark for off-road 3D semantic occupancy prediction. *arXiv preprint arXiv:2410.15792*, 2024. 2
- [33] Shangwei Guo, Hao Shi, Song Wang, Xiaoting Yin, Kailun Yang, and Kaiwei Wang. Event-aided semantic scene completion. *arXiv preprint arXiv:2502.02334*, 2025. 2, 3
- [34] Chen Min, Jilin Mei, Heng Zhai, Shuai Wang, Tong Sun, Fanjie Kong, Haoyang Li, Fangyuan Mao, Fuyang Liu, Shuo Wang, Yiming Nie, Qi Zhu, Liang Xiao, Dawei Zhao, and Yu Hu. Advancing off-road autonomous driving: The large-scale ORAD-3D dataset and comprehensive benchmarks. *arXiv preprint arXiv:2510.16500*, 2025. 2
- [35] Hanlin Wu, Pengfei Lin, Ehsan Javanmardi, Naren Bao, Bo Qian, Hao Si, and Manabu Tsukada. A synthetic benchmark for collaborative 3D semantic occupancy prediction in V2X autonomous driving. *arXiv preprint arXiv:2506.17004*, 2025. 2
- [36] Xin Lin, Xian Ge, Dizhe Zhang, Zhaoliang Wan, Xianshun Wang, Xiangtai Li, Wenjie Jiang, Bo Du, Dacheng Tao, Ming-Hsuan Yang, and Lu Qi. One flight over the gap: A survey from perspective to panoramic vision. *arXiv preprint arXiv:2509.04444*, 2025. 1
- [37] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2
- [38] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. LMSCNet: Lightweight multiscale 3D semantic completion. In *3DV*, 2020. 2, 5, 7
- [39] 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. 2
- [40] 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. 2
- [41] Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. PointOcc: Cylindrical tri-perspective view for point-based 3D semantic occupancy prediction. *arXiv preprint arXiv:2308.16896*, 2023. 2
- [42] Hui Zhou, Xinge Zhu, Xiao Song, Yuexin Ma, Zhe Wang, Hongsheng Li, and Dahua Lin. Cylinder3D: An effective 3D framework for driving-scene LiDAR semantic segmentation. *arXiv preprint arXiv:2008.01550*, 2020. 2
- [43] 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. 2
- [44] Lianqing Zhao, Xiuwei Xu, Ziwei Wang, Yunpeng Zhang, Borui Zhang, Wenzhao Zheng, Dalong Du, Jie Zhou, and Jiwen Lu. LowRankOcc: Tensor decomposition and low-rank recovery for vision-based 3D semantic occupancy prediction. In *CVPR*, 2024. 2
- [45] Yukai Ma, Jianbiao Mei, Xuemeng Yang, Licheng Wen, Weihua Xu, Jiangning Zhang, Xingxing Zuo, Botian Shi, and Yong Liu. LiCROcc: Teach radar for accurate semantic occupancy prediction using LiDAR and camera. *IEEE Robotics and Automation Letters*, 2025. 2, 3
- [46] Junming Wang, Wei Yin, Xiaoxiao Long, Xingyu Zhang, Zebin Xing, Xiaoyang Guo, and Qian Zhang. OccRWKV: Rethinking efficient 3D semantic occupancy prediction with linear complexity. In *ICRA*, 2025. 5, 7
- [47] Xuemeng Yang, Hao Zou, Xin Kong, Tianxin Huang, Yong Liu, Wanlong Li, Feng Wen, and Hongbo Zhang. Semantic segmentation-assisted scene completion for LiDAR point clouds. In *IROS*, 2021.
- [48] Hao Zou, Xuemeng Yang, Tianxin Huang, Chujuan Zhang, Yong Liu, Wanlong Li, Feng Wen, and Hongbo Zhang. Up-to-down network: Fusing multi-scale context for 3D semantic scene completion. In *IROS*, 2021.
- [49] 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. 2

- [50] Anh-Quan Cao and Raoul de Charette. MonoScene: Monocular 3D semantic scene completion. In *CVPR*, 2022. 2, 5, 6, 7
- [51] Yiming Li, Zhiding Yu, Christopher B. Choy, Chaowei Xiao, José M. Álvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. VoxFormer: Sparse voxel transformer for camera-based 3D semantic scene completion. In *CVPR*, 2023. 3, 5, 6, 7
- [52] Yunpeng Zhang, Zheng Zhu, and Dalong Du. OccFormer: Dual-path transformer for vision-based 3D semantic occupancy prediction. In *ICCV*, 2023. 3, 5, 6, 7
- [53] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. RenderOcc: Vision-centric 3D occupancy prediction with 2D rendering supervision. In *ICRA*, 2024. 3
- [54] Zhenxing Ming, Julie Stephany Berrio, Mao Shan, and Stewart Worrall. InverseMatrixVT3D: An efficient projection matrix-based approach for 3D occupancy prediction. In *IROS*, 2024. 3
- [55] Rodrigo Marcuzzi, Lucas Nunes, Elias Marks, Louis Wiesmann, Thomas Läbe, Jens Behley, and Cyrill Stachniss. Sf-mOcc: Vision-based 3D semantic occupancy prediction in urban environments. *IEEE Robotics and Automation Letters*, 2025. 3
- [56] Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. SparseOcc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In *CVPR*, 2024. 3
- [57] Yujie Xue, Ruihui Li, Fan Wu, Zhuo Tang, Kenli Li, and Mingxing Duan. Bi-SSC: Geometric-semantic bidirectional fusion for camera-based 3D semantic scene completion. In *CVPR*, 2024.
- [58] Aleksandar Jevtić, Christoph Reich, Felix Wimbauer, Oliver Hahn, Christian Rupprecht, Stefan Roth, and Daniel Cremers. Feed-forward SceneDINO for unsupervised semantic scene completion. In *ICCV*, 2025. 3
- [59] Enyu Liu, En Yu, Sijia Chen, and Wenbing Tao. Disentangling instance and scene contexts for 3D semantic scene completion. In *ICCV*, 2025. 3
- [60] Dubing Chen, Huan Zheng, Yucheng Zhou, Xianfei Li, Wenlong Liao, Tao He, Pai Peng, and Jianbing Shen. Semantic causality-aware vision-based 3D occupancy prediction. In *ICCV*, 2025. 3
- [61] Hyo-Jun Lee, Yeong Jun Koh, Hanul Kim, Hyunseop Kim, Yonguk Lee, and Jinu Lee. SOAP: Vision-centric 3D semantic scene completion with scene-adaptive decoder and occluded region-aware view projection. In *CVPR*, 2025. 3
- [62] Haoang Lu, Yuanqi Su, Xiaoning Zhang, Longjun Gao, Yu Xue, and Le Wang. VisHall3D: Monocular semantic scene completion from reconstructing the visible regions to hallucinating the invisible regions. In *ICCV*, 2025. 3
- [63] Hao Shi, Song Wang, Jiaming Zhang, Xiaoting Yin, Guangming Wang, Jianke Zhu, Kailun Yang, and Kaiwei Wang. Offboard occupancy refinement with hybrid propagation for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2025. 3
- [64] Hyun-Kurl Jang, Jihun Kim, Hyeokjun Kweon, and Kuk-Jin Yoon. TALoS: Enhancing semantic scene completion via test-time adaptation on the line of sight. In *NeurIPS*, 2024. 3
- [65] Wanshui Gan, Fang Liu, Hongbin Xu, Ningkai Mo, and Naoto Yokoya. GaussianOcc: Fully self-supervised and efficient 3D occupancy estimation with gaussian splatting. *arXiv preprint arXiv:2408.11447*, 2024. 3
- [66] Simon Boeder, Fabian Gigengack, and Benjamin Risse. GaussianFlowOcc: Sparse and weakly supervised occupancy estimation using gaussian splatting and temporal flow. *arXiv preprint arXiv:2502.17288*, 2025. 3
- [67] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. GaussianFormer: Scene as gaussians for vision-based 3D semantic occupancy prediction. In *ECCV*, 2024. 3
- [68] Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. GaussianFormer-2: Probabilistic gaussian superposition for efficient 3D occupancy prediction. In *CVPR*, 2025. 3
- [69] Chaofan Wu, Jiaheng Li, Jinghao Cao, Ming Li, Sidan Du, and Yang Li. OmniOcc: Cylindrical voxel-based semantic occupancy prediction for omnidirectional vision systems. *IEEE Access*, 2025. 3
- [70] Xianghui Pan, Jiayuan Du, Shuai Su, Wenhao Zong, Xiao Wang, Chengju Liu, and Qijun Chen. GenerOcc: Self-supervised framework of real-time 3D occupancy prediction for monocular generic cameras. In *IROS*, 2024. 3
- [71] Zhenxing Ming, Julie Stephany Berrio, Mao Shan, Yaoqi Huang, Hongyu Lyu, Nguyen Hoang Khoi Tran, Tzu-Yun Tseng, and Stewart Worrall. OccCylindrical: Multi-modal fusion with cylindrical representation for 3D semantic occupancy prediction. *arXiv preprint arXiv:2505.03284*, 2025. 3
- [72] Lianqing Zheng, Jianan Liu, Runwei Guan, Long Yang, Shouyi Lu, Yuanzhe Li, Xiaokai Bai, Jie Bai, Zhixiong Ma, Hui-Liang Shen, and Xichan Zhu. Doracamom: Joint 3D detection and occupancy prediction with multi-view 4D radars and cameras for omnidirectional perception. *arXiv preprint arXiv:2501.15394*, 2025. 3
- [73] Jingyi Pan, Zipeng Wang, and Lin Wang. Co-Occ: Coupling explicit feature fusion with volume rendering regularization for multi-modal 3D semantic occupancy prediction. *IEEE Robotics and Automation Letters*, 2024. 3
- [74] Heng Li, Yuenan Hou, Xiaohan Xing, Yuexin Ma, Xiao Sun, and Yanyong Zhang. OccMamba: Semantic occupancy prediction with state space models. In *CVPR*, 2025.
- [75] Long Yang, Lianqing Zheng, Wenjin Ai, Minghao Liu, Sen Li, Qunshu Lin, Shengyu Yan, Jie Bai, Zhixiong Ma, and Xichan Zhu. MetaOcc: Spatio-temporal fusion of surround-view 4D radar and camera for 3D occupancy prediction with dual training strategies. *arXiv preprint arXiv:2501.15384*, 2025. 3
- [76] Chengyang Zhang, Jiashi Zhang, Jun Wu, and Qiuguo Zhu. Vision-assisted localization and terrain reconstruction with quadruped robots. In *IROS*, 2022. 3
- [77] Manthan Patel, Fan Yang, Yuheng Qiu, Cesar Cadena, Sebastian A. Scherer, Marco Hutter, and Wenshan Wang. Tar-

- tanGround: A large-scale dataset for ground robot perception and navigation. *arXiv preprint arXiv:2505.10696*, 2025. 3
- [78] Ganyu Deng, Jianwen Luo, Caiming Sun, Dongwei Pan, Longyao Peng, Ning Ding, and Aidong Zhang. Vision-based navigation for a small-scale quadruped robot pegasus-mini. In *ROBIO*, 2021. 3
- [79] Sebastian Aegidius, Dennis Hadjivelichkov, Jianhao Jiao, Jonathan Embley-Riches, and Dimitrios Kanoulas. Watch your STEPP: Semantic traversability estimation using pose projected features. In *ICRA*, 2025.
- [80] Minh Oh, Byeongho Yu, I Made Aswin Nahrendra, Seoyeon Jang, Hyeonwoo Lee, Dongkyu Lee, Seungjae Lee, Yeeun Kim, Kevin Christiansen Marsim, Hyungtae Lim, and Hyun Myung. TRIP: Terrain traversability mapping with risk-aware prediction for enhanced online quadrupedal robot navigation. *arXiv preprint arXiv:2411.17134*, 2024. 3
- [81] Zhengcai Cao, Junnian Li, Shibo Shao, Dong Zhang, and Mengchu Zhou. Siamese adaptive network-based accurate and robust visual object tracking algorithm for quadrupedal robots. *IEEE Transactions on Cybernetics*, 2025. 3
- [82] Shuo Xin, Zhen Zhang, Liang Liu, Xiaojun Hou, Deye Zhu, Mengmeng Wang, and Yong Liu. A robotic-centric paradigm for 3D human tracking under complex environments using multi-modal adaptation. In *IROS*, 2024. 3
- [83] Yuting Mei, Ye Wang, Sipeng Zheng, and Qin Jin. QuadrupedGPT: Towards a versatile quadruped agent in open-ended worlds. *arXiv preprint arXiv:2406.16578*, 2024. 3
- [84] ByungOk Han, Woo-han Yun, Beom-Su Seo, and Jaehong Kim. Space-aware instruction tuning: Dataset and benchmark for guide dog robots assisting the visually impaired. *arXiv preprint arXiv:2502.07183*, 2025. 3
- [85] Songbo Li, Shixin Luo, Jun Wu, and Qiuguo Zhu. MOVE: Multi-skill omnidirectional legged locomotion with limited view in 3D environments. *arXiv preprint arXiv:2412.03353*, 2024. 3
- [86] Yi Cheng, Hang Liu, Guoping Pan, Houde Liu, and Linqi Ye. Quadruped robot traversing 3D complex environments with limited perception. In *IROS*, 2024.
- [87] Zifan Wang, Teli Ma, Yufei Jia, Xun Yang, Jiaming Zhou, Wenlong Ouyang, Qiang Zhang, and Junwei Liang. Omni-Perception: Omnidirectional collision avoidance for legged locomotion in dynamic environments. In *CoRL*, 2025. 3
- [88] Kai Luo, Hao Shi, Sheng Wu, Fei Teng, Mengfei Duan, Chang Huang, Yuhang Wang, Kaiwei Wang, and Kailun Yang. Omnidirectional multi-object tracking. In *CVPR*, 2025. 3
- [89] Jianping Li, Zhongyuan Liu, Xinhang Xu, Jinxin Liu, Shenghai Yuan, Fang Xu, and Lihua Xie. LiMo-Calib: On-site fast LiDAR-motor calibration for quadruped robot-based panoramic 3D sensing system. *arXiv preprint arXiv:2502.12655*, 2025. 3
- [90] Boyi Liu, Qianyi Zhang, Qiang Yang, Jianhao Jiao, Jagmohan Chauhan, and Dimitrios Kanoulas. The starlink robot: A platform and dataset for mobile satellite communication. *arXiv preprint arXiv:2506.19781*, 2025. 3
- [91] Gang Chen, Zhaoying Wang, Wei Dong, and Javier Alonso-Mora. Particle-based instance-aware semantic occupancy mapping in dynamic environments. *IEEE Transactions on Robotics*, 2025. 3
- [92] Wei Cui, Haoyu Wang, Wenkang Qin, Yijie Guo, Gang Han, Wen Zhao, Jiahang Cao, Zhang Zhang, Jiaru Zhong, Jingkai Sun, Pihai Sun, Shuai Shi, Botuo Jiang, Jiahao Ma, Jiaxu Wang, Hao Cheng, Zhichao Liu, Yang Wang, Zheng Zhu, Guan Huang, Jian Tang, and Qiang Zhang. Humanoid occupancy: Enabling a generalized multimodal occupancy perception system on humanoid robots. *arXiv preprint arXiv:2507.20217*, 2025. 3
- [93] Qiang Zhang, Zhang Zhang, Wei Cui, Jingkai Sun, Jiahang Cao, Yijie Guo, Gang Han, Wen Zhao, Jiaxu Wang, Chenghao Sun, Lingfeng Zhang, Hao Cheng, Yujie Chen, Lin Wang, Jian Tang, and Renjing Xu. HumanoidPano: Hybrid spherical panoramic-LiDAR cross-modal perception for humanoid robots. *arXiv preprint arXiv:2503.09010*, 2025.
- [94] Zhang Zhang, Qiang Zhang, Wei Cui, Shuai Shi, Yijie Guo, Gang Han, Wen Zhao, Jingkai Sun, Jiahang Cao, Jiaxu Wang, Hao Cheng, Xiaozhu Ju, Zhengping Che, Renjing Xu, and Jian Tang. Occupancy world model for robots. *arXiv preprint arXiv:2505.05512*, 2025. 3
- [95] Davide Scaramuzza, Agostino Martinelli, and Roland Siegwart. A toolbox for easily calibrating omnidirectional cameras. In *IROS*, 2006. 4
- [96] Jianbiao Mei, Yu Yang, Mengmeng Wang, Junyu Zhu, Jongwon Ra, Yukai Ma, Laijian Li, and Yong Liu. Camera-based 3D semantic scene completion with sparse guidance network. *IEEE Transactions on Image Processing*, 2024. 5, 6, 7
- [97] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 5