UAVOCC: HIERARCHICAL SPATIOTEMPORAL SEMANTIC OCCUPANCY PREDICTION FOR UAVS

Anonymous authors
Paper under double-blind review

ABSTRACT

Aerial semantic occupancy faces a resolution–coverage trade-off under a fixed voxel budget: fine discretization preserves local geometry but restricts physical extent, while coarse discretization expands coverage at the expense of detail. We formulate *hierarchical aerial semantic occupancy*, representing these competing requirements with complementary fine-local L1 and coarse large-range L2 physical levels of distinct metric resolutions and extents. To make this formulation measurable, we establish a benchmark spanning three virtual and real-world aerial data sources, with a common L1/L2 task definition and *Package-Physical* evaluation for non-redundant joint physical coverage. Modeling this hierarchy requires cross-level interaction beyond ordinary multi-scale fusion, as the two levels differ in metric correspondence and temporal support. We therefore develop UAVOcc with Global–Local Cross-Level Interaction (GLCI), which combines metric-local 3D correspondence with broader global context through bidirectional spatial exchange; pose-aligned same-level temporal fusion is further followed by coarse-to-fine Temporal GLCI to exploit the broader historical support of L2. Across all three benchmarks, UAVOcc consistently outperforms adapted occupancy baselines, improving Package-Physical mIoU by 3.41, 5.47, and 5.27 points over the strongest listed baselines.

1 INTRODUCTION

Semantic occupancy provides a metric 3D representation of scene geometry and semantics for autonomous perception and planning (Cao & de Charette, 2022; Tian et al., 2023; Zheng et al., 2024; Yang et al., 2025). Most existing formulations are developed around ground vehicles, where perception is concentrated within a relatively constrained region around the sensor. UAVs operate differently: large variations in altitude and viewpoint require fine reasoning about nearby structures while the ground and surrounding buildings may extend far beyond a compact local volume. Aerial occupancy therefore exposes a direct tension between *resolution* and *coverage*.

Under a fixed voxel budget, fine voxels preserve local geometry but restrict physical extent, whereas coarse voxels extend coverage at the expense of detail. Enlarging a fine volume instead rapidly increases volumetric cost. Sparse and adaptive representations improve volumetric efficiency (Lu et al., 2024; Liu et al., 2024), while recent aerial occupancy studies investigate multi-UAV collaboration and safety-oriented perception (Lin et al., 2025; Gao et al., 2026). Our focus is complementary: rather than forcing local detail and large-range coverage into a single discretization, we formulate *hierarchical aerial semantic occupancy* with two nested physical levels of distinct metric resolutions and extents: a fine local L1 volume and a coarser, larger-range L2 volume (Fig. 1(c)).

To make this formulation measurable, we establish a hierarchical aerial occupancy benchmark spanning complementary virtual and real-world sources (Yang et al., 2023; Gao et al., 2024; Wang et al., 2025). UrbanBIS-Occupancy, CUS3D-Occupancy, and UAVScenes-Occupancy share a common L1/L2 task definition. Beyond evaluating each level independently, *Package-Physical* combines the complete L1 volume with the non-overlapping outer L2 region, measuring their joint physical coverage without double-counting the shared region.

Modeling this physical hierarchy requires more than ordinary multi-scale feature fusion. L1 and L2 represent different metric resolutions and extents, so identical tensor indices need not correspond to identical physical locations. Temporal modeling introduces a further mismatch: UAV motion

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

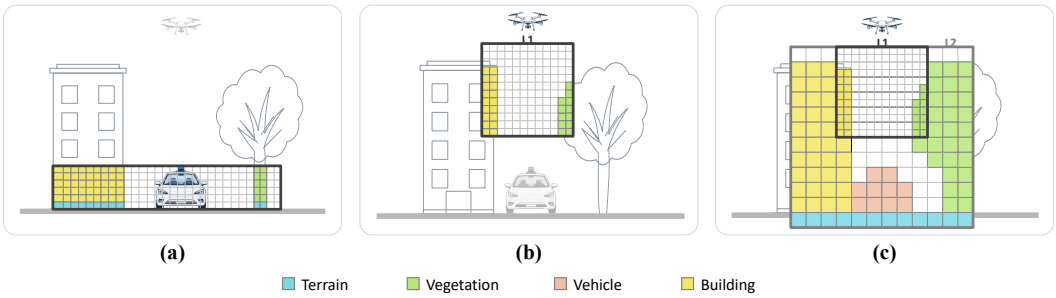

Figure 1: Resolution and physical coverage in aerial semantic occupancy. (a) Ground-vehicle occupancy typically focuses on a constrained region around the sensor. (b) A fine aerial discretization preserves local detail but restricts physical coverage. (c) Our hierarchical formulation represents fine local L1 and coarse large-range L2 as complementary physical levels.

breaks cross-frame voxel correspondence, while the broader L2 extent can retain historical support beyond L1. We develop *UAVOcc* to address these mismatches. Global–Local Cross-Level Interaction (GLCI) combines metric-local 3D correspondence with broader global context and exchanges evidence bidirectionally between L1 and L2. For temporal modeling, historical features are first pose-aligned and fused within each level, followed by coarse-to-fine Temporal GLCI that transfers broader L2 temporal support to L1.

Experiments across all three benchmarks demonstrate consistent improvements over adapted occupancy baselines. Under Package-Physical evaluation, UAVOcc achieves 22.22%, 18.59%, and 20.70% mIoU on UrbanBIS-Occupancy, CUS3D-Occupancy, and UAVScenes-Occupancy, exceeding the strongest listed baselines by 3.41, 5.47, and 5.27 points, respectively. Ablations further show substantial L1 gains with dual-level modeling, while bidirectional global–local interaction and temporal modeling provide additional improvements.

Our contributions are threefold:

- **Hierarchical aerial occupancy.** We formulate aerial semantic occupancy with complementary fine-local L1 and coarse large-range L2 physical levels, representing local detail and extended coverage through distinct metric discretizations.
- **Benchmark and evaluation.** We establish a hierarchical aerial occupancy benchmark spanning three virtual and real-world data sources under a common L1/L2 task definition, together with Package-Physical evaluation for non-redundant joint physical coverage.
- **Hierarchical spatiotemporal modeling.** We develop UAVOcc with GLCI for bidirectional global–local cross-level interaction, pose-aligned temporal fusion, and coarse-to-fine temporal refinement across the complementary physical levels.

2 Related Work

Semantic Occupancy Representation and Temporal Modeling. Semantic occupancy prediction represents geometry and semantics in volumetric space and has been extensively studied for autonomous driving. Early approaches construct dense occupancy volumes from monocular or surround-view images (Cao & de Charette, 2022; Wei et al., 2023), while recent methods improve scalability through structured or sparse representations. OctreeOcc adapts spatial granularity to scene content using octree queries, and SparseOcc replaces dense 3D features with fully sparse representations and queries (Lu et al., 2024; Liu et al., 2024). These methods optimize how a 3D space is represented, but do not explicitly formulate local detail and extended physical coverage as complementary prediction levels. A separate line of work exploits temporal evidence; GDFusion systematically integrates multiple temporal cues across the occupancy pipeline (Chen et al., 2025b). Our setting connects these two axes: it introduces an explicit hierarchy of physical extents and exploits this hierarchy both across levels and over time, with same-level temporal fusion followed by coarse-to-fine spatial retrieval.

Aerial Semantic Occupancy. Aerial datasets such as SynDrone, ECLAIR, UAVScenes, UrbanBIS, and CUS3D provide imagery, point clouds, poses, and semantic annotations that support 3D UAV perception (Rizzoli et al., 2023; Melekhov et al., 2024; Wang et al., 2025; Yang et al., 2023; Gao et al., 2024). More recently, semantic occupancy has emerged as an explicit representation for aerial scenes. MCOP formulates multi-UAV collaborative occupancy prediction, combining spatial feature encoding with cross-agent integration for cooperative perception (Lin et al., 2025). SkyShield instead formulates front-view semantic occupancy for safety-critical low-altitude flight, together with dynamics-aware evaluation (Gao et al., 2026). These works establish aerial occupancy under complementary settings, but target different dimensions of the problem: cross-agent collaboration in MCOP and safety-aware near-field perception in SkyShield. Our work focuses on a distinct axis—the resolution–coverage trade-off—and formulates it explicitly through nested physical levels that preserve fine local geometry while extending spatial coverage. The accompanying benchmark further evaluates both level-wise prediction and their non-redundant joint physical coverage.

3 Proposed Method

Building on the hierarchical aerial occupancy formulation, UAVOcc integrates complementary information along three dimensions: modality, physical level, and time. RGB and LiDAR observations are first fused within each physical level to construct L1 and L2 multimodal representations. Spatial Global–Local Cross-Level Interaction (GLCI) then exchanges evidence between the two physically distinct levels through metric-local and global cross-level retrieval. Finally, historical features are pose-aligned and fused within each level, followed by coarse-to-fine Temporal GLCI that transfers broader temporally aggregated L2 support to L1. Figure 2 summarizes the architecture.

Problem formulation. Given causal observations from $\{t-2, t-1, t\}$, including calibrated front-view RGB images, LiDAR observations, and ego poses, UAVOcc predicts semantic occupancy at time t over two nested UAV-centric volumes, $\Omega_1 \subset \Omega_2$. L1 covers $72 \times 72 \times 72\,\mathrm{m}$ with 0.75-m output voxels, while L2 covers $144 \times 144 \times 144\,\mathrm{m}$ with 1.5-m output voxels. Both levels contain 96^3 output voxels, defining complementary prediction targets for fine local detail and extended spatial coverage. Each valid voxel is labeled as free space or a dataset-specific semantic class, while invalid voxels are ignored during training.

3.1 Dual-Level Multimodal Fusion

For each frame k, UAVOcc constructs multimodal voxel representations for the L1 and L2 physical levels. The Image Encoder extracts visual features from the RGB observation, while the LiDAR Encoder provides sparse geometric support. For each level, Dual-Level Fusion projects the two modalities onto the corresponding predefined 48^3 latent grid and performs voxel-wise multimodal fusion, producing $\mathbf{F}_k^1$ and $\mathbf{F}_k^2$. Although identical in tensor dimensions, the two representations correspond to different physical extents and therefore have different latent-cell sizes. Dual-Level Fusion operates within each physical level to integrate RGB and LiDAR evidence; interaction between L1 and L2 is introduced subsequently by GLCI. Detailed encoder, projection, and voxelization configurations are provided in Appendix C.1.

3.2 Global–Local Cross-Level Interaction (GLCI)

Although $\mathbf{F}_k^1$ and $\mathbf{F}_k^2$ share the same tensor dimensions, they represent different metric resolutions and physical extents. The same tensor index therefore does not generally correspond to the same physical location, making direct index-wise fusion unsuitable for cross-level interaction. GLCI instead performs interaction in metric space. Given a target representation $\mathbf{T}^r$ at level r and a source representation $\mathbf{S}^s$ at level s, we define a generic cross-level operator $\mathcal{G}_{s\to r}(\mathbf{T}^r, \mathbf{S}^s)$ that establishes metric correspondence, retrieves complementary local and global source evidence, and injects the resulting message into the native target representation.

The two retrieval branches serve distinct roles. Local retrieval preserves explicit physical correspondence across discretizations, whereas global retrieval relaxes this locality to capture broader contextual dependencies within the aligned source region. Height-Aware Gated Refinement (HGR) then converts the retrieved planar global context into voxel-specific context and combines it with the

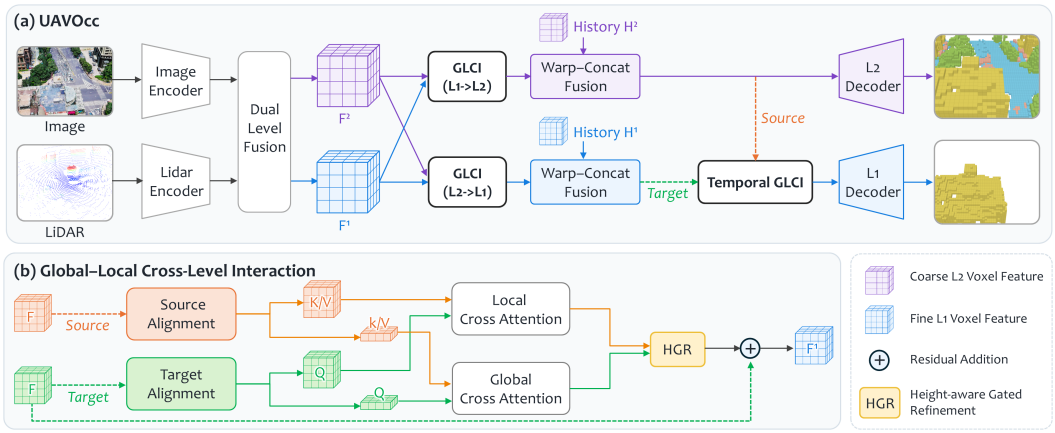

Figure 2: **Overview of UAVOcc.** (a) RGB and LiDAR observations are fused within L1 and L2, followed by bidirectional Spatial GLCI, pose-aligned Warp–Concat Fusion, and coarse-to-fine Temporal GLCI. History H^1/H^2 stores previous Spatial GLCI outputs at the respective levels. (b) GLCI retrieves metric-local 3D and global planar context and combines them through HGR.

local 3D evidence. For current-frame modeling, Spatial GLCI applies this operator bidirectionally:

$$\widetilde{\mathbf{F}}_k^1 = \mathcal{G}_{2\to1}^{\mathrm{sp}}\left(\mathbf{F}_k^1, \mathbf{F}_k^2\right), \qquad \widetilde{\mathbf{F}}_k^2 = \mathcal{G}_{1\to2}^{\mathrm{sp}}\left(\mathbf{F}_k^2, \mathbf{F}_k^1\right). \tag{1}$$

The two directions use independent parameters and operate in parallel on the same pre-update features. The $L2 \to L1$ path supplies broader spatial context to the fine local representation, whereas $L1 \to L2$ transfers finer evidence to the physically overlapping L2 region. We describe $L2 \to L1$ below; the reverse direction follows the same formulation with source and target exchanged.

Metric source–target alignment. Since tensor indices across L1 and L2 do not imply physical correspondence, GLCI maps target cells to the source grid through metric coordinates. Let $\mathbf{a}_\ell$ and $\boldsymbol{\Delta}_\ell$ denote the physical origin and latent-cell size of level ℓ, respectively. For target cell $\mathbf{i}$,

$$\mathbf{x}_i^r = \mathbf{a}_r + \left(\mathbf{i} + \frac{1}{2}\right) \odot \boldsymbol{\Delta}_r, \qquad \mathbf{u}_i^s = \left(\mathbf{x}_i^r - \mathbf{a}_s\right) \oslash \boldsymbol{\Delta}_s - \frac{1}{2}, \tag{2}$$

where $\mathbf{x}_i^r$ is the physical center of target cell $\mathbf{i}$, $\mathbf{u}_i^s$ its continuous source-grid coordinate, and $\oslash$ denotes element-wise division. This mapping establishes correspondence across different metric discretizations without resampling either representation. The physical overlap defines the valid target region, while the source crop is minimally expanded around its boundary to retain complete local neighborhoods. Detailed crop sizes and boundary handling are provided in Appendix C.2.

Local cross-attention. The continuous source coordinate $\mathbf{u}_i^s$ generally lies between discrete source cells. Instead of interpolating them into a single feature, Local Cross-Attention retains the eight enclosing cells $\mathcal{N}_s(\mathbf{u}_i^s)$ as candidates, enabling retrieval based on feature compatibility and relative metric geometry. For target cell i, let $\mathbf{Q}_i$ be its query and $\mathbf{K}_i, \mathbf{V}_i$ the keys and values of $\mathcal{N}_s(\mathbf{u}_i^s)$. Their relative displacements $\mathbf{R}_i \in \mathbb{R}^{8\times3}$ are encoded as a learned attention bias $\boldsymbol{\beta}(\mathbf{R}_i)$ and a geometry-aware value embedding $\mathbf{e}(\mathbf{R}_i)$:

$$\mathbf{L}_i = W_l\ \mathrm{Vec}\left[\mathrm{MaskedSoftmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{d}} + \boldsymbol{\beta}(\mathbf{R}_i)\right)\left(\mathbf{V}_i + \mathbf{e}(\mathbf{R}_i)\right)\right]. \tag{3}$$

Here, d is the per-head dimension, Vec concatenates the attention heads, and W_l projects their output to the target channel dimension. The displacement bias modulates attention over neighboring source cells, while the value embedding injects relative geometry into the retrieved features. Invalid neighbors are masked before normalization. Applying this operation over the target region produces the local volumetric message $\mathbf{L}$ with explicit metric correspondence.

Global cross-attention. Local Cross-Attention is restricted to physically corresponding neighborhoods and therefore cannot capture dependencies beyond this local support. Global Cross-Attention complements it by retrieving broader contextual evidence from the aligned source region. To avoid dense global attention over the full 3D volumes, learned height aggregation first converts the source and target features into compact planar representations, denoted by $\overline{\mathbf{S}}$ and $\overline{\mathbf{T}}$. After adding learned planar positional encodings, the target representation is projected to queries $\mathbf{Q}_g$, while the source representation is projected to keys $\mathbf{K}_g$ and values $\mathbf{V}_g$. Global Cross-Attention computes

$$\mathbf{B}_g = W_g \,\mathrm{Reshape}\left[\mathrm{Softmax}\left(\frac{\mathbf{Q}_g\mathbf{K}_g^\top}{\sqrt{d_g}}\right)\mathbf{V}_g\right], \tag{4}$$

where d_g is the per-head feature dimension and $\mathbf{B}_g$ denotes the retrieved planar context. In contrast to the correspondence-preserving local branch, each global target query can retrieve evidence from the broader aligned source memory, capturing contextual dependencies beyond its enclosing metric neighborhood.

Height-aware gated refinement. Because global retrieval operates on height-aggregated planar features, its output $\mathbf{B}_g$ does not explicitly distinguish target cells at different vertical positions. HGR therefore broadcasts $\mathbf{B}_g$ along the target height dimension and conditions it on the target feature and normalized height coordinates, producing voxel-specific global context $\mathbf{Z}_g$. A learned voxel-wise gate $\mathbf{G}$ then combines this context with the local message:

$$\mathbf{M} = \mathbf{G} \odot \mathbf{L} + (1 - \mathbf{G}) \odot \mathbf{Z}_g. \tag{5}$$

The fused message is further refined and residually injected into the native target representation. For $L2 \to L1$, the complete L1 target volume is refined, whereas $L1 \to L2$ updates only the physically overlapping L2 region. Appendix C.4 details the lifting, gating, and residual updates.

3.3 Temporal Modeling

Temporal modeling across L1 and L2 involves two distinct mismatches. First, UAV motion breaks voxel-wise correspondence across frames. Second, their different physical extents result in unequal historical coverage. For a current L1 location $\mathbf{x}_t$, pose transformation into a historical frame τ may place its correspondence outside the historical L1 volume while it remains inside L2, i.e., $\mathbf{x}_\tau \notin \Omega_1$ but $\mathbf{x}_\tau \in \Omega_2$. UAVOcc addresses these two mismatches sequentially: pose-aligned same-level fusion first compensates for motion, after which Temporal GLCI exploits the remaining coverage asymmetry by transferring broader temporally aggregated L2 support to L1.

Pose-aligned Warp–Concat fusion. For each level $\ell \in \{1, 2\}$, the spatially refined features from frames $t-2$ and $t-1$ are aligned to the current UAV frame using the corresponding ego poses. Given a current voxel center $\mathbf{x}_t$, its location in a historical frame τ is obtained through the relative inter-frame ego transformation,

$$\mathbf{x}_\tau = \mathbf{T}_{\tau \leftarrow t}\mathbf{x}_t, \tag{6}$$

where $\mathbf{T}_{\tau \leftarrow t}$ denotes the transformation from the current frame to frame τ. The transformed coordinate is used to trilinearly sample the corresponding historical feature volume, while a validity mask records whether it lies within the historical level.

The aligned historical features and validity masks are concatenated with the current feature and processed by a level-specific fusion block, producing temporally aggregated representations $\mathbf{U}_t^1$ and $\mathbf{U}_t^2$. Pose alignment restores cross-frame correspondence where valid historical support exists, but cannot create evidence outside the physical extent of a level. Consequently, some current L1 locations may lack valid L1 history while remaining supported by the broader L2 history.

Temporal GLCI. After same-level temporal aggregation, L1 and L2 still differ in their available temporal support because of their distinct physical extents. Temporal GLCI applies the same metric-aware global–local interaction principle to the temporally aggregated representations, using L2 as the source and L1 as the target:

$$\widehat{\mathbf{U}}_t^1 = \mathcal{G}_{2\to1}^{\mathrm{tmp}}\left(\mathbf{U}_t^1, \mathbf{U}_t^2\right). \tag{7}$$

The temporal operator uses parameters independent of Spatial GLCI. Unlike the bidirectional current-frame interaction, Temporal GLCI operates only from L2 to L1. This asymmetry reflects the different roles of the two stages: current-frame L1 and L2 provide complementary spatial evidence

to each other, whereas after same-level temporal aggregation the remaining coverage asymmetry specifically motivates transferring broader L2 support to L1. The refined L1 representation $\widehat{\mathbf{U}}_t^1$ and the temporally aggregated L2 representation $\mathbf{U}_t^2$ are then passed to their respective occupancy decoders for prediction.

3.4 Occupancy Decoding and Supervision

The refined L1 representation $\widehat{\mathbf{U}}_t^1$ and the temporally aggregated L2 representation $\mathbf{U}_t^2$ are decoded by separate level-specific 3D heads. Each head produces semantic occupancy logits on its 48^3 latent grid, which are trilinearly upsampled to the corresponding 96^3 task grid for L1 and L2 prediction. Both physical levels are supervised with cross-entropy loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{L1}} + \mathcal{L}_{\mathrm{L2}}, \tag{8}$$

where invalid voxels are excluded from supervision. No explicit output-level consistency loss is imposed between L1 and L2; their physical correspondence is instead modeled through the metric-aligned spatial and temporal interactions described above.

4 Experimental Results

Benchmarks. We evaluate UAVOcc on three aerial occupancy benchmarks: UrbanBIS-Occupancy and CUS3D-Occupancy from virtual-UAV observations (Yang et al., 2023; Gao et al., 2024), and UAVScenes-Occupancy from real-world UAV observations and annotations (Wang et al., 2025). We use the corresponding training and validation splits. Appendix A details dataset construction, label mappings, and fixed splits.

Evaluation protocol. We report foreground semantic mIoU and binary occupied-space IoU at L1, L2, and Package-Physical. L1 and L2 are evaluated independently, while Package-Physical combines the full L1 volume with the non-overlapping outer L2 region, avoiding double-counting their physical overlap. Ignore labels are excluded. Appendix D.3 specifies the evaluation protocol.

Implementation details. Unless otherwise specified, UAVOcc uses the current and two preceding frames. The image encoder is initialized from ImageNet-pretrained EfficientNet-B0, and the model is optimized with AdamW. For each dataset, we select the checkpoint with the highest Package-Physical validation mIoU.

Comparison protocol. We compare with camera-only SparseOcc (Liu et al., 2024) and ALOcc (Chen et al., 2025a), and camera–LiDAR OpenOccupancy (Wang et al., 2023), BEVFusion (Liu et al., 2023), Co-Occ (Pan et al., 2024), and OccMamba (Li et al., 2025). All baselines are adapted and evaluated on the target benchmarks using the same splits, semantic mappings, ignore masks, L1/L2 geometry, and evaluation implementation; input modalities are reported explicitly. Adaptation details are provided in Appendix D.2.

4.1 Quantitative Comparison

Table 1 compares UAVOcc with adapted occupancy baselines across the three benchmarks. UAVOcc achieves the highest listed performance across L1, L2, and Package-Physical in all three cases. Under Package-Physical, it reaches 22.22%, 18.59%, and 20.70% mIoU on UrbanBIS-Occupancy, CUS3D-Occupancy, and UAVScenes-Occupancy, exceeding the strongest listed baselines by 3.41, 5.47, and 5.27 points, respectively. The gains also hold at both physical levels, with L1/L2 mIoU improvements of 5.24/1.98, 5.27/5.96, and 12.25/8.78 points across the three benchmarks, showing that the improvements extend to both local-detail prediction and extended-range coverage. On UAVScenes-Occupancy, derived from real-world UAV observations, UAVOcc further achieves 48.43%, 46.05%, and 45.21% Binary IoU at L1, L2, and Package-Physical, respectively. Overall, UAVOcc consistently improves prediction across complementary physical levels and their non-redundant joint coverage. Since UAVScenes-Occupancy uses a same-map holdout, these results do not establish cross-map generalization.

Table 1: Comparison with adapted occupancy baselines on UrbanBIS-Occupancy, CUS3D-Occupancy, and UAVScenes-Occupancy. C and C+L denote camera-only and camera–LiDAR inputs, respectively. Results are reported in %. * denotes Package-Physical mIoU recomputed from the archived per-class records in Appendix E.3, not a new evaluation; other entries are unchanged. The two marked CUS3D records remain provisional.

Dataset	Method	L1		L2		Package-Physical	
		mIoU ↑	Binary IoU ↑	mIoU ↑	Binary IoU ↑	mIoU ↑	Binary IoU ↑
UrbanBIS Occupancy	SparseOcc (C)	3.30	9.32	6.76	17.90	4.04	11.41
	ALOcc (C)	2.06	3.87	2.86	5.17	2.33	4.17
	OpenOccupancy (C+L)	11.13	19.18	22.91	40.17	15.83	26.48
	BEVFusion (C+L)	15.73	26.19	23.49	41.90	18.81	31.78
	Co-Occ (C+L)	10.41	19.50	19.67	39.85	13.59	26.39
	OccMamba (C+L)	11.36	19.64	18.25	30.71	12.76	22.52
	UAVOcc (Ours, C+L)	**20.97**	**32.64**	**25.47**	**43.06**	**22.22**	**35.47**
CUS3D Occupancy	SparseOcc (C)	2.38	6.15	7.85	15.51	5.79	10.04
	ALOcc (C)	1.09	3.96	1.60	4.96	1.23	3.96
	OpenOccupancy (C+L)	6.96	15.80	15.24	31.40	8.82*	21.19
	BEVFusion (C+L)	12.41	26.00	13.74	29.83	13.12	27.06
	Co-Occ (C+L)	4.52	12.79	7.99	26.91	5.01*	17.11
	OccMamba (C+L)	9.11	16.47	14.78	27.21	11.08	19.56
	UAVOcc (Ours, C+L)	**17.68**	**31.37**	**21.20**	**39.68**	**18.59**	**32.87**
UAVScenes Occupancy	SparseOcc (C)	2.14	1.60	3.69	4.66	3.40*	3.31
	ALOcc (C)	1.22	0.99	0.60	2.06	0.77	1.62
	OpenOccupancy (C+L)	13.76	47.76	11.32	43.13	15.43*	43.12
	BEVFusion (C+L)	10.62	34.83	11.98	44.10	14.03*	37.12
	Co-Occ (C+L)	11.23	42.62	11.66	44.28	14.22*	41.52
	OccMamba (C+L)	7.50	40.60	7.65	36.71	9.89*	37.40
	UAVOcc (Ours, C+L)	**26.01**	**48.43**	**20.76**	**46.05**	**20.70**	**45.21**

4.2 Qualitative Comparison

Figure 3 qualitatively compares UAVOcc with representative baselines on CUS3D-Occupancy. Across the nested physical volumes, UAVOcc produces more coherent semantic occupancy while preserving local scene structure at L1 and maintaining occupancy over the extended L2 range. The combined physical occupancy further illustrates how the complementary levels jointly represent fine local geometry and broader scene context. Supplementary Figures 10–11 provide additional comparisons on UrbanBIS-Occupancy and UAVScenes-Occupancy.

4.3 Ablation Studies

Effect of dual-level modeling. A larger L2 volume naturally extends spatial coverage, but whether dual-level modeling also benefits prediction within the fine L1 region remains unclear. To separate this effect from the additional outer range of L2, we compare L1-only models with the complete L1+L2 hierarchy while evaluating both on the same L1 physical volume. As shown in Table 2, dual-level modeling improves L1 mIoU by 7.88, 8.11, and 10.86 points on UrbanBIS, CUS3D, and UAVScenes, respectively, with corresponding Binary-IoU gains of 11.24, 12.53, and 17.92 points. Since both variants are evaluated over the same physical region, these gains cannot be attributed to the additional outer L2 evaluation range. The results instead suggest that dual-level modeling also benefits prediction within L1. Together with the extended L2 coverage measured by Package-Physical, the hierarchy therefore supports both fine-range prediction and large-range coverage.

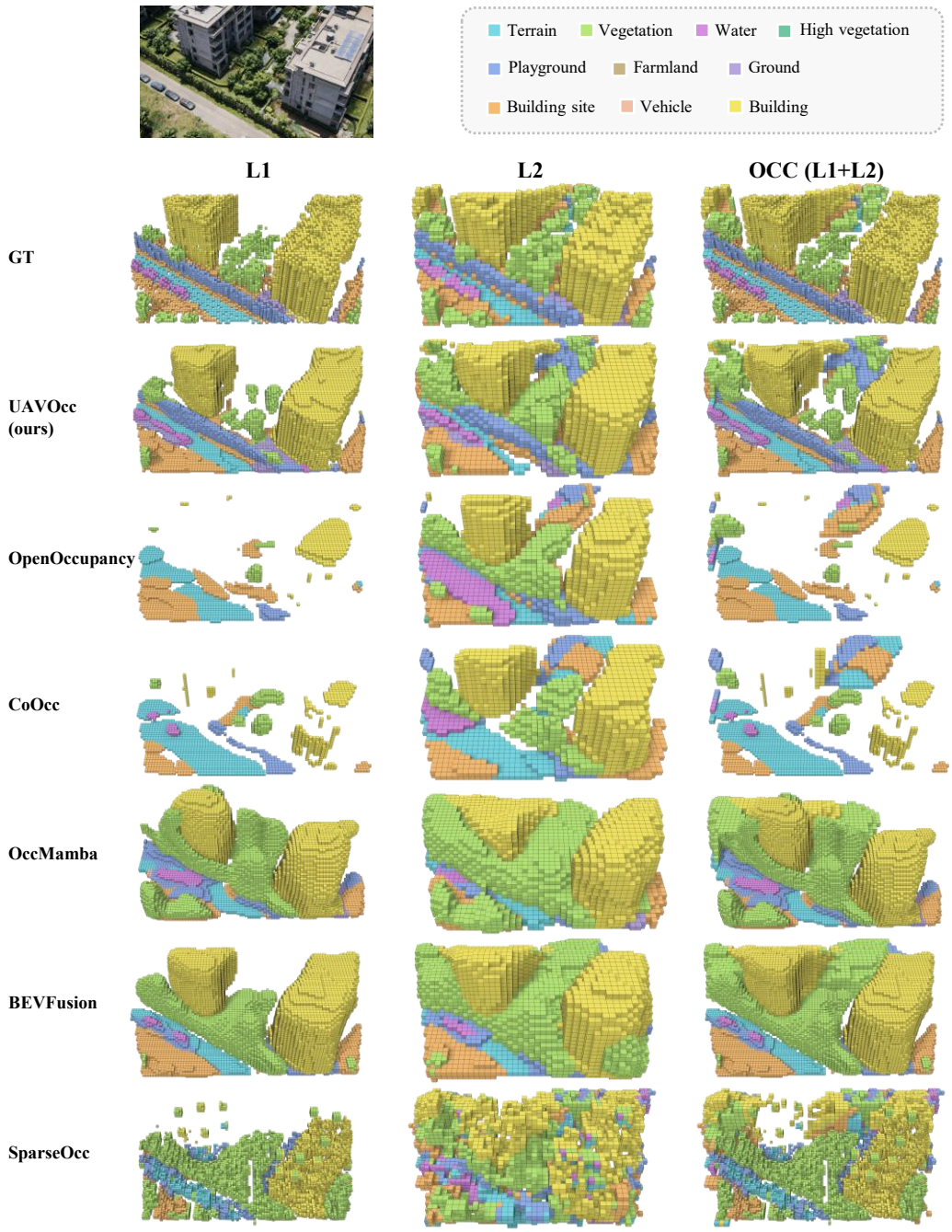

Figure 3: Qualitative comparison on CUS3D-Occupancy. Rows show the ground truth (GT), UAV-Occ, and representative baselines; columns show L1, L2, and their combined physical occupancy (OCC). Colors denote semantic categories.

Spatial GLCI. Table 3 separates interaction direction from retrieval-branch composition. Bidirectional G+L achieves the highest Package-Physical mIoU on all three datasets, improving over no Spatial GLCI by 2.07, 2.27, and 1.01 points on UrbanBIS, CUS3D, and UAVScenes, respectively. Single-direction exchange is less consistent and can degrade performance. Under bidirectional exchange, neither branch alone matches their combination. These results support reciprocal cross-level interaction with complementary global and local evidence.

Table 2: Effect of dual-level modeling evaluated on the same L1 physical region. Values are L1 mIoU / Binary IoU (%).

Variant	UrbanBIS	CUS3D	UAVScenes
L1-only	13.09 / 21.40	9.57 / 18.84	15.15 / 30.51
L1+L2	**20.97 / 32.64**	**17.68 / 31.37**	**26.01 / 48.43**

Table 3: Spatial GLCI ablation. G and L denote the global and local branches. Warp–Concat fusion and Temporal GLCI remain enabled in every row; only the indicated spatial interaction is changed. Values are Package-Physical mIoU / Binary IoU (%).

L1→L2	L2→L1	Spatial branch	UrbanBIS	CUS3D	UAVScenes
×	×	–	20.15 / 33.84	16.32 / 31.26	19.69 / 43.86
✓	×	G+L	21.17 / 33.76	16.60 / 31.31	19.23 / 44.27
×	✓	G+L	21.65 / 33.63	14.41 / 30.37	17.90 / 45.16
✓	✓	G	20.92 / 33.99	15.69 / 31.35	18.27 / 41.48
✓	✓	L	19.20 / 33.18	16.52 / 32.40	18.86 / 43.74
✓	✓	G+L	**22.22 / 35.47**	**18.59 / 32.87**	**20.70 / 45.21**

Spatial and temporal components. We further disentangle current-frame Spatial GLCI, pose-aligned Warp–Concat fusion, and Temporal GLCI. The configurations in Table 4 progressively examine current-frame cross-level interaction, same-level temporal aggregation, and coarse-to-fine temporal interaction. Spatial GLCI alone improves Package-Physical mIoU over the component-free baseline by 0.32, 0.48, and 2.21 points on UrbanBIS, CUS3D, and UAVScenes, respectively. Warp–Concat alone is less consistent, improving UrbanBIS and UAVScenes but slightly decreasing CUS3D. When combined with Spatial GLCI, however, pose-aligned history improves mIoU by 1.26, 1.82, and 2.77 points across the three datasets. Temporal GLCI provides a further 2.60, 1.88, and 1.40 points, yielding the strongest configuration on all three datasets. These results suggest that pose-aligned same-level history and coarse-to-fine temporal interaction provide complementary temporal information on top of current-frame cross-level modeling.

Table 4: Component ablation of Spatial GLCI, Warp–Concat fusion, and Temporal GLCI. Values are Package-Physical mIoU / Binary IoU (%).

Components			Package-Physical mIoU / Binary IoU		
Spatial GLCI	Warp–Concat	Temporal GLCI	UrbanBIS	CUS3D	UAVScenes
×	×	×	18.04 / 29.96	14.41 / 28.27	14.32 / 35.10
✓	×	×	18.36 / 31.69	14.89 / 30.59	16.53 / 37.39
×	✓	×	19.20 / 31.97	14.09 / 29.45	17.13 / 41.06
✓	✓	×	19.62 / 33.06	16.71 / 31.99	19.30 / 41.00
✓	✓	✓	**22.22 / 35.47**	**18.59 / 32.87**	**20.70 / 45.21**

5 Conclusion

We formulate hierarchical aerial semantic occupancy by representing fine local prediction and extended spatial coverage as complementary physical levels with distinct metric extents. To make this formulation measurable, we establish a benchmark spanning UrbanBIS-Occupancy, CUS3D-Occupancy, and UAVScenes-Occupancy, together with L1/L2 evaluation and Package-Physical for their non-redundant joint coverage. Building on this hierarchy, UAVOcc couples bidirectional Global–Local Cross-Level Interaction with pose-aligned temporal fusion and coarse-to-fine Temporal GLCI. Experiments across all three benchmarks show consistent gains from dual-level modeling, spatial cross-level interaction, and temporal integration. Together, these results show that fine-range prediction and large-range coverage can be effectively modeled within a unified hierarchical occupancy representation for UAVs.

GENERATIVE AI DISCLOSURE

Generative AI tools assisted with manuscript editing, reference formatting, code writing, and discussion of result presentation. The authors reviewed and verified the AI-assisted material and take full responsibility for the experimental evidence, its interpretation, and all content of this manuscript.

REFERENCES

Anh-Quan Cao and Raoul de Charette. MonoScene: Monocular 3D semantic scene completion. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pp. 3991–4001, 2022.

Dubing Chen, Jin Fang, Wencheng Han, Xinjing Cheng, Junbo Yin, Chengzhong Xu, Fahad Shahbaz Khan, and Jianbing Shen. ALOcc: Adaptive lifting-based 3D semantic occupancy and cost volume-based flow predictions. In *Proc. Int. Conf. on Computer Vision*, pp. 4156–4166, 2025a.

Dubing Chen, Huan Zheng, Jin Fang, Xingping Dong, Xianfei Li, Wenlong Liao, Tao He, Pai Peng, and Jianbing Shen. Rethinking temporal fusion with a unified gradient descent view for 3D semantic occupancy prediction. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pp. 1505–1515, 2025b.

Jie Gao, Jie Ma, Kaihui Lin, Kai Ye, Miaohui Zhang, Pingyang Dai, and Liujuan Cao. SkyShield: Occupancy as a safety interface for low-altitude UAV autonomy. *arXiv preprint arXiv:2606.00747*, 2026.

Lin Gao, Yu Liu, Xi Chen, Yuxiang Liu, Shen Yan, and Maojun Zhang. CUS3D: A new comprehensive urban-scale semantic-segmentation 3D benchmark dataset. *Remote Sensing*, 16(6):1079, 2024. doi: 10.3390/rs16061079.

Heng Li, Yuenan Hou, Xiaohan Xing, Yuexin Ma, Xiao Sun, and Yanyong Zhang. OccMamba: Semantic occupancy prediction with state space models. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pp. 11949–11959, 2025.

Zefu Lin, Wenbo Chen, Xiaojuan Jin, Yuran Yang, Lue Fan, Yixin Zhang, Yufeng Zhang, and Zhaoxiang Zhang. MCOP: Multi-UAV collaborative occupancy prediction. In *Proc. Int. Conf. on Computer Vision*, pp. 27242–27251, 2025. doi: 10.1109/ICCV51701.2025.02529.

Haisong Liu, Yang Chen, Haiguang Wang, Zetong Yang, Tianyu Li, Jia Zeng, Li Chen, Hongyang Li, and Limin Wang. Fully sparse 3D occupancy prediction. In *Proc. Euro. Conf. on Computer Vision*, pp. 54–71, 2024.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *Proc. IEEE Int. Conf. on Robotics & Automation*, pp. 2774–2781, 2023.

Yuhang Lu, Xinge Zhu, Tai Wang, and Yuexin Ma. OctreeOcc: Efficient and multi-granularity occupancy prediction using octree queries. In *Advances in Neural Information Processing Systems*, volume 37, pp. 79618–79641, 2024.

Iaroslav Melekhov, Anand Umashankar, Hyeong-Jin Kim, Vladislav Serkov, and Dusty Argyle. ECLAIR: A high-fidelity aerial LiDAR dataset for semantic segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition Workshops*, pp. 7627–7637, 2024.

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*, 9(6):5687–5694, 2024.

Giulia Rizzoli, Francesco Barbato, Matteo Caligiuri, and Pietro Zanuttigh. SynDrone – multi-modal UAV dataset for urban scenarios. In *Proc. IEEE/CVF Int. Conf. on Computer Vision Workshops*, pp. 2210–2220, 2023.

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 *Proc. Conf. on Neural Information Processing Systems*, volume 36, pp. 64318–64330, 2023.

Sijie Wang, Siqi Li, Yawei Zhang, Shangshu Yu, Shenghai Yuan, Rui She, Quanjiang Guo, JinXuan Zheng, Ong Kang Howe, Leonrich Chandra, Shrivarshann Srijeyan, Aditya Sivadas, Toshan Aggarwal, Heyuan Liu, Hongming Zhang, Chujie Chen, Junyu Jiang, Lihua Xie, and Wee Peng Tay. UAVScenes: A multi-modal dataset for UAVs. In *Proc. Int. Conf. on Computer Vision*, pp. 28946–28958, 2025.

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 *Proc. Int. Conf. on Computer Vision*, pp. 17850–17859, 2023.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. SurroundOcc: Multi-camera 3D occupancy prediction for autonomous driving. In *Proc. Int. Conf. on Computer Vision*, pp. 21729–21740, 2023.

Guoqing Yang, Fuyou Xue, Qi Zhang, Ke Xie, Chi-Wing Fu, and Hui Huang. UrbanBIS: A large-scale benchmark for fine-grained urban building instance segmentation. In *Proc. SIGGRAPH*, pp. 1–11, 2023.

Yu Yang, Jianbiao Mei, Yukai Ma, Siliang Du, Wenqing Chen, Yijie Qian, Yuxiang Feng, and Yong Liu. Driving in the occupancy world: Vision-centric 4D occupancy forecasting and planning via world models for autonomous driving. In *Proc. AAAI Conf. on Artificial Intelligence*, pp. 9327–9335, 2025.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. OccWorld: Learning a 3D occupancy world model for autonomous driving. In *Proc. Euro. Conf. on Computer Vision*, pp. 55–72, 2024.