

Review

UAV 3D Scene Understanding: A Survey from an Agent-Capability Evolution Perspective

Enze Zhu ^{1,2} , Luxiao Xu ^{1,2} , Zhan Chen ^{1,2} , Jiahui Cui ^{1,2} , Jiayuan Wang ^{1,2,3} , Yongkang Zou ^{1,2,4},
Kaibo Yang ^{1,2} , Xiaoxuan Liu ^{1,2,*} , Xiyu Qi ^{1,2} and Lei Wang ^{1,2} 

¹ Key Laboratory of Target Cognition and Application Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China; zhuzenze23@mails.ucas.ac.cn (E.Z.); xuluxiao20@mails.ucas.ac.cn (L.X.); chenzhan21@mails.ucas.ac.cn (Z.C.); cuijiahui24@mails.ucas.ac.cn (J.C.); wjya@tju.edu.cn (J.W.); zouyk2022@mails.jlu.edu.cn (Y.Z.); yangkaibo25@mails.ucas.ac.cn (K.Y.); qixiyu20@mails.ucas.ac.cn (X.Q.); wanglei002931@aircas.ac.cn (L.W.)

² School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100094, China

³ School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

⁴ College of Communication Engineering, Jilin University, Changchun 130012, China

* Correspondence: liuxiaoxuan@aircas.ac.cn

Highlights

What are the main findings?

- This review introduces an evolutionary three-stage framework from an agent-capability perspective, organizing UAV 3D scene understanding methods into offline interpretation, online understanding and predictive reasoning.
- The reviewed literature shows that UAV 3D scene understanding is evolving from post-flight interpretation toward online understanding and predictive reasoning, enabling UAVs to transition from passive remote sensing platforms toward active embodied agents.

What is the implication of the main finding?

- This perspective reframes UAV 3D scene understanding from isolated perception tasks into an enabling capability for UAV embodied intelligence.
- It guides future research to focus on closed-loop benchmark construction, trustworthy scene-state memory, collaborative 3D understanding, sim-to-real transfer and reliable onboard deployment.

Abstract

Low-altitude economy, fine-grained surveying, emergency response, and autonomous exploration are driving Unmanned Aerial Vehicles (UAVs) from passive data-acquisition platforms toward task-executing aerial agents. This capability transition requires UAVs to operate within complex, open 3D environments under six-degree-of-freedom (6DoF) motion, strict size, weight and power (SWaP) limits, partial observations, onboard computation constraints, and safety-critical action requirements. Therefore, the central scientific problem of UAV 3D scene understanding is how a UAV agent can construct, maintain, and use a spatiotemporally coherent and uncertainty-aware 3D scene state to support localization, planning, and safe interaction. Existing surveys mainly categorize the literature according to sensor types, application scenarios, or generic 3D representations, and thus provide limited analysis of how 3D scene understanding supports agent-capability evolution under embodied aerial constraints. To address this gap, we review UAV 3D scene understanding along an agent-capability evolution from offline interpretation to online understanding



Academic Editor: Massimiliano Pepe

Received: 15 May 2026

Revised: 9 July 2026

Accepted: 9 July 2026

Published: 11 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

and predictive reasoning. This perspective highlights the underlying tensions between representation fidelity and onboard deployability, open-vocabulary semantic coverage and calibrated trustworthiness, post-flight static reconstruction and online scene-state maintenance, and predictive reasoning and safety-bounded decision support. Finally, we discuss open challenges in closed-loop data construction, trustworthy scene-state memory, collaborative fusion, sim-to-real transfer, and reliable onboard deployment.

Keywords: Unmanned Aerial Vehicle; embodied intelligence; intelligent agent; 3D scene understanding; multimodal perception

1. Introduction

Unmanned Aerial Vehicles (UAVs) are now widely employed in the low-altitude economy [1], fine-grained surveying [2], emergency response [3], and autonomous exploration [4]. Despite their diverse task goals, these applications share a fundamental prerequisite: safely completing tasks within a complex, open 3D environment. Operating within such environments requires more than image-level recognition or post-flight geometric reconstruction, because terrain relief, building facades, dynamic objects, free space, occlusion and obstacle clearance jointly determine traversability, observability and risk boundaries. Traditional approaches [5–9] primarily treat UAVs as data collectors for post-flight analysis. However, when scene information must support in-flight localization, path planning, obstacle avoidance and safe interaction, post-acquisition interpretation alone becomes insufficient.

In this review, the core scientific problem is how a UAV agent constructs, maintains and uses a spatiotemporally coherent and uncertainty-aware 3D scene state under embodied aerial constraints. This problem is shaped by six-degree-of-freedom (6DoF) motion, partial and rapidly changing observations, strict size, weight and power (SWaP) limits, onboard computation constraints, open low-altitude environments and safety-critical action requirements [10]. It also exposes several underlying tensions: high-fidelity 3D representation versus real-time onboard deployability, open-world semantic coverage versus trustworthy and calibrated interpretation, static post-flight reconstruction versus online action-oriented scene-state maintenance, and predictive scene reasoning versus safety-bounded decision support. Therefore, UAV 3D scene understanding should not be viewed as a loose collection of perception tasks, but as an agent-oriented 3D scene-understanding capability that links sensing, memory and action in embodied UAV operation.

However, existing surveys typically review prior UAV-centric studies based on specific sensor modalities [11–13] or remote sensing applications [14–16]. Meanwhile, reviews regarding general 3D scene understanding mainly focus on ground-based autonomous vehicles [17–19] or indoor robotics [20–22]. Consequently, the current academic landscape lacks a developmental perspective on agent capability, especially one that explains how UAV 3D scene understanding supports increasingly capable embodied UAV agents.

To address this gap, this review introduces a three-stage agent-capability framework to systematically review the field of UAV 3D scene understanding, as illustrated in Figure 1. Within this framework, our analysis examines how UAV 3D scene understanding capabilities expand from offline static interpretation to online scene-state maintenance and predictive reasoning over hidden or future states. Particular emphasis is placed on research from the past five years that connects isolated perception tasks with closed-loop embodied UAV operation. Moreover, we delineate critical open challenges in multi-UAV collaboration, sim-to-real adaptation, and trustworthy predictive evaluation, outlining a research

roadmap for moving these embodied agents from constrained simulations toward complex real-world operation.

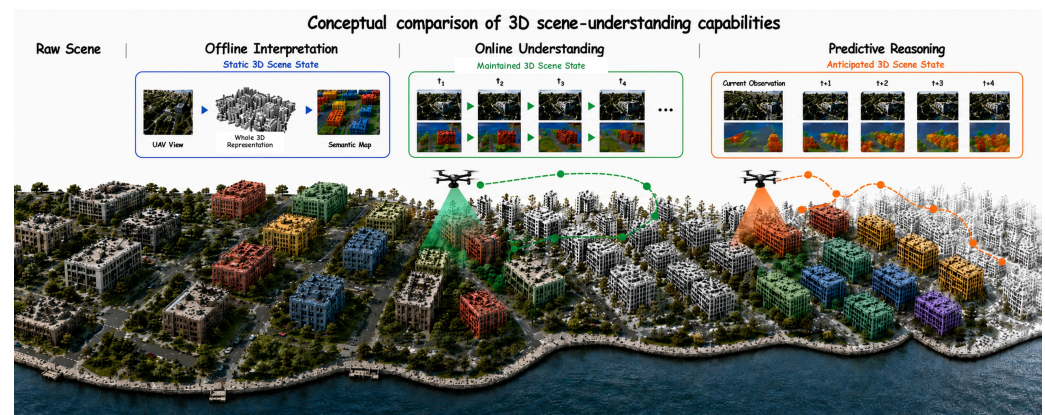


Figure 1. Overview of the proposed agent-capability evolution framework for UAV 3D scene understanding. The framework is organized by increasing observation–state–action coupling rather than by a mandatory serial pipeline. Offline interpretation constructs static scene states from post-acquisition data, online understanding maintains incremental scene states during flight, and predictive reasoning estimates hidden or future scene states to support subsequent sensing and action.

This review is organized as follows: Section 2 defines the conceptual boundary and taxonomy. Section 3 discusses data foundations for UAV 3D scene understanding and their differentiated support for the offline, online and predictive stages. Sections 4–6 detail the methodological evolution across the three agent-capability stages. Section 7 identifies key challenges in benchmark construction, online adaptation, collaborative perception and sim-to-real transfer. Section 8 concludes the review.

2. Conceptual Boundary and Taxonomy

This section defines the conceptual boundary of UAV 3D scene understanding and outlines the survey scope and taxonomy. Moving beyond isolated visual tasks, the central question addressed in this review is how a UAV transforms first-person observations into 3D scene states that support physical actions, and how these states evolve from offline interpretation products into online and predictive capabilities. Accordingly, the survey covers representative studies that directly target UAVs or low-altitude platforms, together with transferable mechanisms from photogrammetry, autonomous driving, indoor robotics and general embodied intelligence when they address geometry, semantic grounding, temporal state maintenance or prediction problems that UAV methods must solve. Such non-UAV studies are not treated as ready-to-deploy aerial solutions; they are interpreted through UAV-specific constraints including 6DoF aerial viewpoints, strict SWaP limits, onboard real-time inference, motion blur and vibration, coupling with flight controllers and safety-critical deployment.

2.1. Definition of UAV 3D Scene Understanding

Scene understanding extracts structural and semantic information from raw observations to interpret an environment. It is closely related to, but distinct from, 3D reconstruction. Reconstruction refers to the estimation of camera poses, depth, point clouds, meshes, radiance fields or Gaussian primitives. In this review, scene understanding further interprets these geometric substrates by assigning semantic categories, object boundaries, spatial relations, traversability and uncertainty. Reconstruction therefore provides metric evidence, whereas scene understanding converts that evidence into a 3D scene description that can be queried or used by an agent. Because UAVs interact with metric 3D space, image-plane

understanding alone is insufficient for reasoning about free space, occlusion, obstacle clearance and spatial reachability. Consequently, 3D scene understanding is defined here as the computational process of constructing a geometrically consistent and semantically enriched three-dimensional representation of the physical world. For example, in remote-sensing intelligent interpretation, such a representation supports hazard assessment, infrastructure inspection, and urban change monitoring by linking 3D geometry with semantics, spatial relationships, and task-relevant scene attributes.

Applying this concept to UAVs introduces additional constraints. UAVs execute 6DoF maneuvers in complex low-altitude environments under strict size, weight and power limits. Their scene understanding methods must therefore handle rapid viewpoint changes, dynamic objects, scale variation caused by oblique viewing angles, incomplete observations and onboard resource constraints. In this review, UAV 3D scene understanding refers to a UAV-centric 3D understanding process that constructs and updates a task-consumable scene state from ego-centric onboard observations, which is intended to provide an interactive spatial representation for subsequent tasks like localization, path planning, and navigation.

More specifically, in the UAV setting, this definition naturally leads to a constrained scene-state modeling problem under incomplete observations and limited airborne resources. Let $\Omega \subset \mathbb{R}^3$ denote the physical 3D workspace in which a UAV operates. At time t , let $\mathcal{O}_t$ denote the current first-person observation, $T_t \in SE(3)$ the UAV pose, u_{t-1} the previously executed control or action input, and $\mathcal{S}_{t-1}$ the scene state maintained before receiving $\mathcal{O}_t$. The continuous construction and updating of the UAV 3D scene state can then be abstracted as

$$\mathcal{S}_t = f_\theta(\mathcal{S}_{t-1}, \mathcal{O}_t, T_t, u_{t-1}), \quad \Omega_t^{\text{obs}} = \Omega_{t-1}^{\text{obs}} \cup \phi(\mathcal{O}_t, T_t), \quad \Omega_t^{\text{unobs}} = \Omega \setminus \Omega_t^{\text{obs}}. \quad (1)$$

where f_θ denotes a scene-state construction or updating model, $\phi(\mathcal{O}_t, T_t)$ denotes the physical region newly observed or sufficiently verified from observation $\mathcal{O}_t$ under pose T_t , Ω_t^{obs} is the accumulated observed or verified region up to time t , and Ω_t^{unobs} denotes the unobserved, occluded, or insufficiently verified region. The update of $\mathcal{S}_t$ is further subject to onboard resource constraints:

$$\text{Mem}(\mathcal{S}_t) \leq B_{\max}, \quad \text{Lat}(f_\theta; \mathcal{S}_{t-1}, \mathcal{O}_t) \leq \tau_{\max}, \quad \text{Pow}(f_\theta; \mathcal{S}_{t-1}, \mathcal{O}_t) \leq P_{\max}, \quad (2)$$

where $\text{Mem}(\mathcal{S}_t)$ denotes the memory required to store and maintain the scene state, $\text{Lat}(f_\theta; \mathcal{S}_{t-1}, \mathcal{O}_t)$ denotes the latency of one scene-state update, and $\text{Pow}(f_\theta; \mathcal{S}_{t-1}, \mathcal{O}_t)$ denotes the corresponding power consumption. $B_{\max}$, $\tau_{\max}$, and $P_{\max}$ represent the allowable onboard memory, real-time latency, and power budgets, respectively.

The above resource constraints further imply a representation-capacity boundary for the maintained scene state. Let N_t denote the number of active representation units in $\mathcal{S}_t$. Let b_s , c_s , and e_s denote the average memory cost, computational cost, and energy cost required to update one representation unit, respectively. Let η_c denote the onboard computational throughput, ν the scene-state update frequency, and Π the UAV payload configuration. Following UAV system studies that emphasize the joint influence of payload, onboard computation, and energy efficiency on autonomous aerial operation [23,24], the feasible representation capacity can be abstracted as

$$N_t \leq N_{\max}(\Pi) = \min\left(\frac{B_{\max}(\Pi)}{b_s}, \frac{\eta_c(\Pi)\tau_{\max}}{c_s}, \frac{P_{\max}(\Pi)}{e_s\nu}\right). \quad (3)$$

which extends the memory, latency, and power constraints above from scene-state updating to scene-state capacity. It indicates that the density and semantic richness of $\mathcal{S}_t$ are bounded

by payload-dependent onboard resources rather than by representation design alone. Observation geometry imposes an additional resolution-side boundary. Let H denote the flight altitude, f the effective focal length, ψ the off-nadir viewing angle, and κ a sensor-dependent scale factor. In accordance with UAV photogrammetry studies on off-nadir imaging and flight-geometry effects [25], the finest reliably observable spatial detail can be abstracted as

$$\Delta_{\text{obs}}(H, \psi) \gtrsim \kappa \frac{H}{f} \chi(\psi), \quad \chi(0) = 1, \quad \chi(\psi) \geq 1, \quad (4)$$

where $\chi(\psi)$ denotes the degradation factor induced by oblique viewing. If V_t denotes the physical extent represented by $\mathcal{S}_t$, the achievable representation granularity is therefore bounded by

$$\Delta_t \gtrsim \max \left(\Delta_{\text{obs}}(H, \psi), \left(\frac{V_t}{N_{\text{max}}(\Pi)} \right)^{1/3} \right). \quad (5)$$

which clarifies that UAV 3D representation fidelity is jointly limited by observation geometry and payload-dependent onboard resources. Higher altitude, more oblique viewpoints, larger represented space, or tighter resource budgets all increase the minimum feasible representation granularity. Therefore, the core contradiction is to maintain a scene state that is geometrically consistent, semantically informative, and uncertainty-aware, while the UAV observes only a partial subset of Ω and must satisfy strict resource constraints.

It should also be noted that UAV 3D scene understanding is not equivalent to the full autonomy stack. Localization, planning, decision-making and control remain downstream modules with their own objectives and constraints. The role of 3D scene understanding is to provide these modules with a task-consumable and uncertainty-aware scene state, including geometric structure, semantic labels, spatial relations, traversability, affordances and confidence estimates. In this sense, the 3D scene state can be regarded as a belief-like representation of the environment, although this review focuses on scene-state construction and use rather than complete belief-space planning or flight control [26].

This definition also provides the basis for the three-stage taxonomy used in this review. The taxonomy is organized according to the dominant agent-capability enabled by a stage, rather than according to its training source, implementation form, or possible downstream reuse. In this review, direct embodiment is defined by whether scene-state modeling is coupled with the UAV agent's in-flight perception–action requirements. The three stages therefore do not denote mutually exclusive algorithmic paradigms. Instead, they describe an agent-capability evolution in which UAV 3D scene understanding progresses from retrospective scene-level interpretation to situated in-flight understanding and further to anticipatory scene-state reasoning, with increasing coupling among observation, scene-state construction, and action.

The defining marker of offline interpretation is post-acquisition scene interpretation. Methods in this stage operate on scene-level 3D representations that have already been acquired or reconstructed, and transform them into task-consumable scene knowledge. This capability enables retrospective scene analysis, agent knowledge organization, map construction, and model pretraining. However, it does not constitute a directly embodied closed-loop capability, because the scene representation is not maintained in response to the UAV's ongoing perception–action demands in a previously unseen environment. Offline interpretation remains integral to the evolutionary chain because it establishes the data foundations, representation backbones, and semantic priors from which the following online understanding and predictive reasoning are developed.

The defining marker of online understanding is in-flight scene-state maintenance. Methods in this stage maintain an actionable 3D scene state from streaming onboard observations and UAV poses, enabling the agent to use the updated spatial state during

autonomous flight. In this role, UAV 3D scene understanding becomes an operational perception component within the perception–action loop, supporting flight decision-making and task execution. The use of pretrained models or offline-learned priors does not alter the stage assignment when the method is designed for online scene-state maintenance during UAV operation.

The defining marker of predictive reasoning is anticipatory scene-state reasoning beyond direct observation. Methods in this stage estimate unobserved, uncertain, or future scene states to support subsequent sensing or action. This capability enables the UAV agent to move beyond reactive use of the currently maintained scene state toward proactive autonomy, including uncertainty-aware sensing, risk avoidance, and forward-looking decision support. Representation fields, pretrained priors, and uncertainty estimates are therefore treated as enabling resources across stages, whereas stage assignment is determined by the dominant UAV capability that the method provides.

We first examine data foundations because datasets determine which scene states are observable, which supervision signals are available, and which capability boundaries can be evaluated. On this basis, recent methodologies are reviewed according to the three capability stages defined above. Table 1 summarizes their stage logic, input–output forms, operational subdimensions and capability boundaries.

Table 1. Summary of the three capability stages for UAV 3D scene understanding.

Capability Stage	Stage Logic	Input–Output Form	Operational Subdimensions	Capability Boundary
Offline Interpretation	Post-acquisition scene interpretation with weak perception–action coupling	Pre-acquired observations or reconstructed 3D representations → Static or queryable scene knowledge	Static Scene Interpretation; Single-UAV Training	Retrospective Analysis; Scene Data Organization and Annotation
Online Understanding	In-flight scene-state maintenance within perception–action loop	Streaming onboard observations and UAV poses → Incrementally maintained scene state	Dynamic Scene Awareness; Multi-UAV Online Perception	In-Flight 3D Perception; Online Decision Support
Predictive Reasoning	Anticipatory scene-state reasoning beyond direct observation	Maintained or partial scene states with sensing/action conditions → Hidden, uncertain, or future state estimates	Dynamic Scene Anticipation; Multi-UAV Predictive and Active Perception	Proactive 3D Sensing; Predictive Decision Support

The transitions among these stages are driven by increasing task demands and deployment constraints. Offline interpretation becomes insufficient when scene information must be updated or consumed during flight, such as in obstacle avoidance, target search or safety-critical navigation. Online understanding becomes insufficient when occlusion, latency, dynamic objects or unobserved space require the UAV to reason beyond the currently observed scene state. Predictive reasoning extends this boundary by estimating hidden or future scene states, but it remains constrained by uncertainty calibration, physical feasibility, onboard computation, action-state data availability and safety verification. It is worth noting that LLM/VLM-based modules mainly contribute to the online and predictive stages, where they support open-vocabulary scene querying, language-conditioned reasoning and vision–language–action coupling.

2.2. Survey Scope and Technical Taxonomy

To establish a rigorous survey scope, the literature selection primarily targets the recent research transition in which embodied-AI methods are increasingly coupled with UAV platforms. Foundational works in multi-view geometry, active vision, semantic scene completion, and world models are also traced back to establish theoretical roots [27–31]. Earlier studies mainly provided stable geometric and localization support through photogrammetry, point cloud learning, and visual-inertial estimation [32–35]. Recently, notable shifts have occurred as open-vocabulary mapping, vision-language-action models, and predictive world models have begun to directly target UAV task loops [36–40]. We therefore include work directly targeting UAV modeling alongside transferable technologies from adjacent fields, while evaluating how each work supports the 3D scene-understanding capabilities required at different UAV agent-capability stages. Figure 2 further visualizes these stage-wise shifts through representative milestone works, showing how UAV 3D scene understanding has evolved from post-acquisition offline scene interpretation to in-flight online scene-state maintenance and anticipatory scene-state prediction.

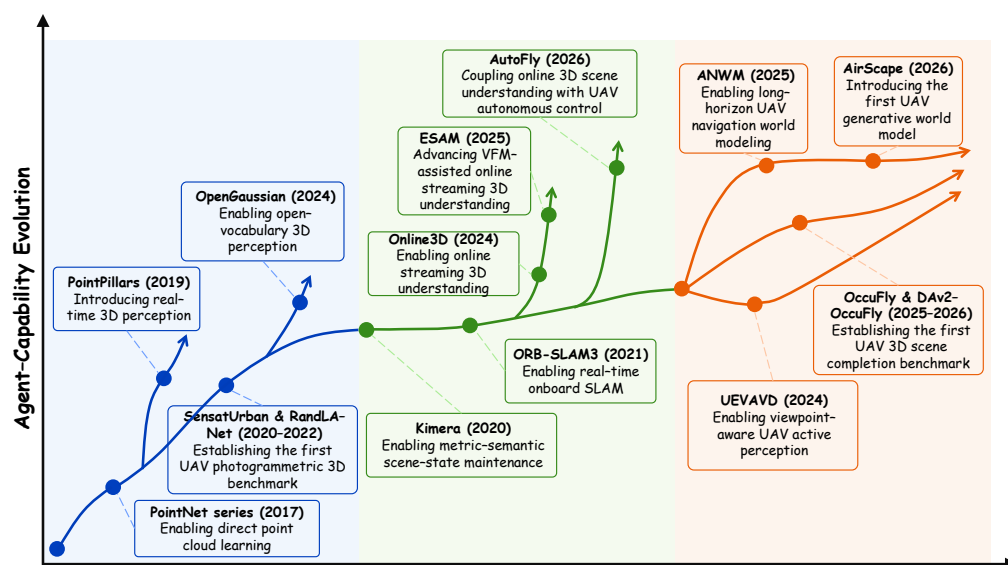


Figure 2. Stage-wise technical timeline of representative milestone works for UAV 3D scene understanding.

Existing UAV-oriented surveys provide important but different organizational views. Reviews on deep learning for UAVs and UAV remote sensing commonly organize the literature according to application scenarios, sensor types, task formulations and learning techniques [11,15]. Surveys on autonomous UAVs, visual SLAM and UAV-based photogrammetric 3D mapping further emphasize navigation pipelines, localization and mapping modules, data acquisition, image matching, aerial triangulation, dense reconstruction and deployment challenges [41–43]. These taxonomies are valuable for summarizing UAV methods, datasets and system components, but they do not explicitly describe how 3D scene-understanding requirements change as UAV agents move from post-acquisition analysis to in-flight updating and anticipatory reasoning. Therefore, the taxonomy in this review is intended as a complementary agent-capability-oriented view: it organizes UAV 3D scene understanding by the capabilities that methods provide at different agent stages and by the degree of observation–scene state–action coupling, rather than by application domain, sensor type, task label or reconstruction module alone.

Within this scope, cross-domain transfer is analyzed at the level of capability mechanisms rather than application domains. Photogrammetry and SLAM contribute metric consistency; autonomous-driving perception contributes BEV, occupancy and action-

conditioned world-modeling mechanisms; indoor robotics contributes active perception, scene graphs and language-grounded interaction; neural rendering contributes compact appearance and language fields. These mechanisms become UAV-relevant only after adaptation to aerial observation geometry, altitude-dependent scale changes, sparse or oblique coverage, sensor timing errors, communication and computation limits and flight-safety margins. This criterion is used throughout Sections 4–6 to distinguish transferable representation or reasoning mechanisms from UAV-ready deployment systems.

As detailed in Figure 3, we introduce a comprehensive taxonomy outlining the agent-capability-oriented organization of UAV 3D scene understanding methods. This taxonomy categorizes current methodologies into three stages. The offline stage covers foundational 3D representation construction, point cloud semantic learning, object-level understanding, and prompt-conditioned 3D semantic mapping, which are evaluated primarily for their capacity to achieve holistic 3D scene understanding. The online stage transitions into dynamic flight environments, detailing online semantic mapping, structured scene states, language-conditioned reasoning, and onboard implicit understanding to elevate scene understanding to a capability stage synchronized with UAV action execution. The predictive stage further extends the capability to anticipatory scene-state modeling by exploring 3D semantic scene completion, active perception, and action-conditioned embodied world models to support predictive reasoning over occluded and future scenes.

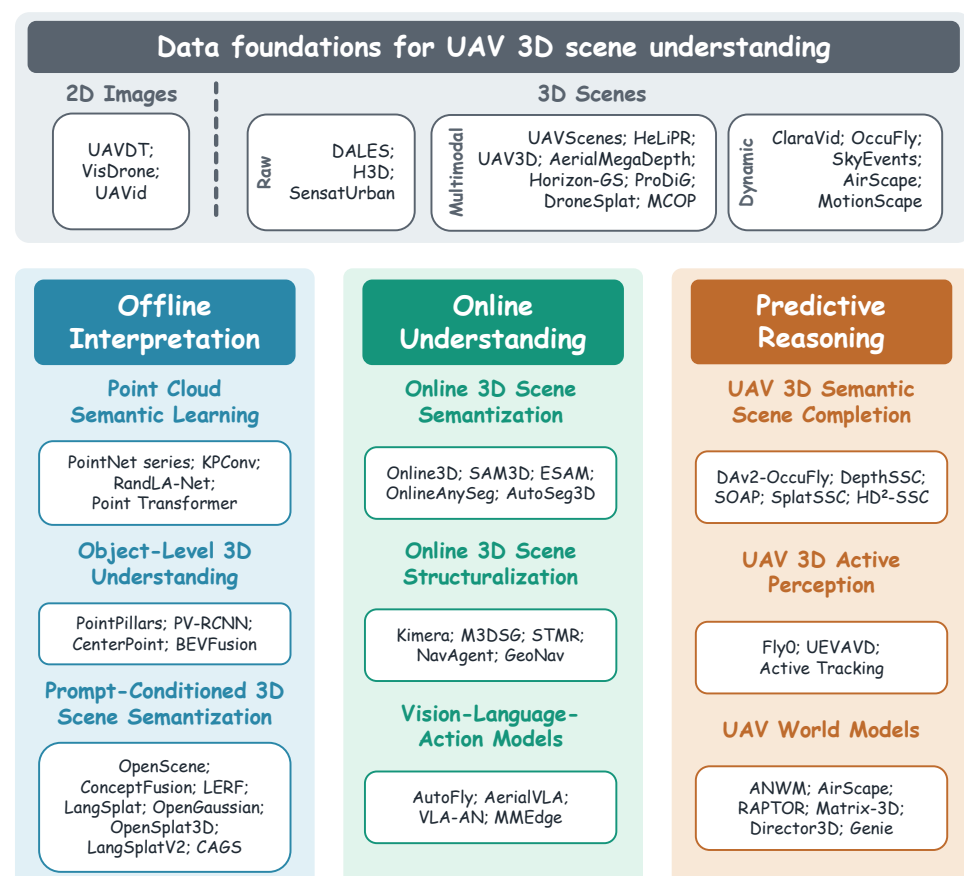


Figure 3. Agent-capability-oriented taxonomy and review structure of UAV 3D scene understanding, covering data foundations and the three stages of offline interpretation, online understanding and predictive reasoning.

3. Data Foundations for UAV 3D Scene Understanding

Data foundations determine the observable scene states, supervision signals and evaluation protocols available to UAV 3D scene understanding methods. Compared

with general 3D vision data, UAV data are characterized by large scale variation, rapid viewpoint changes, oblique observation angles, wide fields of view and complex low-altitude scene content. In the UAV literature, early learning-oriented benchmarks were predominantly built around 2D aerial imagery, because image-level annotations naturally supported detection, tracking and semantic segmentation tasks that were central to UAV remote sensing applications. In parallel, photogrammetry, LiDAR mapping and multi-view reconstruction datasets provided important 3D geometric resources, but these resources were mainly designed for reconstruction or mapping rather than for action-oriented UAV scene understanding.

Recent advances in 3D vision, multimodal foundation models and agent technology have shifted the data requirement from isolated image interpretation to unified scene-state learning. The UAV scene dataset increasingly needs to connect image-plane semantics with metric 3D geometry, temporally ordered observations, platform poses, uncertainty and, in more advanced settings, action-conditioned future states. Therefore, as summarized in Table 2, this section reviews the development of UAV data foundations from 2D aerial imagery-dominated perception benchmarks to UAV-view 3D scene data, and further analyzes their expansion toward multimodal and dynamic forms.

Table 2. Representative datasets, benchmarks and resources supporting the capability evolution of UAV 3D scene understanding.

Data Type	Representative Datasets	Research Focus	Key Assets	Capability Potential
2D Aerial Images	UAVDT [44], VisDrone [45], UAVid [46]	Aerial Detection, Object Tracking, Semantic Segmentation	RGB Images, 2D Labels, Visual Priors	Offline Recognition, 2D Semantic Priors, Image-Plane Evaluation
UAV-View 3D Scenes	DALES [47], H3D [48], SensatUrban [49]	Point Cloud Semantics, Static Mapping	3D Scenes, 3D Static Semantics	Offline 3D Interpretation, Static Scene States, Geometric Scene Priors
Multimodal 3D Scene Contexts	UAVScenes [50], HeLiPR [51], DroneSplat [52], AerialMegaDepth [53], Horizon-GS [54], ProDiG [55]	Multimodal Reconstruction, Pose-Aware Mapping	RGB/RGB-D Images, LiDAR Data, Poses, Reconstruction Targets	Online Mapping, Pose-Aware Scene States, Cross-Modal Fusion
	UAV3D [56], MCOP [57]	Collaborative Perception, Distributed Occupancy	Multi-Agent Views, Inter-Agent Poses, Communication Records	Collaborative 3D Perception, Distributed Scene States, Cooperative Occupancy
Dynamic Scene Streams	ClaraVid [58], OccuFly [59], SkyEvents [60], AirScape [61], MotionScape [62]	Dynamic Understanding, Future Prediction	3D Scenes, Video Streams, Action Trajectories, Future States	Predictive Reasoning, Action-Conditioned Forecasting, World Modeling

Rather than treating the three stages as a purely conceptual route, the dataset and benchmark evolution summarized in Table 2 provide empirical evidence for the proposed capability transition. The representative resources listed in this table cover 2D aerial image benchmarks, UAV-view 3D scene datasets, multimodal and collaborative 3D scene resources, and dynamic or predictive scene streams. This progression indicates

that UAV-related data foundations have expanded from image-plane detection, tracking and segmentation toward static 3D semantic interpretation, multimodal pose-aware mapping, collaborative 3D perception, semantic scene completion and action-conditioned future-state modeling.

3.1. From 2D Aerial Imagery to UAV-View 3D Scene

Early learning-oriented UAV datasets were predominantly organized around 2D aerial imagery. UAVDT [44], VisDrone [45] and UAVid [46] promote low-altitude traffic tracking, aerial object detection and high-resolution semantic segmentation, respectively. The DOTA series [63,64] extends large-scale object detection from aerial viewpoints, while ISPRS Vaihingen and Potsdam [65,66] provide standard aerial semantic annotation protocols. These datasets are valuable for UAV scene perception, but their annotations are defined on the image plane. As a result, they cannot fully describe metric geometry, free space, occluded regions or continuous 3D structural boundaries, which are necessary for UAV 3D scene understanding.

As UAV tasks increasingly require metric spatial reasoning, 3D annotations and 3D scene representations become necessary data foundations. Photogrammetry and multi-view geometry provide the basic production route for converting overlapping UAV images into camera poses, sparse point clouds, dense point clouds, meshes and textured surfaces [27,28,33,67]. These reconstruction techniques do not by themselves constitute scene understanding, but they determine the geometric substrate on which semantic labels, object boundaries, spatial relations and traversability attributes can be defined. Large-scale 3D datasets further support this transition. DALES [47] introduces large-scale aerial LiDAR data for point-level semantic categorization. H3D [48] provides high-density point clouds and textured meshes for detailed urban semantic segmentation. SensatUrban [49] further scales this paradigm by offering urban-scale photogrammetric point clouds with rich semantic annotations. Although these datasets advance static 3D scene understanding, they usually provide limited temporal continuity and weak action-state coupling.

The quality of UAV-view 3D scene data also depends on the robustness of the underlying representation construction pipeline. Feature matching methods such as SuperPoint [68], SuperGlue [69], LoFTR [70] and LightGlue [71] improve correspondence estimation under low texture, illumination variation and large viewpoint changes. Learning-based multi-view stereo methods including MVSNet [72], Recurrent MVSNet [73], CascadeMVSNet [74] and PatchmatchNet [75] strengthen dense depth recovery from multi-view observations. Monocular depth estimators such as Monodepth2 [76], DPT [77], Depth Pro [78] and Depth Anything [79] provide useful geometric priors when route overlap is insufficient or viewpoints are constrained. Neural radiance fields and 3D Gaussian Splatting further provide continuous and renderable scene representations [36,80–85]. In this review, these techniques are treated as data and representation foundations rather than as UAV 3D scene interpretation methods, because their main role is to provide pose-consistent and geometrically reliable 3D substrates for subsequent object-level interpretation, prompt-conditioned querying, and spatial memory construction for multimodal foundation models.

3.2. The Trend Toward Multimodal and Dynamic Data

To move beyond isolated image interpretation and static mapping, UAV 3D scene understanding data are increasingly expanding toward multimodal, multi-platform and dynamic recording. Multimodal datasets provide conditions closer to deployable UAV systems, where images, LiDAR, depth, poses, temporal information and cross-platform observations must be jointly maintained. UAVScenes [50] organizes synchronized RGB images, LiDAR and accurate poses into a unified annotation framework. HeLiPR [51]

targets heterogeneous LiDAR place recognition under spatiotemporal changes, while Paris-Luco3D [86] highlights cross-platform distribution shifts using target-domain LiDAR data. UAV3D [56] and MCOP [57] further extend UAV data toward multi-agent observation by providing benchmarks for multi-UAV 3D detection, tracking and collaborative occupancy prediction. Although OPV2V [87] and DAIR-V2X [88] are mainly designed for autonomous driving, they offer transferable paradigms for cross-agent perception, pose alignment and selective information sharing. These resources extend UAV scene data from single-modality perception toward pose-aware, temporally organized and multi-platform 3D scene states.

Recent UAV-oriented reconstruction resources also strengthen this data foundation by improving the completeness, consistency and cross-view usability of 3D representations. DroneSplat [52] improves 3D Gaussian reconstruction under wide baselines, sparse textures and exposure differences in in-the-wild drone imagery. AerialMegaDepth [53] emphasizes depth and appearance correspondence between aerial and ground views, supporting aerial–ground reconstruction and view synthesis. Horizon-GS [54] targets unified 3D Gaussian Splatting for large-scale aerial-to-ground scenes and mitigates inconsistency caused by scale and viewpoint changes. ProDiG [55] uses progressive diffusion guidance for Gaussian reconstruction from aerial to ground views. These resources are important because they convert UAV observations into more complete and transferable 3D scene substrates, which can support semantic annotation, prompt-conditioned spatial querying, online map initialization and predictive world-model training.

A further trend is the shift from static scene recording to dynamic and predictive data. When UAVs operate in closed-loop scenarios, conventional frame-based datasets may not fully capture rapid motion, illumination changes, occlusion evolution and action-conditioned future observations. ClaraVid [58] provides high-resolution multi-view aerial images with depth, panoptic segmentation and dynamic-object masks for dynamic scene reconstruction. OccuFly [59] provides a real camera-based aerial benchmark for semantic scene completion, enabling models to infer 3D semantic occupancy from partial UAV observations. SkyEvents [60] combines event cameras with UAV 3D reconstruction and provides event-enhanced data for high-speed motion and strong illumination changes. AirScape [61] emphasizes the relation between aerial motion controllability and future-view generation. MotionScape [62] provides large-scale real-world UAV videos with accurate pose trajectories and language descriptions to support action-conditioned world modeling. These datasets indicate that UAV data foundations are moving from static visual annotation toward multimodal, multi-platform and action-aware scene-state learning.

3.3. Stage-Specific Data Requirements

The purpose of reviewing data foundations is not merely to list available datasets, but to clarify what kind of scene state each capability stage can learn and evaluate. Offline interpretation, online understanding and predictive reasoning methods require different forms of observation, supervision and temporal organization. This subsection therefore connects the preceding dataset review to the three-stage framework by summarizing the data conditions needed for each stage.

The core data requirement for offline UAV 3D scene interpretation lies in the completeness of spatial representation and the diversity of scene types. To support generalization in complex environments, datasets should provide high-density geometric ground truth and fine-grained semantic labels while covering diverse scene categories, such as urban buildings, vegetation, infrastructure and indoor–outdoor transitions. The data collection process should provide sufficient route overlap to support stable metric reconstruction and reduce severe geometric holes. For frontier tasks such as prompt-conditioned mapping, datasets should also include well-aligned 3D spatial attributes and text descriptions. Such

comprehensive static data provide the foundation for learning global 3D scene structure and semantics.

Data for online UAV 3D scene understanding should follow a streaming sequential format and provide 3D information at both frame and scene levels. This requires not only single-frame 3D ground truth for instantaneous obstacle avoidance and local perception, but also continuous scene states for incremental mapping and global state maintenance. Unlike offline datasets that can exploit global visibility after acquisition, online datasets should preserve temporal causality so that models only access observations available at or before a given timestamp. These data are typically organized as continuous video streams, frame-by-frame LiDAR scans or synchronized multimodal sequences with temporally aligned UAV poses.

Training data for predictive UAV 3D scene reasoning should capture the relations among scene context, UAV actions and future scene evolution. For semantic scene completion, datasets should provide voxel-level masks that distinguish observed regions, inferred occluded regions and unknown spaces. For action-conditioned world models, datasets should record current first-person observations, flight control commands, future views and resulting trajectories in a synchronized form. Such paired action-state data help models estimate physical feasibility and reduce unsupported future-state hallucination, which refers to predictions that are inconsistent with observed geometry or physical constraints.

Overall, these data requirements show that the proposed capability evolution is reflected not only in method design, but also in benchmark construction. Offline interpretation is mainly supported by static geometry and semantic annotations, online understanding requires temporally ordered and pose-aware scene states, and predictive reasoning further requires supervision over hidden regions, future states and action-state transitions.

Correspondingly, evaluation criteria should also be stage-specific. Offline interpretation is mainly assessed by geometric accuracy, reconstruction completeness, semantic precision, object-level interpretation, mean Intersection over Union (mIoU), average precision (AP), occupancy accuracy and cross-view consistency. Online understanding should additionally consider temporal consistency, pose consistency, update frequency, memory cost, runtime latency, onboard energy consumption, localization drift and robustness under motion blur, vibration and communication constraints. Predictive reasoning requires evaluation of hidden-state completion, future-state consistency, uncertainty calibration, action feasibility, collision-risk prediction, safety violation rate and downstream task benefit. These criteria connect low-level perception quality with the capability evolution toward deployable UAV agents.

4. Offline UAV 3D Scene Interpretation

From an agent-capability perspective, the offline stage corresponds to interpreting a 3D scene after data acquisition. Its unifying criterion is not a single algorithmic paradigm, but the capability of converting post-acquisition UAV observations or reconstructed geometry into interpretable static 3D scene states. Figure 4 outlines this offline interpretation pipeline, in which the 3D scene representation is further interpreted into point-level semantics, object-level attributes and prompt-conditioned semantic maps. As summarized in Table 3, sensing, computation and action are temporally decoupled: the UAV collects multi-view observations along planned routes, and ground-side processing subsequently constructs and interprets a global 3D scene using relatively abundant computation and storage resources. In this stage, offline UAV 3D scene interpretation usually takes a complete 3D representation of the target scene as input and outputs semantic categories, object boundaries, spatial relations and task-related attributes for the entire scene.

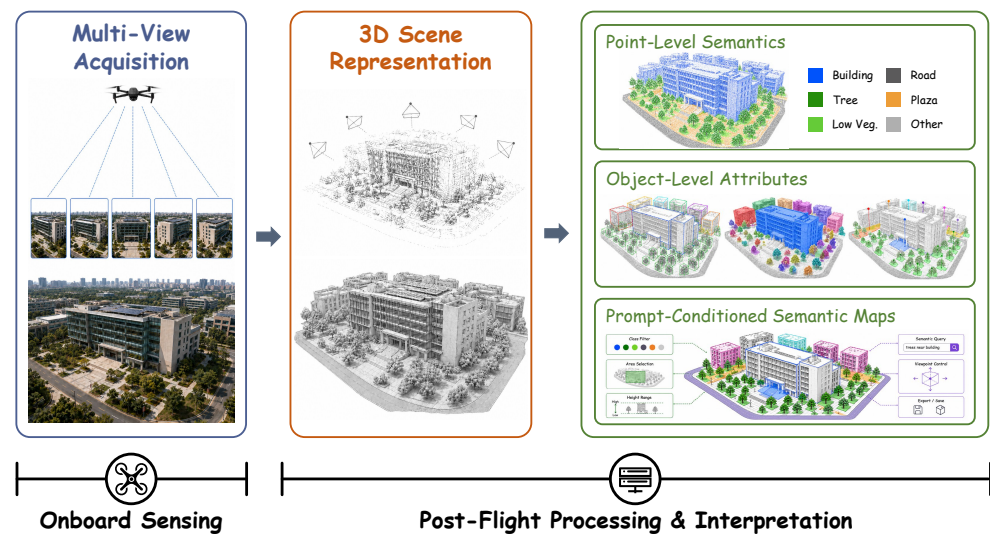


Figure 4. The technical pipeline of offline UAV 3D scene interpretation methods.

Table 3. Summary of offline UAV 3D scene interpretation methods, comparing their core capabilities and capability limitations.

Method Category	Representative Methods	Core Capabilities	Key Limitations
Point Cloud Semantic Learning	PointNet [89], PointNet++ [90], KPConv [91], RandLA-Net [92], Point Transformer [93]	Dense Geometric Parsing; Point-Level Labeling; Local Structure Learning; Scene-Scale Semantics	Large-Scale Processing Cost; Density-Shift Sensitivity; Closed-Set Semantics; Limited Temporal Updating
Object-Level 3D Understanding	PointPillars [94], PV-RCNN [95], CenterPoint [96], BEVFusion [97]	Object Detection; Metric Localization; BEV Spatial Abstraction; Target-Level Scene Indexing	Fusion Computation Overhead; Viewpoint-Occlusion Sensitivity; Sparse Scene Relations; Weak Flight Coupling
Prompt-Conditioned 3D Scene Semantization	OpenScene [98], ConceptFusion [99], LERF [100], LangSplat [37], OpenGaussian [101], OpenSplat3D [102], LangSplatV2 [103], CAGS [104]	Open-Vocabulary Grounding; Language-Conditioned Querying; Cross-View Feature Association; Interactive Map Access	Dense Feature Storage; Pose-Drift Sensitivity; Cross-View Semantic Pollution; Limited Onboard Validation

Accordingly, the method-oriented discussion below is anchored by three capability-transition markers: geometric-semantic parsing, object-centric spatial abstraction and queryable scene knowledge. These markers trace how offline methods make reconstructed UAV scenes progressively more task-consumable. Point cloud semantic learning assigns dense semantic meaning to geometric evidence, object-level understanding condenses this evidence into localized entities and target-level indices, and prompt-conditioned semanti-

zation makes static 3D maps accessible through language or task prompts. This progression clarifies that the offline stage advances the semantic accessibility of reconstructed scenes, but its scene state remains temporally decoupled from in-flight perception–action loops. Unlike pure 3D reconstruction, which primarily recovers geometry and appearance, offline UAV 3D scene interpretation emphasizes how the reconstructed scene is semantically organized, how objects and structures are identified, and how the resulting representation can provide static but comprehensive scene knowledge for later online understanding and predictive reasoning.

Table 4 reports numerical indicators for representative offline 3D scene interpretation methods relevant to UAV or low-altitude 3D scene understanding. The reported metrics include mean Intersection over Union (mIoU), overall accuracy (OA), average precision (AP), mean average precision (mAP), and mean accuracy (mAcc), together with latency, footprint and onboard adaptability. Because these methods are evaluated on different datasets and tasks, the numbers indicate reported performance evidence rather than direct rankings across methods.

Table 4. Quantitative indicators of representative offline UAV 3D scene interpretation methods.

Method	Primary Performance	Latency	Footprint	Onboard Adaptability
PointNet [89]	49.7% mIoU, 90.6% OA (SensatUrban)	>1 M points/s	3.5 M parameters; 440 M FLOPs/sample	Medium
PointNet++ [90]	58.1% mIoU, 93.1% OA (SensatUrban)	–	<2 M parameters	Medium
RandLA-Net [92]	58.6% mIoU, 91.6% OA (SensatUrban); 77.4% mIoU, 94.8% OA (Semantic3D); 70.0% mIoU, 88.0% OA (S3DIS)	185 s averaged by 4071 frames	1.24 M parameters	High
PointPillars [94]	59.2% mAP (KITTI); 44.5 mAP@0.5 (Ruby-128, V2U4Real); 30.6 mAP@0.5 (M1-Plus, V2U4Real)	62 Hz; 105 Hz with limited loss of accuracy	–	High
PV-RCNN [95]	52.4 mAP@0.5 (Ruby-128, V2U4Real); 31.6 mAP@0.5 (M1-Plus, V2U4Real)	99.0 ms/frame	5.4 GB training memory	Medium
CenterPoint [96]	48.1 mAP@0.5 (Ruby-128, V2U4Real); 33.6 mAP@0.5 (M1-Plus, V2U4Real)	11.0–16.0 FPS	153.5 G MACs; 4.6–8.7 GB training memory	Medium
BEVFusion [97]	53.6% mAP (ResNet-101, UAV3D); 48.7% mAP (ResNet-50, UAV3D)	8.4 FPS	253.2 G MACs	Medium
CAGS [104]	50.79% mIoU, 69.62% mAcc (LERF-OVS); 32.6% mIoU, 48.9% mAcc (ScanNet)	–	2.1 GB with 2-layer contextual propagation	Low–Medium

Note: “–” denotes contents not reported in the original or relevant papers. Onboard adaptability is qualitatively summarized according to latency, footprint, sensor dependence, and deployment evidence.

4.1. Point Cloud Semantic Learning

The transition from geometric reconstruction to 3D scene understanding first requires assigning semantic meaning to reconstructed aerial geometry. Early approaches projected

2D segmentations into 3D coordinate frames, which provided useful static annotations but remained dependent on image-level predictions and cross-view consistency [105–107]. Point cloud deep learning subsequently established a more direct path for processing 3D aerial structures. PointNet [89] and PointNet++ [90] learn point-wise features from unordered point sets, while KPConv [91], RandLA-Net [92] and Point Transformer [93] improve local geometric encoding, computational efficiency and long-range context modeling for large-scale point clouds. For UAV 3D scene understanding, these methods are important because they convert reconstructed surfaces into closed-set semantic layers such as buildings, roads, vegetation and infrastructure. However, their outputs are usually static labels over a completed map and therefore provide limited support for object-level reasoning, temporal updating and closed-loop flight decisions.

4.2. Object-Level 3D Understanding

Object-level understanding further organizes aerial scenes into countable and locatable entities. In low-altitude UAV applications such as facility inspection, disaster assessment and urban inventory, point-level labels are often insufficient because downstream tasks need object extents, metric positions and spatial relations. PointPillars [94], PV-RCNN [95] and CenterPoint [96] transform point clouds or voxelized space into object proposals and metric bounding boxes. BEVFusion [97] further illustrates how multi-sensor evidence can be projected into bird's-eye-view representations for object-centric reasoning. Although many of these architectures were originally developed for autonomous driving, their voxel and BEV formulations are transferable to UAV top-down analysis because they explicitly encode metric location and object scale. This transfer is nevertheless partial: driving systems often assume ground-level sensors, gravity-aligned motion and lane-constrained dynamics, whereas UAVs encounter oblique or top-down viewpoints, altitude-dependent scale changes, full 6DoF motion, vertical clearance requirements and thin aerial obstacles. BEV or voxel detectors are therefore treated here as object-centric representation mechanisms rather than UAV-ready scene-understanding systems. Their limitation is that the scene is still represented as a set of task-specific objects rather than as a complete, queryable and temporally actionable 3D scene state.

4.3. Prompt-Conditioned 3D Scene Semantization

Prompt-conditioned 3D scene semantization further extends this offline capability from closed-set categorization to language-queryable spatial semantics. Compared with point-level and object-level interpretation, this line of work aims to make reconstructed UAV scenes searchable by natural language prompts or open-vocabulary categories. Existing methods mainly follow two routes. The first route lifts 2D vision-language or foundation-model outputs into 3D geometry. Supported by SAM [108], CLIP [109], BLIP-2 [110], Grounding DINO [111] and LLaVA [112], OpenScene [98] and ConceptFusion [99] fuse language-aligned image features with 3D geometry to support point-level and region-level semantic queries. This route is relatively compatible with existing UAV image collections, but its reliability depends on camera pose accuracy, viewpoint coverage and cross-view consistency of the lifted 2D masks or features.

The second route embeds language-aligned features into neural fields, Gaussian primitives or other renderable 3D memories. CLIP-Fields [113], LERF [100] and LangSplat [37] improve prompt-conditioned localization by storing multi-scale language features in neural or Gaussian representations, while OpenGaussian [101], OpenSplat3D [102] and CAGS [104] further exploit explicit Gaussian primitives and instance boundaries to handle complex spatial topologies. This route provides richer queryable spatial memory, but increases storage, rendering and feature-fusion costs. For UAVs, both routes can produce useful preflight

semantic maps for critical elements such as wires or temporarily parked vehicles; however, they may also amplify the UAV error chain. Unstable 2D masks caused by top-down occlusion can lead to cross-view semantic pollution when projected into 3D space, and pose drift can write correct semantics into erroneous geometric positions. Neural fields and Gaussian language fields are therefore introduced here as representation foundations for queryable maps, not as evidence that generic open-vocabulary 3D grounding can be directly deployed onboard. UAV-oriented prompt-conditioned 3D understanding must couple semantic similarity with cross-view consistency, geometric visibility and quantified uncertainty.

4.4. Capability Boundary of the Offline Stage

Although the offline stage establishes important representation and semantic foundations for UAV scene understanding, its ability to support active flight remains limited. First, offline paradigms usually assume post-flight access to complete observations, meaning that the scene state relies on long-window posterior integration rather than real-time incremental updates. Second, the generated semantic geometries mainly support static mapping, annotation, semantic querying and model pretraining. They cannot directly replace onboard localization, obstacle avoidance, continuous language interaction or flight control. Third, offline methods often rely on ground-station computation, large storage capacity and human verification, which are difficult to reconcile with the size, weight and power constraints of onboard UAV platforms. While offline models can describe what a scene contains and how its geometry is organized, they have limited ability to answer such queries during active flight. This explains why many offline methods can perform strongly on completed point clouds, reconstructed maps or prompt-conditioned static queries, yet remain limited when the UAV must update uncertain scene states under changing viewpoints, motion blur, onboard resource limits and time-critical decisions. Post-flight reconstruction can identify low-coverage regions after processing all captured images, whereas a UAV agent must decide whether to turn, ascend or descend as new observations stream in. This gap between posterior static understanding and instantaneous in-flight decision-making motivates the transition from offline processing to deployable online 3D scene understanding.

This boundary can be further clarified through a cross-method comparison within the offline stage. Point cloud semantic learning provides the most direct geometric-semantic parsing of reconstructed UAV scenes, but its computational cost increases rapidly with large-scale, sparse or uneven point clouds, and its semantic reliability is usually constrained by closed-set labels and density shifts. Object-level 3D understanding offers a more compact representation for target localization, facility inspection and BEV-level spatial abstraction, yet this compactness comes at the cost of weaker fine-grained scene relations and reduced sensitivity to non-object structural cues. Prompt-conditioned 3D semantization improves open-vocabulary access to static 3D maps and makes offline representations more queryable for downstream tasks, but it relies heavily on stable pose registration, cross-view feature consistency and semantic calibration. These comparisons show that offline methods are most effective when high-quality post-acquisition reconstruction and semantic indexing are available, whereas their practical boundary appears when the scene state must be continuously updated, queried and verified under onboard latency, viewpoint change and flight-safety constraints.

5. Online UAV 3D Scene Understanding

The online stage represents a shift from post-acquisition interpretation to scene understanding during active flight. Its unifying criterion is the capability of maintaining an incremental 3D scene state. Figure 5 illustrates this in-flight understanding pipeline, where

streaming observations are continuously translated into scene states before being connected to downstream action interfaces. Rather than relying on offline processing, online 3D scene understanding serves immediate state estimation and task grounding. As summarized in Table 5, the input encompasses streaming sensory data including RGB imagery, depth, LiDAR, inertial measurements and historical states, while the output is an incrementally updated 3D scene state.

Within the online stage, the capability transition is marked by how a UAV transforms streaming observations into an action-usable scene state. Online neural representations first provide pose-consistent spatial memory; online semantization then attaches temporally maintained semantic and open-vocabulary cues to this memory; scene structuralization further converts maintained states into object-relation and topological-metric abstractions; and vision–language–action models expose the scene state to instruction-conditioned action interfaces. These markers connect the method categories to a common progression from spatial maintenance to semantic enrichment, relational organization and action-facing use. They also clarify the central boundary of online understanding: richer semantic and language interfaces improve task grounding, but require more reliable temporal association, uncertainty tracking, onboard computation and safety-aware updating under time-constrained flight.

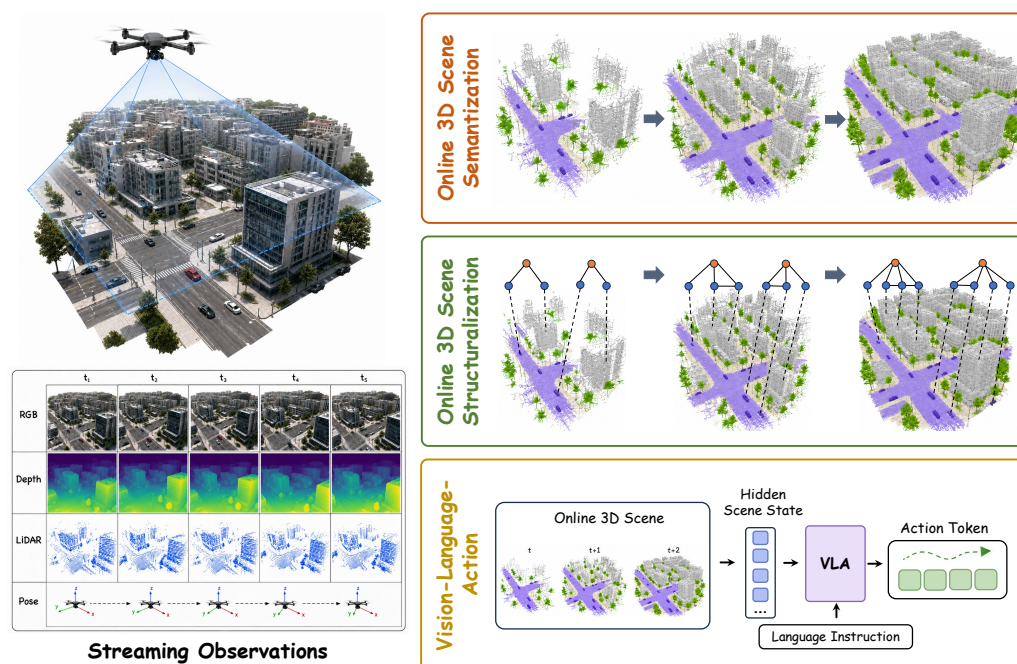


Figure 5. The technical pipeline of online UAV 3D scene understanding methods.

Table 6 summarizes quantitative indicators for representative online UAV 3D scene understanding methods. The reported metrics cover average trajectory error (ATE), average precision (AP), success rate (SR), object success rate (OSR), collision rate (CR), path efficiency ratio (PER), success weighted by path length (SPL), latency, and footprint, reflecting the operational requirements of in-flight scene-state maintenance.

Table 5. Summary of online UAV 3D scene understanding methods, comparing their core capabilities and capability limitations.

Method Category	Representative Methods	Core Capabilities	Key Limitations
Online Neural Representation	VINS-Mono [35], LIO-SAM [114], ORB-SLAM3 [34], Splat-Nav [115], SAGE-3D [116]	Incremental Pose Tracking; Metric Map Updating; Neural Spatial Memory; Pose-Consistent View Synthesis	Memory-Latency Tradeoffs; Dynamic-Scene Drift; Weak Semantic Grounding; Limited Safety Verification
Online 3D Scene Semantization	Online3D [117], SAM3D [118], ESAM [119], OnlineAnySeg [120], AutoSeg3D [121]	Streaming Semantic Updating; Instance-Level Association; Open-Vocabulary Grounding; Task-Conditioned Querying	Segmentation Inference Overhead; Pose-Noise Sensitivity; Category Hallucination; Limited Closed-Loop Validation
Online 3D Scene Structuralization	Kimera [122], M3DSG [123], STMR [124], NavAgent [125], GeoNav [126]	Scene Graph Construction; Metric-Topological Reasoning; Relation-Aware State Abstraction; Navigation-Oriented Structuring	Graph Update Cost; Front-End Detection Dependence; Accumulated Relation Errors; Uncertain Controller Integration
Vision–Language–Action Models	AutoFly [38], AerialVLA [127], VLA-AN [128], MMEdge [129]	Language–Action Grounding; Instruction Following; Action Proposal Generation; Edge-Control Interface	High Inference Latency; Platform-Generalization Gaps; Ambiguous 3D Grounding; Difficult Safety Verification

Table 6. Quantitative indicators of representative online UAV 3D scene understanding methods.

Method	Primary Performance	Latency	Footprint	Onboard Adaptability
ORB-SLAM3 [34]	3.6 cm ATE (EuRoC drone)	30–40 frames/s	<32 GB memory	High
AutoFly [38]	47.9% SR, 21.9% CR, 77.3% PER (simulation); 55.0% SR, 35.0% CR, 75.1% PER (real-world)	15 FPS; sub-100 ms response times	7B-parameter VLM backbone	High
ESAM [119]	42.2% AP (ScanNet200); 42.6% AP (ScanNet); 4.4% AP (ClaraVid)	7.3 FPS (ScanNet200); 9.8 FPS (ScanNet); 4.7 FPS (ClaraVid)	–	Medium
AerialVLA [127]	48.0% SR, 57.7% OSR, 38.5% SPL (UAV-Need-Help)	0.38 s/step	17 GB VRAM	High
VLA-AN [128]	94.6% SR (spatial grounding), 98.1% SR (object navigation), 85.7% SR (long-horizon navigation)	2–3 Hz onboard on Jetson Orin NX	Jetson Orin NX 16 GB, 30 W mode	Medium–High

Note: “–” denotes contents not reported in the original or relevant papers. Onboard adaptability is qualitatively summarized according to latency, footprint, sensor dependence, and deployment evidence.

5.1. Online Neural Representation

To achieve incremental updating and actionability, online understanding first requires robust geometric state estimation and efficient 3D representations. This layer does not by itself constitute semantic scene understanding, but it provides the pose-consistent scaf-

fold on which a UAV can maintain a scene state during flight. The state-update logic originates from visual SLAM, visual-inertial odometry, LiDAR-inertial fusion and incremental mapping. ORB-SLAM [130] and ORB-SLAM3 [34] demonstrate how visual features, loop closure and multi-map management support long-term localization. OKVIS [131] and VINS-Mono [35] utilize tightly coupled visual-inertial optimization to mitigate scale drift under fast aerial motion. LOAM [132], FAST-LIO [133] and LIO-SAM [114] further improve real-time geometry and state consistency through LiDAR odometry and factor-graph smoothing. These methodologies allow scene evidence to be accumulated during continuous observation rather than only after mission completion.

The representation choice determines what the maintained state can safely support. Probabilistic voxel maps explicitly represent occupied, free and unknown regions, making them suitable for collision checking and exploration under partial observations [134]. Euclidean Signed Distance Fields provide distance and gradient queries for local trajectory optimization, and aerial-ground systems such as Voxblox and FIESTA demonstrate that ESDFs can be incrementally maintained for onboard planning [135,136]. Bird's-eye-view maps compress multi-view observations into planning-friendly top-down layouts, although they may discard vertical structures such as wires, poles and tree crowns [137]. Neural implicit maps and 3D Gaussian maps provide appearance memory and view synthesis, but they often require additional geometric safety layers before their outputs can be used for control [138]. Methods transferred from ground robotics, indoor reconstruction or neural rendering also require high-rate inertial cues, calibrated camera-IMU timing and conservative occupancy interfaces before they can support aerial action. For UAV agents, a layered representation is therefore often more practical than a single unified map: visual-inertial odometry supplies high-frequency pose estimates, a local geometric map provides near-term safety constraints, a neural map stores appearance evidence, and a semantic layer supports language-level reasoning.

5.2. Online 3D Scene Semantization

Once pose-consistent spatial memory is available, geometric mapping alone remains insufficient because UAV actions depend not only on where surfaces are, but also on what they are and how reliable the semantic evidence remains over time. Online 3D understanding methods transferable to UAVs are therefore expanding from tracking basic spatial features to maintaining dynamic semantic and open-vocabulary 3D states, although many of them were originally designed for RGB-D streams, indoor scenes or general robotic perception rather than aerial platforms. Online3D [117] studies how 3D perception models can be transferred to streaming RGB-D inputs by using memory adapters to mitigate incomplete observations and category shifts. SAM3D [118] lifts masks generated by foundational models into 3D scenes through cross-view projection and point cloud aggregation, providing a basis for UAV multi-view semantic lifting. ESAM [119] targets efficient online scene segmentation with geometry-aware queries and lightweight decoding to reduce repeated calls to heavy 2D segmentation models. OnlineAnySeg [120] updates 3D instances as observations stream in through online mask association and voxel memory. AutoSeg3D [121] formulates online 3D segmentation as continuous instance tracking to maintain temporal consistency. Their UAV relevance depends on whether semantic updates remain stable under sparse depth, long-range scale changes, rolling-shutter distortion and intermittently observed objects.

For UAV platforms, these online semantization methods bridge metric 3D understanding and high-level language tasks. Traditional semantic mapping depends on predefined closed category sets, whereas prompt-conditioned methods allow the system to retrieve scene objects with task instructions. However, dynamically updated open-vocabulary

semantics do not inherently satisfy flight-safety requirements. Mask projection errors, onboard pose drift and category hallucination may be written into the maintained 3D state. Therefore, online semantic maintenance should be coupled with geometric consistency checks, cross-view label verification, temporal confidence tracking, and uncertainty calibration for open-vocabulary segmentation and language-grounded 3D labels.

5.3. Online 3D Scene Structuralization

Streaming semantic labels still do not by themselves form a task-consumable scene state. When a UAV performs long-horizon task execution, isolated geometric fragments and semantic points are insufficient for precise decision-making. The maintained scene state must organize maps and semantic labels into object instances, spatial relations and topological structures. The input of this layer consists of the metric and semantic observations maintained by the online mapping module, while the output is an object-level and topological scene state. Kimera [122] unifies metric maps and 3D dynamic scene graphs, enabling online perception to organize states in terms of actionable objects and spatial relations. M3DSG [123] demonstrates how 3D scene graphs serve as explicit memory for policy learning. In UAV-oriented systems, STMR [124] compresses semantic, topological and metric information into a structured representation directly parsable by large language models. NavAgent [125] emphasizes multi-scale urban spatial fusion, indicating that aerial agents require structural memory across local observations and broader geospatial context. GeoNav [126] introduces dual-scale geospatial reasoning, showing that map-level structure and local visual evidence can be jointly organized for 3D spatial grounding. These methodologies convert continuous perception into intermediate interfaces usable by symbolic planning, multimodal reasoning and language-conditioned scene-state interpretation.

Scene structuralization is critical for aerial platforms because UAV missions frequently involve long-range spatial expressions and viewpoint-dependent constraints. A frame-level image caption or a flattened object list is insufficient for such tasks. The onboard system must continuously maintain the localization of geometric shapes, the identity of object instances and the validity of topological relations under accumulative pose uncertainty. This structured representation allows the system to determine whether a spatial relation originates from direct observation or inference, thereby providing a more reliable foundation for further reasoning.

5.4. Vision–Language–Action Models

Vision–language–action (VLA) models expose the maintained and structured scene state to language-conditioned action interfaces. They connect online scene understanding with executable control by transforming visual observations and language instructions into action commands under vehicle dynamics and environmental constraints. In the context of online UAV 3D scene understanding, VLA models are relevant because they require the maintained scene state to be not only perceptually accurate but also actionable for navigation, inspection and interaction. Unlike offline language grounding, UAV VLA systems must interpret instructions while continuously updating spatial context during flight. AutoFly [38] represents progress in complex outdoor UAV navigation by utilizing coarse language directions coupled with visual observations, and introduces pseudo-depth features to compensate for the geometric insufficiency of monocular RGB. AerialVLA [127] introduces progress awareness and online questioning, allowing a UAV to request clarification when instructions face environmental ambiguity. VLA-AN [128] connects geometric safety correction to the action generation chain, focusing on making actions executable under flight dynamics and collision constraints.

The online deployment of these VLA models is restricted by SWaP limits and by the need to couple language-conditioned decisions with flight controllers. General techniques such as model compression, quantization and low-rank adaptation are adopted [139–143], while preserving structural safety margins. MMEdge [129] demonstrates that multimodal edge inference is limited by the sequential pipeline from sensing to encoding and decision-making. Its speculative skipping mechanism trades slight modality completeness for reduced closed-loop latency. Unlike desktop or indoor VLA settings, aerial systems cannot treat generated textual or action tokens as executable commands unless they are mapped to controller-feasible trajectories and checked by geometric safety layers. Thus, online closed-loop scene understanding is a coupled problem encompassing vision, hardware systems and flight control.

5.5. Capability Boundary of the Online Stage

The online stage integrates scene understanding into the active flight process, but its primary focus is current scene contexts rather than proactive prediction. Current methodologies focus on stable interpretation of the observed environment. They update the geometric shapes, topological layout and object instances of what has already been detected. However, they frequently show limited capability in inferring hidden structures behind immediate occlusions or extrapolating the future spatial consequences of candidate maneuvers. In complex low-altitude aerial scenes, this boundary poses limitations. A UAV may accurately maintain a local map of visible building facades, yet fail to anticipate whether a traversable corridor exists behind an upcoming corner. It may segment a structural pole in its current field of view, but fail to predict that its next turn will expose a thin cable crossing its flight path.

This boundary also explains a common gap between benchmark performance and UAV deployment. Online semantic or relational models may perform well when camera poses, frame rates and object categories are controlled, but their maintained scene states can degrade under fast 6DoF motion, vibration, motion blur, intermittent observations, association errors and onboard latency. Therefore, online understanding should be evaluated not only by semantic accuracy, but also by whether the maintained state remains temporally valid and usable for downstream navigation or interaction.

This observation defines the boundary between online understanding and predictive reasoning. Online methods primarily answer how to maintain a reliable representation of the currently observed scene during flight, whereas predictive methods further ask what may exist beyond the visible horizon and how the scene may evolve if a specific action is executed. Consistent with the stage-specific evaluation criteria discussed in Section 3, online UAV 3D scene understanding should not be evaluated only by frame-level or map-level perception metrics, such as mean Intersection over Union (mIoU) or average precision (AP). Since online scene states are expected to support closed-loop flight, they should also be assessed by compact task- and system-level indicators, including success rate (SR), success weighted by path length (SPL), collision rate (CR), intervention rate (IR), runtime latency (LAT), and onboard energy consumption (EC). The pursuit of predictive understanding therefore naturally emerges once the limitations of current-view online cognition are exposed in such closed-loop settings [144–148].

The practical boundary of the online stage can also be specified by comparing how different method families balance state maintenance, semantic reliability and deployability. Online neural representations provide the geometric backbone for pose-consistent spatial memory, but their efficiency is limited by incremental map updates, memory growth and rendering or optimization latency, especially under fast UAV motion. Online 3D scene semantization improves the semantic usability of the maintained state by introducing

labels, instances and open-vocabulary concepts, but its robustness depends on whether segmentation, association and language grounding remain stable under pose noise and view changes. Scene structuralization compresses the maintained state into objects, relations and topological-metric graphs that are more accessible to planners, although its accuracy remains coupled to front-end detection quality and accumulated relation errors. Vision–language–action models further expose the scene state to instructions and action proposals, but their onboard use is constrained by inference latency, platform generalization and safety verification. Overall, online methods improve the actionability of UAV scene states, while their deployability depends on whether geometric updating, semantic maintenance and controller-facing interfaces can remain reliable in closed-loop flight.

6. Predictive UAV 3D Scene Reasoning

The predictive stage extends online UAV 3D scene understanding from direct scene-state maintenance to anticipatory scene-state reasoning under partial observations, which is achieved in a broad sense that includes both spatial and temporal prediction. It is therefore defined as anticipatory UAV 3D scene understanding: a capability stage in which heterogeneous methods estimate hidden, uncertain, or future 3D scene states beyond direct observation to support subsequent sensing or action. This stage is not a single predictive task or algorithmic paradigm. Figure 6 illustrates such predictive capabilities from partial observations to semantic scene completion, active viewpoint selection and world modeling. The input of this stage consists of current or historical observations, maintained scene states and candidate trajectories. The output includes conditional estimates of hidden structures, uncertainty-reduction cues, or future scene states.

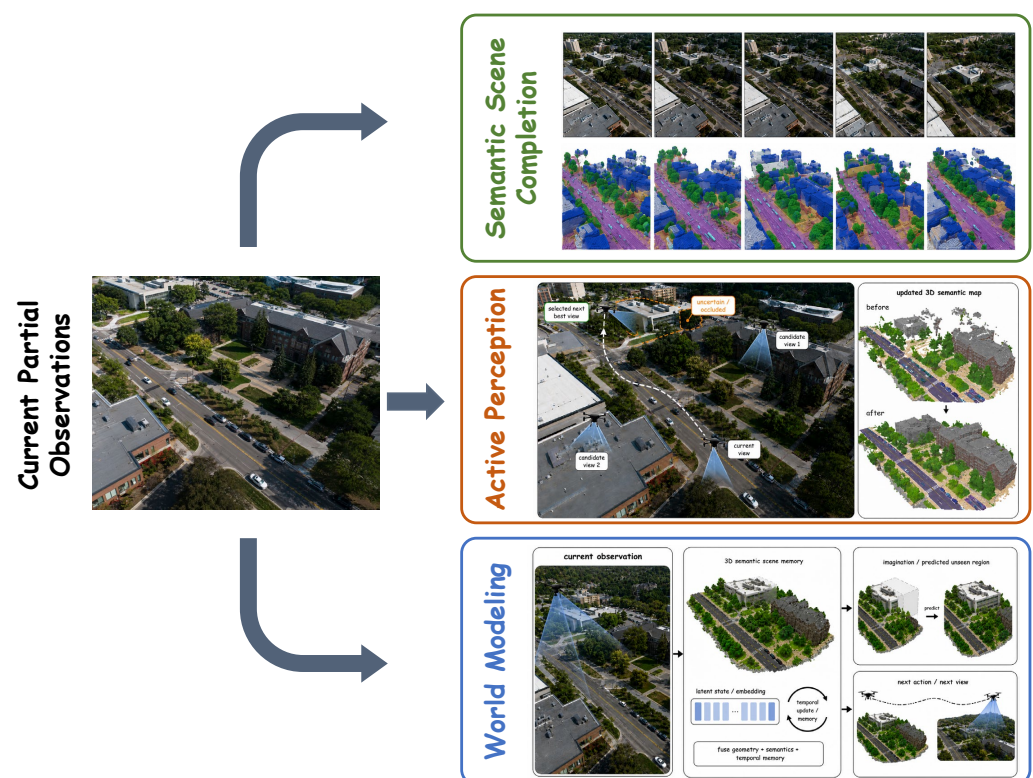


Figure 6. The technical pipeline of predictive UAV 3D scene reasoning methods.

Following the general formulation above, S_t denotes the maintained UAV 3D scene state, Ω_t^{obs} and Ω_t^{unobs} denote the observed and unobserved regions, and u_t denotes a candidate sensing or motion action. Semantic scene completion aims to predict the currently

unobserved part of the scene state. Consistent with the standard definition of semantic scene completion as predicting volumetric occupancy and semantic labels from partial observations [30], it can be abstracted as

$$\hat{\mathcal{S}}_t^{\text{unobs}} = \arg \max_{\mathcal{Z}} p_{\theta}(\mathcal{Z} \mid \mathcal{S}_t, \Omega_t^{\text{unobs}}), \quad (6)$$

where $\hat{\mathcal{S}}_t^{\text{unobs}}$ denotes the inferred occupancy-semantic state of unobserved or occluded regions, and $\mathcal{Z}$ denotes a candidate hidden scene-state completion. Active perception instead selects a sensing or motion action to improve subsequent scene-state estimation [149]:

$$u_t^* = \arg \max_{u_t \in \mathcal{U}_t} [\mathcal{G}(\mathcal{S}_t, u_t) - \lambda C(u_t)], \quad (7)$$

where $\mathcal{U}_t$ is the feasible action set, $\mathcal{G}(\mathcal{S}_t, u_t)$ denotes the expected gain in scene-state quality, such as information gain, uncertainty reduction, or coverage improvement, and $C(u_t)$ denotes the action cost. In contrast, a world model predicts how the scene state evolves based on state-action dynamics under partial observability [26]:

$$p_{\theta}(\mathcal{S}_{t+k} \mid \mathcal{S}_t, u_{t:t+k-1}), \quad k \geq 1. \quad (8)$$

As summarized in Table 7, the predictive stage is marked by increasingly anticipatory forms of scene-state reasoning. Semantic scene completion extends the current scene state by inferring hidden occupancy and semantics, making partially observed space more usable for risk-aware planning. Active perception shifts the marker from completion to uncertainty-guided sensing, where candidate viewpoints or sensing actions are evaluated according to their expected contribution to future state estimation. World models further move the marker toward action-conditioned future-state reasoning by estimating scene evolution or action consequences before physical execution. This progression distinguishes hidden-state inference, observation selection and future prediction, while making the internal capability transition of the predictive stage visible in the detailed discussion.

Table 7. Summary of predictive UAV 3D scene reasoning methods, comparing their core capabilities and capability limitations.

Method Category	Representative Methods	Core Capabilities	Key Limitations
UAV Semantic Scene Completion	DAv2-OccuFly [59], DepthSSC [39], SOAP [150], SplatSSC [151], HD ² -SSC [152]	Hidden Geometry Inference; Occupancy-Semantic Completion; Occlusion-Aware Reasoning; Risk-Aware Planning Support	Volumetric Prediction Cost; Altitude-Viewpoint Sensitivity; Unknown-Space Overconfidence; Limited Planning Validation
UAV Active Perception	Fly0 [153], UEVAVD [154], Active Tracking [155–157]	Information-Gain Planning; Uncertainty-Reducing View Selection; Active Scene Exploration; Sensing-Motion Tradeoff Modeling	Online Optimization Cost; Reachability-Energy Constraints; Utility Estimation Uncertainty; Risk–Reward Validation Gaps
UAV World Models	ANWM [40], AirScape [61], RAPTOR [158], Matrix-3D [159], Director3D [160], Genie [161]	Action-Conditioned Prediction; Consequence Evaluation; Long-Horizon Imagination; Future-State Planning Support	High Model Cost; Physical Consistency Gaps; Long-Horizon Error Accumulation; Insufficient Safety Calibration

Table 8 reports quantitative indicators for representative predictive UAV 3D scene reasoning methods. Since predictive reasoning includes hidden-state completion, active perception, and future-state prediction, the reported metrics cover absolute relative error (AbsRel), mean absolute error (MAE), Intersection over Union (IoU), mean Intersection over Union (mIoU), recognition accuracy, average trajectory error (ATE), relative pose error (RPE), navigation error (NE), Fréchet inception distance (FID), Fréchet video distance (FVD), and intention alignment rate (IAR).

Table 8. Quantitative indicators of representative predictive UAV 3D scene reasoning methods.

Method	Primary Performance	Latency	Footprint	Onboard Adaptability
DAv2-OccuFly [59]	0.1 AbsRel, 2.7 MAE (OccuFly 30 m); 0.1 AbsRel, 3.8 MAE (OccuFly 40 m); 0.1 AbsRel, 5.3 MAE (OccuFly 50 m)	–	24.8 M parameters	Medium
SplatSSC [151]	62.83% IoU, 51.83% mIoU (Occ-ScanNet); 61.47% IoU, 48.87% mIoU (Occ-ScanNet-mini)	115.56 ms in the 1200-primitive setting	3.0 GB memory in the 1200-primitive setting	Medium
UEVAVD [154]	0.610 recognition accuracy, −0.059 return, 4.065 path length (UEVAVD hard set)	maximum 5 decision steps per episode	ResNet-50 + GRU + DuelingDQN	Medium
ANWM [40]	60.0% SR, 8.13 ATE, 1.06 RPE, 14.12 NE (3D navigation)	High (Generative inference)	8 CDiT blocks	Low–Medium
AirScape [61]	111.89 FID, 1043.24 FVD, 84.51% IAR (future-view prediction)	High (Generative inference)	CogVideoX-i2v-5B foundation model	Low–Medium

Note: “–” denotes contents not reported in the original or relevant papers. Onboard adaptability is qualitatively summarized according to latency, footprint, sensor dependence, and deployment evidence.

6.1. UAV 3D Semantic Scene Completion

Within the predictive stage, semantic scene completion corresponds primarily to hidden current-state inference rather than full action-conditioned forecasting. It aims to infer both occupancy and semantic labels of a 3D volume from partial observations [30]. Its input typically comprises monocular images, depth priors or voxelized local observations, and its output provides occupancy probabilities and semantic categories within a complete volume. In autonomous driving, architectures including SSCNet [30], SketchAware [162], LMSCNet [163], MonoScene [164], TPVFormer [165] and OccFormer [166] established mainstream paths from sensor inputs to voxel or tri-plane representations. For aerial platforms, this task requires explicitly handling top-down viewpoints that observe roofs or upper facades while missing lower facades, interior spaces and local free passages.

Recent studies adapt mainstream completion baselines to evaluate their structural reasoning capability under oblique aerial observations. The OccuFly benchmark introduces DAv2-OccuFly [59], which combines foundational depth estimation with occupancy prediction to evaluate how monocular aerial observations can be lifted into semantic 3D volumes. Its results reveal that aerial completion performance strongly depends on flight altitude, viewing angle and the frequency distribution of semantic categories. DepthSSC [39] employs depth-spatial alignment and voxel adaptation to reduce geometric distortion during the lifting of monocular UAV observations to 3D voxels. SOAP [150] employs scene-adaptive decoding and occlusion-aware projection to suppress the erroneous mapping of visible aerial features into hidden regions. SplatSSC [151] introduces Gaussian primitives

into the completion pipeline by using depth-guided initialization and decoupled aggregation to minimize initialization artifacts. HD²-SSC [152] provides mechanisms for bridging high-dimensional inputs and high-density semantic outputs under sparse observations. These methods collectively indicate a shift from describing observed UAV-view surfaces to probabilistically inferring unobserved 3D spaces.

For UAVs, the utility of 3D semantic scene completion extends beyond estimating hidden surfaces to defining traversability and risk boundaries. Unobserved aerial spaces frequently result from top-down occlusion, height variations and changing platform attitudes. Completion models must therefore distinguish observed occupancy, inferred occupancy and unknown spaces to expose geometric uncertainty to planners rather than overwriting unknown regions with unverified estimates.

6.2. UAV 3D Active Perception

Active perception addresses how to select the subsequent observation to reduce spatial uncertainty [29]. It is included in the predictive stage not because it is equivalent to task planning, but because it evaluates how candidate viewpoints may change the future observable scene state. For UAVs, this paradigm is inherently suitable because the platform can actively adjust altitude, viewpoint and distance. Early next-best-view and active reconstruction studies demonstrated that viewpoint selection improves geometric coverage and reconstruction quality [167,168]. In UAV scenarios, active perception couples geometric information gain with flight risk, energy consumption and dynamic feasibility.

Recent methodologies introduce semantic and geometric priors into active perception frameworks. Fly0 [153] decouples semantic grounding and geometric planning. Foundation models identify task-relevant targets, while a geometric planner computes observation paths that satisfy safety constraints and information-gain objectives. UEVAVD [154] implements deep reinforcement learning baselines equipped with scene pre-decomposition and memory-based state estimation to search for viewpoints that reduce aerial occlusion. End-to-end active tracking [155], reinforcement learning for active tracking [156] and semantics-aware exploration planning [157] further demonstrate the relevance of active sensing behavior to predictive reasoning, where prediction involves reorganizing future observations rather than passively completing invisible regions.

Active perception requires balancing information acquisition with operational costs. For UAVs, extended geometric coverage necessitates complex maneuvers and greater near-obstacle proximity. Next-view selection is therefore a multi-objective optimization problem encompassing information gain, flight safety and resource constraints. A robust active-perception system must evaluate viewpoint reachability, obstacle margin preservation and whether the expected uncertainty reduction justifies the dynamic motion cost.

6.3. UAV World Models

World models form the most action-coupled family within the predictive stage because they connect spatial prediction with executable actions. They evaluate how the future scene will evolve if a specific action is executed [31,169,170]. The input incorporates historical observations, the maintained scene state and candidate trajectories, while the output generates a sequence of future images, spatial grids or latent states. For UAVs, these architectures compare the potential consequences of multiple candidate trajectories prior to physical execution to support long-horizon decision-making.

Recent implementations adapt world modeling paradigms to highly dynamic aerial environments. Among UAV-oriented studies, Aerial Navigation World Model [40] generates visual observations from historical frames and planned control actions, and injects geometric priors into prediction through future-frame projection. AirScape [61] binds

future-view generation to aerial motion controllability, making the generated view sequence more relevant to UAV action evaluation. RAPTOR [158] targets high-resolution UAV video prediction and addresses the tension between prediction fidelity and real-time inference through efficient video attention.

Transferable generative world models broaden this direction at the mechanism level rather than providing directly deployable UAV systems. Matrix-3D [159] is relevant to UAV research because omnidirectional explorable 3D world generation can provide controllable environments for aerial viewpoint planning and future-view synthesis. Director3D [160] contributes camera-trajectory-conditioned 3D scene generation, offering a transferable mechanism for evaluating UAV viewpoint sequences before real execution. Genie [161] demonstrates generative interactive environments, which inspire action-conditioned rollout interfaces for autonomous agents even though its original setting is not aerial. DriveDreamer [171] and UniSim [172] further provide transferable ideas for trajectory-conditioned simulation and interactive real-world modeling. However, driving-oriented world models often assume ground-constrained trajectories, while interactive simulators may use abstract actions that are not tied to UAV flight dynamics. Their use for UAVs therefore requires adapting action representations to 6DoF pose evolution, inertial feasibility, altitude-dependent scale and safety-certified controller interfaces. Predictive research is therefore shifting from passive next-frame prediction toward active action-effect prediction grounded in 3D space.

UAV world models must address the gap between abstract action spaces and physical flight dynamics. Navigation systems frequently use simplified control interfaces such as horizontal displacement and yaw, whereas actual UAV motion involves full 6DoF pose evolution, together with coupled velocity, roll, pitch and actuation constraints. Models restricted to statistical relations between simplified actions and future images may generate physically infeasible rollouts under aggressive maneuvers. UAV world models therefore require pose propagation, inertial consistency and geometric projection priors before their outputs can be treated as actionable world states. The utility of world models in UAV research extends beyond future-video generation to the provision of state estimates related to 3D space, traversability and task risk. If an architecture generates smooth visual frames but fails to preserve metric geometry or share an explicit uncertainty interface with the planner, its contribution to structural UAV autonomy remains limited.

6.4. Capability Boundary of the Predictive Stage

The predictive stage moves UAV 3D scene understanding from maintaining the current scene toward inferring spaces outside the field of view and evaluating future events. Its main value is to help systems assess risks and organize actions under structural occlusion and sparse observations. However, geometrically inconsistent future rollouts may introduce more severe planning risks than explicitly marked unknown regions, because they can transform uncertain hypotheses into apparently actionable scene states.

Within the scope of this review, predictive reasoning is the most anticipatory form of UAV 3D scene understanding, but it is not the endpoint of autonomous UAV intelligence. It provides scene-state estimates, uncertainty cues and possible action consequences for downstream modules, while task planning, decision-making, control execution and safety verification remain outside the primary focus of this survey.

We use inference to denote estimating hidden current states from available observations, prediction to denote estimating future or counterfactual states, planning to denote selecting actions to achieve task objectives, and reasoning to denote the structured use of inferred or predicted scene states for decision support. Active perception therefore lies at

the interface between prediction and planning, but it is not treated as full task planning in this review.

Predictive 3D understanding requires calibrated constraints rather than unconstrained generation. The main trade-off in this stage is between generative flexibility and verifiable reliability: richer future-state hypotheses can improve anticipation, but they become useful for UAV agents only when their geometric consistency, physical feasibility and uncertainty are explicit. The outputs of semantic scene completion, active viewpoint selection and action-conditioned world models should satisfy geometric consistency, physical feasibility and uncertainty quantification. For semantic completion, systems should differentiate observed structures, inferred geometry and unknown spaces. For active perception, selected viewpoints should balance expected information gain with reachability, dynamic feasibility and energy constraints. For world models, predicted future frames or latent states should remain consistent with camera pose, spatial scale and flight dynamics to avoid misleading the planning module.

Predictive architectures should support decision comparison rather than replace planning, decision-making or low-level control. Because this review focuses on 3D scene understanding as an upstream perception and scene-state modeling capability, predictive outputs should be treated as inputs to planners and controllers, not as executable commands. A reliable predictive model should indicate which structural regions a candidate trajectory may expose, distinguish risks derived from verified observations versus model hypotheses, and determine whether uncertainty warrants additional sensing, deceleration, trajectory replanning or human intervention. Prediction meaningfully expands UAV 3D scene understanding when its outputs are interpretable by planners, compatible with controller interfaces, bounded by safety layers and verified through metric and closed-loop evaluation.

The practical boundary of the predictive stage differs according to the form of anticipatory scene state being produced. Semantic scene completion is relatively direct because it estimates hidden occupancy and semantics from partial observations, thereby supporting more complete risk-aware planning, but dense volumetric or Gaussian-based completion can be computationally expensive and may produce overconfident predictions in unobserved space. Active perception shifts the objective from completing the current state to selecting observations that reduce uncertainty; this can improve robustness to viewpoint change when feasible viewpoints are available, but it introduces online optimization, reachability, energy and flight-risk constraints. World models provide the most action-aware form of prediction by estimating future scene states or action consequences, yet they impose stronger requirements on physical consistency, long-horizon stability and uncertainty calibration. Thus, semantic scene completion mainly improves the completeness of the current 3D state, active perception improves the reliability of future observations, and world models support planning-oriented consequence evaluation. Their shared deployment boundary lies in whether predictive outputs can be converted into calibrated planning evidence rather than being treated as unverified executable decisions.

7. Key Challenges and Future Directions

The preceding sections reviewed the data foundations of UAV 3D scene understanding and the methodological evolution across offline interpretation, online understanding and predictive reasoning. The remaining gaps should be distinguished at scientific, technical and system levels. At the scientific level, UAV 3D scene understanding requires uncertainty-aware scene-state representation, a principled boundary between reconstruction and understanding, open-world semantic grounding and action-conditioned prediction under partial observability. At the technical level, it requires real-time multimodal fusion, temporal

confidence tracking, cross-view consistency, representation compression, collaborative scene-state alignment and physically plausible prediction. At the system level, it must satisfy SWaP constraints, onboard latency, communication bandwidth, controller interfaces, safety requirements and sim-to-real deployment robustness.

Based on this three-level view, this section summarizes four cross-stage bottlenecks that prevent UAV 3D scene understanding from moving toward reliable and actionable autonomous systems. First, current data resources remain insufficient for learning and validating online and predictive agent capabilities, especially because closed-loop flight records and action-conditioned supervision are still scarce. Second, online and predictive methods require trustworthy scene-state memory to maintain geometric, semantic and temporal consistency over long horizons. Third, collaborative 3D understanding requires distributed scene-state fusion across multiple UAVs and heterogeneous air-ground platforms. Finally, sim-to-real transfer and physical deployment reliability determine whether these capabilities can remain valid under real sensor, timing and onboard resource constraints. These bottlenecks should be assessed not only by perception accuracy, but also by stage-specific and system-level criteria that reflect reliability, safety and deployability. As summarized in Table 9, these challenges provide a structured basis for identifying future directions, while the following synthesis clarifies how they jointly constrain transitions across capability stages.

Table 9. Summary of key challenges, future directions and system-level assessment priorities for UAV 3D scene understanding methods.

Key Challenges	Future Directions	Assessment Focus
Closed-Loop Data Scarcity	Flight-Centered Records, Action-State Supervision, Observed/Inferred/Unknown-Region Labels, Failure-Recovery Cases	Task Success Rate, Safety Violation Rate, Intervention Frequency, Recovery Success Rate
Trustworthy and Interpretable Scene-State Memory	Temporal State Maintenance, Structural Provenance, Uncertainty Calibration, Online Adaptation, Traceable Foundation-Model Outputs	Distribution-Shift Robustness, State Traceability, Confidence Calibration, Explanation Consistency, Long-Term Consistency
Collaborative Scene-State Fusion	Distributed Scene-State Alignment, Cross-Platform Registration, Selective Information Exchange, Bandwidth-Adaptive Representation	Fusion Accuracy, Pose Consistency, Communication Efficiency, Scalability Across Agents
Real-World Deployment Robustness	Geometry-Aware Transfer, Sensor Calibration, Temporal Synchronization, Controller Interfaces, Resource-Aware Onboard Execution	Transfer Robustness, Metric Consistency, Runtime Latency, Energy Consumption

7.1. Cross-Stage Synthesis of Capability Transitions

The four bottlenecks in Table 9 are not isolated issues, but recurring constraints that determine whether UAV 3D scene understanding can progress across capability stages. Moving from offline interpretation to online understanding requires more than applying static perception models to streaming inputs. It depends on real-time processing, pose-

consistent scene-state updating, latency-aware multimodal fusion, temporal memory and uncertainty exposure. Without these capabilities, geometrically accurate offline reconstructions may remain insufficient for in-flight navigation, interaction or safety-critical decision support. Moving from online understanding to predictive reasoning further exposes the coupling among closed-loop data scarcity, trustworthy scene-state memory, collaborative fusion and sim-to-real robustness. Closed-loop data provide the empirical basis for learning and validating how observations, actions and scene states evolve during flight. When such data are insufficient, memory calibration, multi-agent interaction modeling and deployment-failure diagnosis all become underconstrained. As a result, errors in one bottleneck may be propagated to others rather than being isolated within a single module.

Trustworthy memory serves as the state carrier through which these bottlenecks interact. If the maintained scene state cannot preserve temporal validity, uncertainty and structural provenance, prediction may amplify erroneous inferred states, collaboration may fuse inconsistent scene representations, and transfer methods may misinterpret sensing or environmental shifts as reliable scene evidence. Collaborative fusion can mitigate the limited field of view of a single UAV, but it also introduces pose uncertainty, communication delay and representation-alignment errors that must be reflected in closed-loop data and memory design. Sim-to-real validation closes this coupling loop: real-flight failures reveal missing data modes and memory-calibration errors, whereas successful deployment logs provide supervision signals for improving state maintenance, collaborative alignment and transfer robustness. These relations indicate that the four bottlenecks should be addressed as a coupled system rather than as independent challenges. Progress therefore requires the co-design of data protocols, memory structures, collaborative fusion mechanisms and deployment validation, so that predictive UAV 3D scene understanding can remain reliable under real sensing, timing and onboard resource constraints.

To make these directions more actionable, future research should translate the above bottlenecks into stage-specific evaluation records, representation mechanisms and deployment protocols. For trustworthy evaluation, UAV benchmarks should report not only final perception accuracy, but also state-update traces, observed, inferred and unknown regions, uncertainty calibration, failure-recovery cases and safety-related violations. For collaborative understanding, multi-UAV and air-ground systems should exchange compact and uncertainty-aware scene-state abstractions rather than raw observations when bandwidth, latency or viewpoint overlap is limited. For deployment robustness, simulation-to-real validation should explicitly account for UAV-specific effects such as 6DoF motion, rolling shutter, motion blur, vibration, wind disturbance, calibration drift, temporal synchronization and onboard computation limits.

7.2. Closed-Loop Data Scarcity

UAV 3D scene understanding remains strongly constrained by the limited availability of closed-loop and action-conditioned data. Compared with offline reconstruction datasets, deployable UAV systems require time-synchronized flight records that include sensory observations, platform poses, candidate actions, task states, failure cases, occlusion changes and calibration parameters. However, existing datasets still predominantly target offline detection, segmentation or mapping. Data resources that support online scene-state updating, action-conditioned prediction and collaborative perception remain limited. Future benchmarks should therefore move from static scene annotation toward flight-centered data collection that captures how a UAV perceives, acts, fails and recovers during active missions.

Data protocols should also be aligned with the capability stage under study. Offline interpretation mainly requires geometric accuracy and semantic precision. Online un-

derstanding additionally requires temporally ordered observations, pose traces, runtime constraints and records of state updates during flight. Predictive reasoning further requires action-state pairs, future observations, inferred unknown regions and failure-recovery cases. A model with high accuracy on generic 3D benchmarks may still fail in deployment if its training data do not reflect onboard resource limits, temporal causality or the uncertainty that must be exposed to planning and control modules. From a theoretical perspective, these data should support the learning and evaluation of scene-state transitions rather than only isolated perception outputs, including the distinction among directly observed evidence, inferred structures and explicitly unknown regions.

Future data construction should explicitly distinguish observed, inferred and unknown regions. Predictive systems need supervision and validation signals for geometric consistency, uncertainty calibration, action feasibility and downstream task benefit. High-level generative policies should remain bounded by geometric verification, dynamics constraints and safety filters [128,173]. For autonomous UAVs, a visually plausible but metrically inconsistent future prediction may be more hazardous than an explicitly unknown region, because it can mislead the planner into treating an unverified hypothesis as an actionable scene state.

7.3. Trustworthy and Interpretable Scene-State Memory

Current UAV action-conditioned world models are often trained offline and deployed with limited adaptation capability, which makes them vulnerable to environmental distribution shifts. Future research should prioritize lightweight online adaptation during flight to account for local wind disturbance, sensor degradation and task-specific environmental variation [143,174,175]. A robust world model should not remain a fixed offline predictor, but should support bounded correction of predictive bias during deployment. This adaptation should operate not only within network parameters, but also within the explicitly maintained scene-state representation.

UAV 3D scene understanding is intrinsically temporal. The offline stage uses long observation windows for global optimization, the online stage relies on continuous state updates, and the predictive stage performs conditional reasoning over future horizons. A trustworthy scene state should therefore record current system beliefs together with structural provenance, so that the system can distinguish direct observations, inferred structures and unknown spaces. For dynamic aerial environments, temporal state memory is essential for maintaining map continuity, handling temporary occlusions, revisiting structural elements and resolving cross-time natural language references. When open-vocabulary models, VLMs or LLM-based interfaces are used to update or query this memory, provenance should further indicate whether a semantic label, spatial relation or language-grounded hypothesis is supported by visual evidence, geometric observations, temporal records or only by model inference. The core bottleneck is therefore not only memory capacity, but also whether the maintained state can preserve provenance, uncertainty and temporal validity when observations are incomplete, delayed or semantically ambiguous.

Uncertainty should be treated as a core component of agent-level scene understanding rather than auxiliary metadata. Errors from depth extrapolation, relation inference, occupancy completion and world-model rollout should be exposed to planners and safety layers. This issue becomes more critical in foundation-model-assisted settings, where prompt-conditioned mapping, VLA coupling and world-model prediction may produce ambiguous grounding, biased semantic associations or unsupported generative outputs. Language interfaces make uncertainty handling more important, because natural language instructions express high-level goals that depend on current viewpoints, historical observations and spatial constraints. A language-conditioned UAV system should therefore integrate intent

parsing, 3D spatial grounding, ambiguity detection, confidence calibration and, when necessary, human clarification before transferring candidate goals to geometric perception and control modules. In this way, trustworthy scene-state memory should provide not only long-term geometric and semantic consistency, but also traceable, calibrated and interpretable state estimates for downstream planning and safety-aware decision support.

7.4. Collaborative Scene-State Fusion

The limitations of a single UAV in field of view, payload capacity and flight endurance motivate collaborative frameworks for wide-area exploration and complex structural reasoning. Collaborative 3D scene understanding goes beyond basic sensor fusion by requiring the alignment of distributed scene states. The system must jointly address cross-view geometric fusion, communication bandwidth constraints, distributed pose uncertainty and semantic consistency. While benchmarks such as MCOP have introduced multi-UAV collaborative occupancy prediction to extend local observations into unified volumetric forecasting [57], scalable and geometrically accurate collaborative understanding in unconstrained environments remains an open challenge.

Multi-UAV systems may still provide limited lateral and ground-level observations, especially for lower facades, occluded passages and interior structures. Integrating unmanned ground vehicles establishes a complementary air-ground collaborative understanding paradigm. UAVs provide global top-down geometric coverage and route priors, while ground units provide fine-grained lateral structures and interior topology. Although mechanisms derived from vehicle-to-everything architectures offer transferable paradigms [87,88,176,177], they primarily target object-level bounding box perception rather than dense 3D scene understanding. Advanced collaborative understanding requires spatial confidence estimation [178], enabling nodes to transmit structural summaries or latent states selectively rather than redundant raw point clouds.

A central challenge in collaborative scene-state fusion is determining the appropriate representation for shared scene information. Point clouds, volumetric maps, neural fields, scene graphs and latent world states encode different levels of geometric and semantic abstraction. Some representations are effective for metric free-space delineation, whereas others better support symbolic language interaction and long-horizon cooperative planning. In heterogeneous air-ground systems, representation selection should account for asymmetric bandwidth, processing latency and task benefit. Consequently, collaborative 3D scene understanding should evolve from centralized data aggregation toward distributed and task-oriented scene-state negotiation.

7.5. Real-World Deployment Robustness

Simulation is an important tool for training world models and testing high-risk policies, yet UAVs face a pronounced sim-to-real gap. Illumination variation, motion blur, rolling shutter effects, propeller vibration, wind disturbance, LiDAR noise and GPS multipath can substantially alter how scene states are generated and consumed [179,180]. Recent efforts such as RA3T show that sim-to-real transfer requires geometric structure alignment and cross-modal consistency rather than only visual style translation [181]. Future methodologies should jointly model occlusion, spatial scale, elevation and sensor timing while incorporating geometric consistency and dynamic feasibility into transfer objectives.

A primary engineering bottleneck across all capability stages is coordinate and temporal consistency. UAV 3D scene understanding involves camera intrinsics, camera-IMU extrinsics, LiDAR-camera extrinsics, gimbal poses, hardware timestamps and real-time kinematic priors. Camera calibration, multi-sensor spatiotemporal calibration and inertial preintegration form the geometric basis for keeping these heterogeneous observations in

a common coordinate frame [182–184]. Offline photogrammetry can partially absorb calibration errors through global bundle adjustment, whereas online pipelines and predictive models have limited access to post-hoc correction. Visual-inertial and LiDAR-inertial systems therefore remain essential for maintaining physically meaningful coordinate frames during active flight [35,114,185]. A minor timestamp offset during aggressive yaw motion or a marginal extrinsic error may cause spatial misalignment, semantic misprojection and safety-critical free-space misjudgment, and such errors may further propagate into predicted future trajectories. Therefore, scene states used in active task loops should remain traceable to calibrated sensors, verified timestamps, deterministic state estimators and explicit confidence representations.

In addition to geometric consistency, deployment-oriented studies should report runtime latency, memory footprint and energy consumption under realistic payload and power constraints. Advanced perception and prediction modules become actionable only when they satisfy the size, weight and power limits of UAV platforms. Future systems should therefore couple algorithmic accuracy with resource-aware assessment, so that improvements in 3D understanding can be translated into reliable onboard execution.

8. Conclusions

This review has organized UAV 3D scene understanding through an agent-capability framework supported by data foundations and developed across offline interpretation, online understanding and predictive reasoning. Data foundations determine which scene states can be observed, supervised and evaluated. Offline interpretation methods transform post-acquisition UAV observations into semantically enriched 3D scenes through point cloud semantic learning, object-level interpretation and prompt-conditioned semanticization. Online methods then move these scene states into active flight by maintaining pose-consistent geometric memory, streaming semantic states, structured object relations and language-conditioned control interfaces. Predictive methods further extend UAV 3D scene understanding beyond the currently visible scene by estimating hidden occupancy, selecting informative future viewpoints and evaluating action-conditioned future states.

Beyond predictive reasoning, autonomous UAV systems still require task-level planning, decision-making under uncertainty, controller integration, safety-layer verification, human–agent collaboration, multi-agent coordination and mission execution. Predictive 3D scene understanding contributes to this broader autonomy stack by supplying uncertainty-aware scene states and action-consequence cues, while final action selection and execution remain governed by task objectives, vehicle dynamics, communication conditions and safety constraints.

This capability trajectory shows that UAV 3D scene understanding is no longer an isolated perception task, but a system-level scene-understanding capability for UAV agents. The central challenge is to maintain scene states that are geometrically consistent, semantically meaningful, temporally traceable, uncertainty-aware and executable under onboard constraints. Future research should therefore focus on closed-loop and stage-aware benchmarks, trustworthy scene-state memory, collaborative air–ground 3D understanding, sim-to-real transfer and resource-aware deployment. As edge computing, multimodal foundation models and distributed collaborative perception continue to develop, robust 3D scene understanding is likely to become a key enabling capability for UAVs operating safely and effectively in open physical environments.

Author Contributions: Conceptualization, E.Z., X.L. and Z.C.; methodology, E.Z., L.X., Z.C. and J.C.; investigation, E.Z., L.X., J.C., J.W., Y.Z. and K.Y.; resources, X.L., X.Q. and L.W.; data curation, E.Z., J.C., J.W. and Y.Z.; writing—original draft preparation, E.Z., L.X. and Z.C.; writing—review and editing, E.Z., L.X., J.W., Y.Z., K.Y. and X.L.; visualization, E.Z., J.C. and K.Y.; supervision, X.L. and X.Q.; project administration, X.L.; funding acquisition, X.L. and L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Disruptive Technology Program, grant number 2025-AIRCAS-SDTP-01 and 2024-AIRCAS-SDTP-03, and the Key Program of the Chinese Academy of Sciences, grant numbers KGFZD-145-25-38.

Data Availability Statement: No new data were created or analyzed in this review.

Acknowledgments: During the preparation of this manuscript, the authors used AI-assisted tools for visualization, bibliographic organization and language polishing. The authors reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Maes, W.H.; Steppe, K. Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture. *Trends Plant Sci.* **2019**, *24*, 152–164. [[CrossRef](#)] [[PubMed](#)]
2. Nex, F.; Remondino, F. UAV for 3D Mapping Applications: A Review. *Appl. Geomat.* **2014**, *6*, 1–15. [[CrossRef](#)]
3. Butilă, E.V.; Boboc, R.G. Urban traffic monitoring and analysis using unmanned aerial vehicles (UAVs): A systematic literature review. *Remote Sens.* **2022**, *14*, 620. [[CrossRef](#)]
4. Ye, Y.; Wang, X.; Gou, G.; Zhang, H.; Li, H.; Sui, H. Autonomous Exploration-Oriented UAV Approach for Real-Time Spatial Mapping in Unknown Environments. *Drones* **2025**, *9*, 844. [[CrossRef](#)]
5. Remondino, F. Heritage Recording and 3D Modeling with Photogrammetry and 3D Scanning. *Remote Sens.* **2011**, *3*, 1104–1138. [[CrossRef](#)]
6. Xu, Z.; Wu, L.; Gerke, M.; Wang, R.; Yang, H. Skeletal camera network embedded structure-from-motion for 3D scene reconstruction from UAV images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *121*, 113–127. [[CrossRef](#)]
7. Jiang, C.; Shao, H. Fast 3d reconstruction of uav images based on neural radiance field. *Appl. Sci.* **2023**, *13*, 10174. [[CrossRef](#)]
8. Yi, S.; Liu, X.; Li, J.; Chen, L. UAVformer: A composite transformer network for urban scene segmentation of UAV images. *Pattern Recognit.* **2023**, *133*, 109019. [[CrossRef](#)]
9. Qian, J.; Yan, Y.; Gao, F.; Ge, B.; Wei, M.; Shangguan, B.; He, G. C3DGS: Compressing 3D Gaussian model for surface reconstruction of large-scale scenes based on multiview UAV images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 4396–4409. [[CrossRef](#)]
10. Kumar, V.; Michael, N. Opportunities and Challenges with Autonomous Micro Aerial Vehicles. *Int. J. Robot. Res.* **2012**, *31*, 1279–1291. [[CrossRef](#)]
11. Carrio, A.; Sampedro, C.; Rodriguez-Ramos, A.; Campoy, P. A review of deep learning methods and applications for unmanned aerial vehicles. *J. Sens.* **2017**, *2017*, 3296874. [[CrossRef](#)]
12. Samaras, S.; Diamantidou, E.; Ataloglou, D.; Sakellariou, N.; Vafeiadis, A.; Magoulanis, V.; Lalas, A.; Dimou, A.; Zarpalas, D.; Votis, K.; et al. Deep learning on multi sensor data for counter UAV applications—A systematic review. *Sensors* **2019**, *19*, 4837. [[CrossRef](#)] [[PubMed](#)]
13. Rahman, M.H.; Sejan, M.A.S.; Aziz, M.A.; Tabassum, R.; Baik, J.I.; Song, H.K. A comprehensive survey of unmanned aerial vehicles detection and classification using machine learning approach: Challenges, solutions, and future directions. *Remote Sens.* **2024**, *16*, 879. [[CrossRef](#)]
14. Shakhathreh, H.; Sawalmeh, A.H.; Al-Fuqaha, A.; Dou, Z.; Almaita, E.; Khalil, I.; Othman, N.S.; Khreishah, A.; Guizani, M. Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges. *IEEE Access* **2019**, *7*, 48572–48634. [[CrossRef](#)]
15. Osco, L.P.; Junior, J.M.; Ramos, A.P.M.; de Castro Jorge, L.A.; Fatholahi, S.N.; Silva, J.d.A.; Matsubara, E.T.; Pistori, H.; Gonçalves, W.N.; Li, J. A review on deep learning in UAV remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102456. [[CrossRef](#)]
16. Chen, Z.; Zhu, E.; Guo, Z.; Zhang, P.; Liu, X.; Wang, L.; Zhang, Y. Predictive autonomy for UAV remote sensing: A survey of video prediction. *Remote Sens.* **2025**, *17*, 3423. [[CrossRef](#)]
17. Mao, J.; Shi, S.; Wang, X.; Li, H. 3D object detection for autonomous driving: A comprehensive survey. *Int. J. Comput. Vis.* **2023**, *131*, 1909–1963. [[CrossRef](#)]

18. Pravalika, A.; Hashmi, M.F.; Gupta, A. Deep learning frontiers in 3D object detection: A comprehensive review for autonomous driving. *IEEE Access* **2024**, *12*, 173936–173980. [\[CrossRef\]](#)
19. Xu, H.; Chen, J.; Meng, S.; Wang, Y.; Chau, L.P. A survey on occupancy perception for autonomous driving: The information fusion perspective. *Inf. Fusion* **2025**, *114*, 102671. [\[CrossRef\]](#)
20. Roldao, L.; De Charette, R.; Verroust-Blondet, A. 3D semantic scene completion: A survey. *Int. J. Comput. Vis.* **2022**, *130*, 1978–2005. [\[CrossRef\]](#)
21. Benallal, H.; Abdallah Saab, N.; Tairi, H.; Alfalou, A.; Riffi, J. Advancements in Semantic Segmentation of 3D Point Clouds for Scene Understanding Using Deep Learning. *Technologies* **2025**, *13*, 322. [\[CrossRef\]](#)
22. Nguyen, T.A.Q.; Bourki, A.; Macudzinski, M.; Brunel, A.; Bennamoun, M. Semantically-aware neural radiance fields for visual scene understanding: A comprehensive review. *Int. J. Comput. Vis.* **2026**, *134*, 109. [\[CrossRef\]](#)
23. Krishnan, S.; Wan, Z.; Bhardwaj, K.; Whatmough, P.; Faust, A.; Neuman, S.; Wei, G.Y.; Brooks, D.; Reddi, V.J. Automatic domain-specific soc design for autonomous unmanned aerial vehicles. In Proceedings of the 55th IEEE/ACM International Symposium on Microarchitecture, Chicago, IL, USA, 1–5 October 2022; pp. 300–317.
24. Boroujerdian, B.; Genc, H.; Krishnan, S.; Duisterhof, B.P.; Plancher, B.; Mansoorshahi, K.; Almeida, M.; Cui, W.; Faust, A.; Reddi, V.J. The role of compute in autonomous micro aerial vehicles: Optimizing for mission time and energy efficiency. *ACM Trans. Comput. Syst. (TOCS)* **2022**, *39*, 1–44.
25. Agüera-Vega, F.; Ferrer-González, E.; Martínez-Carricondo, P.; Sánchez-Hermosilla, J.; Carvajal-Ramírez, F. Influence of the Inclusion of Off-Nadir Images on UAV-Photogrammetry Projects from Nadir Images and AGL or AMSL Flights. *Drones* **2024**, *8*, 662. [\[CrossRef\]](#)
26. Kurniawati, H. Partially Observable Markov Decision Processes and Robotics. *Annu. Rev. Control Robot. Auton. Syst.* **2022**, *5*, 253–277. [\[CrossRef\]](#)
27. Snavely, N.; Seitz, S.M.; Szeliski, R. Photo Tourism: Exploring Photo Collections in 3D. In Proceedings of the ACM SIGGRAPH, Boston, MA, USA, 30 July–3 August 2006; pp. 835–846. [\[CrossRef\]](#)
28. Furukawa, Y.; Ponce, J. Accurate, Dense, and Robust Multi-View Stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1362–1376. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Aloimonos, J.; Weiss, I.; Bandyopadhyay, A. Active vision. *Int. J. Comput. Vis.* **1988**, *1*, 333–356. [\[CrossRef\]](#)
30. Song, S.; Yu, F.; Zeng, A.; Chang, A.X.; Savva, M.; Funkhouser, T. Semantic Scene Completion from a Single Depth Image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1746–1754. [\[CrossRef\]](#)
31. Ha, D.; Schmidhuber, J. World models. *arXiv* **2018**, arXiv:1803.10122. [\[CrossRef\]](#)
32. Colomina, I.; Molina, P. Unmanned Aerial Systems for Photogrammetry and Remote Sensing: A Review. *ISPRS J. Photogramm. Remote Sens.* **2014**, *92*, 79–97. [\[CrossRef\]](#)
33. Schonberger, J.L.; Frahm, J.M. Structure-from-Motion Revisited. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4104–4113. [\[CrossRef\]](#)
34. Campos, C.; Elvira, R.; Rodriguez, J.J.G.; Montiel, J.M.M.; Tardos, J.D. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial, and Multimap SLAM. *IEEE Trans. Robot.* **2021**, *37*, 1874–1890. [\[CrossRef\]](#)
35. Qin, T.; Li, P.; Shen, S. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. *IEEE Trans. Robot.* **2018**, *34*, 1004–1020. [\[CrossRef\]](#)
36. Kerbl, B.; Kopanas, G.; Leimkuehler, T.; Drettakis, G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* **2023**, *42*, 1–14. [\[CrossRef\]](#)
37. Qin, M.; Li, W.; Zhou, J.; Wang, H.; Pfister, H. Langsplat: 3d language gaussian splatting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 20051–20060. [\[CrossRef\]](#)
38. Sun, X.; Si, W.; Ni, W.; Li, Y.; Wu, D.; Xie, F.; Guan, R.; Xu, H.Y.; Ding, H.; Wu, Y.; et al. AutoFly: Vision-Language-Action Model for UAV Autonomous Navigation in the Wild. *arXiv* **2026**, arXiv:2602.09657. [\[CrossRef\]](#)
39. Yao, J.; Zhang, J.; Pan, X.; Wu, T.; Xiao, C. DepthSSC: Monocular 3D semantic scene completion via depth-spatial alignment and voxel adaptation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Tucson, AZ, USA, 28 February–4 March 2025; pp. 2154–2163. [\[CrossRef\]](#)
40. Zhang, W.; Tang, P.; Zeng, X.; Man, F.; Yu, S.; Dai, Z.; Zhao, B.; Chen, H.; Shang, Y.; Wu, W.; et al. Aerial World Model for Long-Horizon Visual Generation and Navigation in 3D Space. *arXiv* **2025**, arXiv:2512.21887. [\[CrossRef\]](#)
41. Elmokadem, T.; Savkin, A.V. Towards Fully Autonomous UAVs: A Survey. *Sensors* **2021**, *21*, 6223. [\[CrossRef\]](#) [\[PubMed\]](#)
42. Tian, Y.; Yue, H.; Yang, B.; Ren, J. Unmanned Aerial Vehicle Visual Simultaneous Localization and Mapping: A Survey. *J. Phys. Conf. Ser.* **2022**, *2278*, 012006. [\[CrossRef\]](#)
43. Jiang, S.; Jiang, W.; Wang, L. Unmanned Aerial Vehicle-Based Photogrammetric 3D Mapping: A Survey of Techniques, Applications, and Challenges. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 135–171. [\[CrossRef\]](#)

44. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 370–386. [\[CrossRef\]](#)
45. Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q. Vision meets drones: A challenge. *arXiv* **2018**, arXiv:1804.07437. [\[CrossRef\]](#)
46. Lyu, Y.; Vosselman, G.; Xia, G.S.; Yilmaz, A.; Yang, M.Y. UAVid: A Semantic Segmentation Dataset for UAV Imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *165*, 108–119. [\[CrossRef\]](#)
47. Varney, N.; Asari, V.K.; Graehling, Q. DALES: A large-scale aerial LiDAR data set for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Virtual, 14–19 June 2020; pp. 186–187. [\[CrossRef\]](#)
48. Kölle, M.; Laupheimer, D.; Schmohl, S.; Haala, N.; Rottensteiner, F.; Wegner, J.D.; Ledoux, H. The Hessigheim 3D (H3D) benchmark on semantic segmentation of high-resolution 3D point clouds and textured meshes from UAV LiDAR and Multi-View-Stereo. *ISPRS Open J. Photogramm. Remote Sens.* **2021**, *1*, 100001. [\[CrossRef\]](#)
49. Hu, Q.; Yang, B.; Khalid, S.; Xiao, W.; Trigoni, N.; Markham, A. Sensaturban: Learning semantics from urban-scale photogrammetric point clouds. *Int. J. Comput. Vis.* **2022**, *130*, 316–343. [\[CrossRef\]](#)
50. Wang, S.; Li, S.; Zhang, Y.; Yu, S.; Yuan, S.; She, R.; Guo, Q.; Zheng, J.; Howe, O.K.; Chandra, L.; et al. Uavscenes: A multi-modal dataset for uavs. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 28946–28958. [\[CrossRef\]](#)
51. Jung, M.; Yang, W.; Lee, D.; Gil, H.; Kim, G.; Kim, A. HeLiPR: Heterogeneous LiDAR dataset for inter-LiDAR place recognition under spatiotemporal variations. *Int. J. Robot. Res.* **2024**, *43*, 1867–1883. [\[CrossRef\]](#)
52. Tang, J.; Gao, Y.; Yang, D.; Yan, L.; Yue, Y.; Yang, Y. DroneSplat: 3D Gaussian Splatting for Robust 3D Reconstruction from In-the-Wild Drone Imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2025; pp. 833–843. [\[CrossRef\]](#)
53. Vuong, K.; Ghosh, A.; Ramanan, D.; Narasimhan, S.; Tulsiani, S. Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2025; pp. 21674–21684. [\[CrossRef\]](#)
54. Jiang, L.; Ren, K.; Yu, M.; Xu, L.; Dong, J.; Lu, T.; Zhao, F.; Lin, D.; Dai, B. Horizon-GS: Unified 3D Gaussian Splatting for Large-Scale Aerial-to-Ground Scenes. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2025; pp. 26789–26799. [\[CrossRef\]](#)
55. Mitra, S.; Rawat, Y.S. ProDiG: Progressive Diffusion-Guided Gaussian Splatting for Aerial to Ground Reconstruction. *arXiv* **2026**, arXiv:2604.02003. [\[CrossRef\]](#)
56. Ye, H.; Sunderraman, R.; Ji, S. Uav3d: A large-scale 3d perception benchmark for unmanned aerial vehicles. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 55425–55442. [\[CrossRef\]](#)
57. Lin, Z.; Chen, W.; Jin, X.; Yang, Y.; Fan, L.; Zhang, Y.; Zhang, Y.; Zhang, Z. MCOP: Multi-UAV Collaborative Occupancy Prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 27242–27251. [\[CrossRef\]](#)
58. Beche, R.; Nedeveschi, S. ClaraVid: A Holistic Scene Reconstruction Benchmark From Aerial Perspective with Delentropy-Based Complexity Profiling. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 26015–26025. [\[CrossRef\]](#)
59. Gross, M.; Matha, S.B.; Fahmy, A.; Song, R.; Cremers, D.; Meess, H. OccuFly: A 3D Vision Benchmark for Semantic Scene Completion from the Aerial Perspective. *arXiv* **2025**, arXiv:2512.20770. [\[CrossRef\]](#)
60. Ma, W.; Li, Z.; Zhu, J.; Hua, T.; Chen, K.; Cao, Z.; Yang, D.; Shi, P.; Zhou, Y.; Zhao, E.; et al. SkyEvents: A Large-Scale Event-Enhanced UAV Dataset for Robust 3D Scene Reconstruction. In Proceedings of the International Conference on Learning Representations, Rio de Janeiro, Brazil, 23–27 April 2026.
61. Zhao, B.; Tang, R.; Jia, M.; Wang, Z.; Man, F.; Zhang, X.; Shang, Y.; Zhang, W.; Wu, W.; Gao, C.; et al. AirScape: An Aerial Generative World Model with Motion Controllability. In Proceedings of the the 33rd ACM International Conference on Multimedia, Dublin, Ireland, 27–31 October 2025; pp. 12519–12528. [\[CrossRef\]](#)
62. Guo, Z.; Chen, Z.; Zhu, E.; Wei, K.; Zou, Y.; Liu, X.; Wang, L. MotionScape: A Large-Scale Real-World Highly Dynamic UAV Video Dataset for World Models. *arXiv* **2026**, arXiv:2604.07991. [\[CrossRef\]](#)
63. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983. [\[CrossRef\]](#)
64. Ding, J.; Xue, N.; Xia, G.S.; Bai, X.; Yang, W.; Yang, M.Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7778–7796. [\[CrossRef\]](#) [\[PubMed\]](#)

65. Rottensteiner, F.; Sohn, G.; Jung, J.; Gerke, M.; Baillard, C.; Benitez, S.; Breitkopf, U. The ISPRS Benchmark on Urban Object Classification and 3D Building Reconstruction. In Proceedings of the ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Melbourne, Australia, 25 August–1 September 2012; pp. 293–298. [\[CrossRef\]](#)
66. Gerke, M. Use of the Stair Vision Library within the ISPRS 2D Semantic Labeling Benchmark. In Proceedings of the ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Zurich, Switzerland, 5–7 September 2014.
67. Schönberger, J.L.; Zheng, E.; Frahm, J.M.; Pollefeys, M. Pixelwise view selection for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 501–518. [\[CrossRef\]](#)
68. DeTone, D.; Malisiewicz, T.; Rabinovich, A. Superpoint: Self-supervised interest point detection and description. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 224–236. [\[CrossRef\]](#)
69. Sarlin, P.E.; DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperGlue: Learning Feature Matching with Graph Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 4938–4947. [\[CrossRef\]](#)
70. Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; Zhou, X. LoFTR: Detector-Free Local Feature Matching with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 8922–8931. [\[CrossRef\]](#)
71. Lindenberger, P.; Sarlin, P.E.; Pollefeys, M. Lightglue: Local feature matching at light speed. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 17627–17638. [\[CrossRef\]](#)
72. Yao, Y.; Luo, Z.; Li, S.; Fang, T.; Quan, L. MVSNet: Depth Inference for Unstructured Multi-view Stereo. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 767–783. [\[CrossRef\]](#)
73. Yao, Y.; Luo, Z.; Li, S.; Shen, T.; Fang, T.; Quan, L. Recurrent MVSNet for High-Resolution Multi-View Stereo Depth Inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5525–5534. [\[CrossRef\]](#)
74. Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; Tan, P. Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 2495–2504. [\[CrossRef\]](#)
75. Wang, F.; Galliani, S.; Vogel, C.; Speciale, P.; Pollefeys, M. PatchmatchNet: Learned Multi-View Patchmatch Stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 14194–14203. [\[CrossRef\]](#)
76. Godard, C.; Aodha, O.M.; Firman, M.; Brostow, G.J. Digging into Self-Supervised Monocular Depth Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 27 October–2 November 2019; pp. 3828–3838. [\[CrossRef\]](#)
77. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision Transformers for Dense Prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 12179–12188. [\[CrossRef\]](#)
78. Bochkovskiy, A.; Delaunoy, A.; Germain, H.; Santos, M.; Zhou, Y.; Richter, S.; Koltun, V. Depth pro: Sharp monocular metric depth in less than a second. *Proc. Int. Conf. Learn. Represent.* **2025**, 2025, 75602–75637. [\[CrossRef\]](#)
79. Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; Zhao, H. Depth anything v2. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 21875–21911. [\[CrossRef\]](#)
80. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 405–421. [\[CrossRef\]](#)
81. Barron, J.T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; Srinivasan, P.P. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 5855–5864. [\[CrossRef\]](#)
82. Müller, T.; Evans, A.; Schied, C.; Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.* **2022**, *41*, 1–15. [\[CrossRef\]](#)
83. Chen, A.; Xu, Z.; Geiger, A.; Yu, J.; Su, H. Tensorf: Tensorial radiance fields. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 333–350. [\[CrossRef\]](#)
84. Fridovich-Keil, S.; Yu, A.; Tancik, M.; Chen, Q.; Recht, B.; Kanazawa, A. Plenoxels: Radiance fields without neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 5501–5510. [\[CrossRef\]](#)
85. Tancik, M.; Weber, E.; Ng, E.; Li, R.; Yi, B.; Wang, T.; Kristoffersen, A.; Austin, J.; Salahi, K.; Ahuja, A.; et al. Nerfstudio: A modular framework for neural radiance field development. In Proceedings of the ACM SIGGRAPH, Los Angeles, CA, USA, 6–10 August 2023; pp. 1–12. [\[CrossRef\]](#)
86. Sanchez, J.; Soum-Fontez, L.; Deschaud, J.E.; Goulette, F. Parislu3d: A high-quality target dataset for domain generalization of lidar perception. *IEEE Robot. Autom. Lett.* **2024**, *9*, 5496–5503. [\[CrossRef\]](#)

87. Xu, R.; Xiang, H.; Tu, Z.; Xia, X.; Yang, M.H.; Ma, J. OPV2V: An Open Benchmark Dataset and Fusion Pipeline for Perception with Vehicle-to-Vehicle Communication. In Proceedings of the IEEE International Conference on Robotics and Automation, Philadelphia, PA, USA, 23–27 May 2022; pp. 2583–2589. [\[CrossRef\]](#)
88. Yu, H.; Luo, Y.; Shu, M.; Huo, Y.; Yang, Z.; Shi, Y.; Guo, Z.; Li, H.; Hu, X.; Yuan, J.; et al. DAIR-V2X: A Large-Scale Dataset for Vehicle-Infrastructure Cooperative 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 21361–21370. [\[CrossRef\]](#)
89. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660. [\[CrossRef\]](#)
90. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5099–5108. [\[CrossRef\]](#)
91. Thomas, H.; Qi, C.R.; Deschaud, J.E.; Marcotegui, B.; Goulette, F.; Guibas, L.J. KPConv: Flexible and Deformable Convolution for Point Clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 27 October–2 November 2019; pp. 6411–6420. [\[CrossRef\]](#)
92. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 11108–11117. [\[CrossRef\]](#)
93. Zhao, H.; Jiang, L.; Jia, J.; Torr, P.H.; Koltun, V. Point Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 16259–16268. [\[CrossRef\]](#)
94. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. PointPillars: Fast Encoders for Object Detection from Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12697–12705. [\[CrossRef\]](#)
95. Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; Li, H. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 10529–10538. [\[CrossRef\]](#)
96. Yin, T.; Zhou, X.; Krahenbuhl, P. Center-Based 3D Object Detection and Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 11784–11793. [\[CrossRef\]](#)
97. Liang, T.; Xie, H.; Yu, K.; Xia, Z.; Lin, Z.; Wang, Y.; Tang, T.; Wang, B.; Tang, Z. Bevfusion: A simple and robust lidar-camera fusion framework. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 10421–10434. [\[CrossRef\]](#)
98. Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T. Openscene: 3d scene understanding with open vocabularies. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 815–824. [\[CrossRef\]](#)
99. Jatavallabhula, K.M.; Kuwajerwala, A.; Gu, Q.; Omama, M.; Chen, T.; Maalouf, A.; Li, S.; Iyer, G.; Saryazdi, S.; Keetha, N.; et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv* **2023**, arXiv:2302.07241. [\[CrossRef\]](#)
100. Kerr, J.; Kim, C.M.; Goldberg, K.; Kanazawa, A.; Tancik, M. Lerf: Language embedded radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 19729–19739. [\[CrossRef\]](#)
101. Wu, Y.; Meng, J.; Li, H.; Wu, C.; Shi, Y.; Cheng, X.; Zhao, C.; Feng, H.; Ding, E.; Wang, J.; et al. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 19114–19138. [\[CrossRef\]](#)
102. Piekenbrinck, J.; Schmidt, C.; Hermans, A.; Vaskevicius, N.; Linder, T.; Leibe, B. OpenSplat3D: Open-Vocabulary 3D Instance Segmentation Using Gaussian Splatting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2025; pp. 5246–5255. [\[CrossRef\]](#)
103. Li, W.; Zhao, Y.; Qin, M.; Liu, Y.; Cai, Y.; Gan, C.; Pfister, H. Langsplatv2: High-dimensional 3d language gaussian splatting with 450+ fps. *Adv. Neural Inf. Process. Syst.* **2025**, *38*, 174306–174330. [\[CrossRef\]](#)
104. Sun, W.; Li, Y.; Jiao, J. CAGS: Open-Vocabulary 3D Scene Understanding with Context-Aware Gaussian Splatting. *Image Vis. Comput.* **2026**, *165*, 105830. [\[CrossRef\]](#)
105. Niemeyer, J.; Rottensteiner, F.; Soergel, U. Contextual classification of lidar data and building object detection in urban areas. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 152–165. [\[CrossRef\]](#)
106. Shapovalov, R.; Velizhev, A.; Barinova, O. Non-associative Markov networks for 3D point cloud classification. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives, Saint-Mandé, France, 1–3 September 2010; pp. 103–108.
107. Lawin, F.J.; Danelljan, M.; Tosteberg, P.; Bhat, G.; Khan, F.S.; Felsberg, M. Deep Projective 3D Semantic Segmentation. In Proceedings of the International Conference on Computer Analysis of Images and Patterns, Ystad, Sweden, 22–24 August 2017; pp. 95–107. [\[CrossRef\]](#)

108. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 4015–4026. [\[CrossRef\]](#)
109. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763. [\[CrossRef\]](#)
110. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 19730–19742. [\[CrossRef\]](#)
111. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 38–55. [\[CrossRef\]](#)
112. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 34892–34916. [\[CrossRef\]](#)
113. Shafiullah, N.M.M.; Paxton, C.; Pinto, L.; Chintala, S.; Szlam, A. Clip-fields: Weakly supervised semantic fields for robotic memory. *arXiv* **2022**, arXiv:2210.05663. [\[CrossRef\]](#)
114. Shan, T.; Englot, B.; Meyers, D.; Wang, W.; Ratti, C.; Rus, D. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, NV, USA, 25–29 October 2020; pp. 5135–5142. [\[CrossRef\]](#)
115. Chen, T.; Shorinwa, O.; Bruno, J.; Swann, A.; Yu, J.; Zeng, W.; Nagami, K.; Dames, P.; Schwager, M. Splat-nav: Safe real-time robot navigation in gaussian splatting maps. *IEEE Trans. Robot.* **2025**, *41*, 2765–2784. [\[CrossRef\]](#)
116. Miao, B.; Wei, R.; Ge, Z.; Sun, X.; Gao, S.; Zhu, J.; Wang, R.; Tang, S.; Xiao, J.; Tang, R.; et al. Towards Physically Executable 3D Gaussian for Embodied Navigation. *arXiv* **2025**, arXiv:2510.21307. [\[CrossRef\]](#)
117. Xu, X.; Xia, C.; Wang, Z.; Zhao, L.; Duan, Y.; Zhou, J.; Lu, J. Memory-Based Adapters for Online 3D Scene Perception. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 21604–21613. [\[CrossRef\]](#)
118. Yang, Y.; Wu, X.; He, T.; Zhao, H.; Liu, X. SAM3D: Segment Anything in 3D Scenes. *arXiv* **2023**, arXiv:2306.03908. [\[CrossRef\]](#)
119. Xu, X.; Chen, H.; Zhao, L.; Wang, Z.; Zhou, J.; Lu, J. Embodiedsam: Online segment any 3d thing in real time. In Proceedings of the International Conference on Learning Representations, Singapore, 24–28 April 2025; pp. 38020–38035. [\[CrossRef\]](#)
120. Tang, Y.; Zhang, J.; Lan, Y.; Guo, Y.; Dong, D.; Zhu, C.; Xu, K. OnlineAnySeg: Online Zero-Shot 3D Segmentation by Visual Foundation Model Guided 2D Mask Merging. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2025; pp. 3676–3685. [\[CrossRef\]](#)
121. Wang, H.; Zijian, C.; Gao, J.; Zhang, Y.; Hu, W.; Wang, K.; Zhang, Z. Online Segment Any 3D Thing as Instance Tracking. *Adv. Neural Inf. Process. Syst.* **2025**, *38*, 49076–49095. [\[CrossRef\]](#)
122. Rosinol, A.; Violette, A.; Abate, M.; Hughes, N.; Chang, Y.; Shi, J.; Gupta, A.; Carlone, L. Kimera: From SLAM to Spatial Perception with 3D Dynamic Scene Graphs. *Int. J. Robot. Res.* **2021**, *40*, 1510–1546. [\[CrossRef\]](#)
123. Ravichandran, Z.; Peng, L.; Hughes, N.; Griffith, J.D.; Carlone, L. Hierarchical representations and explicit memory: Learning effective navigation policies on 3d scene graphs using graph neural networks. In Proceedings of the IEEE International Conference on Robotics and Automation, Philadelphia, PA, USA, 23–27 May 2022; pp. 9272–9279. [\[CrossRef\]](#)
124. Gao, Y.; Wang, Z.; Jing, L.; Wang, D.; Li, X.; Zhao, B. Exploring Spatial Representation to Enhance LLM Reasoning in Aerial Vision-Language Navigation. *arXiv* **2024**, arXiv:2410.08500. [\[CrossRef\]](#)
125. Liu, Y.; Yao, F.; Yue, Y.; Xu, G.; Sun, X.; Fu, K. NavAgent: Multi-scale Urban Street View Fusion For UAV Embodied Vision-and-Language Navigation. *arXiv* **2024**, arXiv:2411.08579. [\[CrossRef\]](#)
126. Xu, H.; Hu, Y.; Gao, C.; Zhu, Z.; Zhao, Y.; Yin, Q. GeoNav: Empowering MLLMs with dual-scale geospatial reasoning for language-goal aerial navigation. *Pattern Recognit.* **2026**, *177*, 113365. [\[CrossRef\]](#)
127. Chen, J.; Li, H.; Tang, Z.; Li, X.; Wu, W.; Liu, S. AerialVLA: A Vision-Language-Action Model for Aerial Navigation with Online Dialogue. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 18161–18169. [\[CrossRef\]](#)
128. Wu, Y.; Zhu, M.; Li, X.; Du, Y.; Fan, Y.; Li, W.; Han, Z.; Zhou, X.; Gao, F. VLA-AN: An Efficient and Onboard Vision-Language-Action Framework for Aerial Navigation in Complex Environments. *arXiv* **2025**, arXiv:2512.15258. [\[CrossRef\]](#)
129. Huang, R.; Yu, M.; Tsoi, M.; Ouyang, X. MMEdge: Accelerating On-Device Multimodal Inference via Pipelined Sensing and Encoding. *arXiv* **2025**, arXiv:2510.25327. [\[CrossRef\]](#)
130. Mur-Artal, R.; Montiel, J.M.M.; Tardos, J.D. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Trans. Robot.* **2015**, *31*, 1147–1163. [\[CrossRef\]](#)
131. Leutenegger, S.; Lynen, S.; Bosse, M.; Siegwart, R.; Furgale, P. Keyframe-Based Visual-Inertial Odometry Using Nonlinear Optimization. *Int. J. Robot. Res.* **2015**, *34*, 314–334. [\[CrossRef\]](#)
132. Zhang, J.; Singh, S. LOAM: Lidar Odometry and Mapping in Real-Time. *Proc. Robot. Sci. Syst.* **2014**, *2*, 1–9. [\[CrossRef\]](#)

133. Xu, W.; Zhang, F. FAST-LIO: A Fast, Robust LiDAR-Inertial Odometry Package by Tightly-Coupled Iterated Kalman Filter. *IEEE Robot. Autom. Lett.* **2021**, *6*, 3317–3324. [[CrossRef](#)]
134. Hornung, A.; Wurm, K.M.; Bennewitz, M.; Stachniss, C.; Burgard, W. OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees. *Auton. Robot.* **2013**, *34*, 189–206. [[CrossRef](#)]
135. Oleynikova, H.; Taylor, Z.; Fehr, M.; Siegwart, R.; Nieto, J. Voxblox: Incremental 3D Euclidean Signed Distance Fields for On-Board MAV Planning. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Vancouver, BC, Canada, 24–28 September 2017; pp. 1366–1373. [[CrossRef](#)]
136. Han, L.; Gao, F.; Zhou, B.; Shen, S. FIESTA: Fast Incremental Euclidean Distance Fields for Online Motion Planning of Aerial Robots. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Macau, China, 3–8 November 2019; pp. 4423–4430. [[CrossRef](#)]
137. Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Yu, Q.; Dai, J. BEVFormer: Learning Bird’s-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 1–18. [[CrossRef](#)]
138. Keetha, N.; Karhade, J.; Jatavallabhula, K.M.; Yang, G.; Scherer, S.; Ramanan, D.; Luiten, J. SplatTAM: Splat, Track & Map 3D Gaussians for Dense RGB-D SLAM. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 21357–21366. [[CrossRef](#)]
139. Han, S.; Pool, J.; Tran, J.; Dally, W. Learning both weights and connections for efficient neural network. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1135–1143. [[CrossRef](#)]
140. Frankle, J.; Carbin, M. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019; pp. 8954–8995. [[CrossRef](#)]
141. Jacob, B.; Kligys, S.; Chen, B.; Zhu, M.; Tang, M.; Howard, A.; Adam, H.; Kalenichenko, D. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2704–2713. [[CrossRef](#)]
142. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531. [[CrossRef](#)]
143. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022; pp. 12513–12525. [[CrossRef](#)]
144. Shah, S.; Dey, D.; Lovett, C.; Kapoor, A. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In Proceedings of the Field and Service Robotics: Results of the 11th International Conference, Zurich, Switzerland, 12–15 September 2017; pp. 621–635. [[CrossRef](#)]
145. Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; Koltun, V. CARLA: An open urban driving simulator. In Proceedings of the Conference on Robot Learning, Mountain View, CA, USA, 13–15 November 2017; pp. 1–16. [[CrossRef](#)]
146. Savva, M.; Kadian, A.; Maksymets, O.; Zhao, Y.; Wijmans, E.; Jain, B.; Straub, J.; Liu, J.; Koltun, V.; Malik, J.; et al. Habitat: A Platform for Embodied AI Research. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 27 October–2 November 2019; pp. 9339–9347. [[CrossRef](#)]
147. Xia, F.; Shen, W.B.; Li, C.; Kasimbeg, P.; Tchapmi, M.E.; Toshev, A.; Martin-Martin, R.; Savarese, S. Interactive Gibson Benchmark: A Benchmark for Interactive Navigation in Cluttered Environments. *IEEE Robot. Autom. Lett.* **2020**, *5*, 713–720. [[CrossRef](#)]
148. Li, C.; Xia, F.; Martín-Martín, R.; Lingelbach, M.; Srivastava, S.; Shen, B.; Vainio, K.; Gokmen, C.; Dharan, G.; Jain, T.; et al. Igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv* **2021**, arXiv:2108.03272. [[CrossRef](#)]
149. Placed, J.A.; Strader, J.; Carrillo, H.; Atanasov, N.; Indelman, V.; Carlone, L.; Castellanos, J.A. A Survey on Active Simultaneous Localization and Mapping: State of the Art and New Frontiers. *IEEE Trans. Robot.* **2023**, *39*, 1686–1705. [[CrossRef](#)]
150. Lee, H.J.; Koh, Y.J.; Kim, H.; Kim, H.; Lee, Y.; Lee, J. SOAP: Vision-Centric 3D Semantic Scene Completion with Scene-Adaptive Decoder and Occluded Region-Aware View Projection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2025; pp. 17145–17154. [[CrossRef](#)]
151. Qian, R.; Cao, H.; Deng, T.; Yuan, S.; Xie, L. SplatSSC: Decoupled Depth-Guided Gaussian Splatting for Semantic Scene Completion. In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2026; Volume 40, pp. 8520–8528. [[CrossRef](#)]
152. Yang, Z.; Peng, Y. HD²-SSC: High-Dimension High-Density Semantic Scene Completion for Autonomous Driving. In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2026; Volume 40, pp. 11829–11837. [[CrossRef](#)]
153. Xu, Z.; Lu, B.; Bao, W.; Zhu, Z.; Zhang, J.; Yan, H.; Lu, W.; Wang, J. Fly0: Decoupling Semantic Grounding from Geometric Planning for Zero-Shot Aerial Navigation. *arXiv* **2026**, arXiv:2602.15875. [[CrossRef](#)]
154. Jiang, X.; Liu, T.; Liu, L.; Liu, Z.; Liu, Y. UEVAVD: A Dataset for Developing UAV’s Eye View Active Object Detection. *IEEE Robot. Autom. Lett.* **2025**, *10*, 6272–6279. [[CrossRef](#)]
155. Luo, W.; Sun, P.; Zhong, F.; Liu, W.; Zhang, T.; Wang, Y. End-to-end active object tracking and its real-world deployment via reinforcement learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 1317–1332. [[CrossRef](#)] [[PubMed](#)]

156. Jeong, H.; Hassani, H.; Morari, M.; Lee, D.D.; Pappas, G.J. Deep reinforcement learning for active target tracking. In Proceedings of the IEEE International Conference on Robotics and Automation, Xi'an, China, 30 May–5 June 2021; pp. 1825–1831. [\[CrossRef\]](#)
157. Dharmadhikari, M.; Alexis, K. Semantics-Aware Exploration and Inspection Path Planning. In Proceedings of the IEEE International Conference on Robotics and Automation, London, UK, 29 May–2 June 2023; pp. 3360–3367. [\[CrossRef\]](#)
158. Chen, Z.; Guo, Z.; Zhu, E.; Zhang, P.; Liu, X.; Wang, L.; Zhang, Y. RAPTOR: Real-Time High-Resolution UAV Video Prediction with Efficient Video Attention. In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2026; Volume 40, pp. 3183–3190. [\[CrossRef\]](#)
159. Yang, Z.; Ge, W.; Li, Y.; Chen, J.; Li, H.; An, M.; Kang, F.; Xue, H.; Xu, B.; Yin, Y.; et al. Matrix-3D: Omnidirectional Explorable 3D World Generation. *arXiv* **2025**, arXiv:2508.08086. [\[CrossRef\]](#)
160. Li, X.; Lai, Z.; Xu, L.; Qu, Y.; Cao, L.; Zhang, S.; Dai, B.; Ji, R. Director3d: Real-world camera trajectory and 3d scene generation from text. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 75125–75151. [\[CrossRef\]](#)
161. Bruce, J.; Dennis, M.D.; Edwards, A.; Parker-Holder, J.; Shi, Y.; Hughes, E.; Lai, M.; Mavalankar, A.; Steigerwald, R.; Apps, C.; et al. Genie: Generative interactive environments. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024; pp. 4603–4623. [\[CrossRef\]](#)
162. Chen, X.; Lin, K.Y.; Qian, C.; Zeng, G.; Li, H. 3D Sketch-Aware Semantic Scene Completion via Semi-Supervised Structure Prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 4193–4202. [\[CrossRef\]](#)
163. Roldao, L.; De Charette, R.; Verroust-Blondet, A. LMSCNet: Lightweight Multiscale 3D Semantic Completion. In Proceedings of the International Conference on 3D Vision, Virtual, 25–28 November 2020; pp. 111–119. [\[CrossRef\]](#)
164. Cao, A.Q.; de Charette, R. MonoScene: Monocular 3D Semantic Scene Completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 3991–4001. [\[CrossRef\]](#)
165. Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; Lu, J. Tri-perspective view for vision-based 3d semantic occupancy prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 9223–9232. [\[CrossRef\]](#)
166. Zhang, Y.; Zhu, Z.; Du, D. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 9433–9443. [\[CrossRef\]](#)
167. Scott, W.R.; Roth, G.; Rivest, J.F. View Planning for Automated Three-Dimensional Object Reconstruction and Inspection. *ACM Comput. Surv.* **2003**, *35*, 64–96. [\[CrossRef\]](#)
168. Kriegel, S.; Rink, C.; Bodenmüller, T.; Suppa, M. Efficient next-best-scan planning for autonomous 3D surface reconstruction of unknown objects. *J. Real-Time Image Process.* **2015**, *10*, 611–631. [\[CrossRef\]](#)
169. Hafner, D.; Lillicrap, T.; Fischer, I.; Villegas, R.; Ha, D.; Lee, H.; Davidson, J. Learning latent dynamics for planning from pixels. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 2555–2565. [\[CrossRef\]](#)
170. Hafner, D.; Pasukonis, J.; Ba, J.; Lillicrap, T. Mastering Diverse Domains through World Models. *arXiv* **2023**, arXiv:2301.04104. [\[CrossRef\]](#)
171. Wang, X.; Zhu, Z.; Huang, G.; Chen, X.; Zhu, J.; Lu, J. Drivedreamer: Towards real-world-drive world models for autonomous driving. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 55–72. [\[CrossRef\]](#)
172. Yang, S.; Du, Y.; Ghasemipour, K.; Tompson, J.; Kaelbling, L.; Schuurmans, D.; Abbeel, P. Learning interactive real-world simulators. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024; pp. 16504–16528. [\[CrossRef\]](#)
173. Ames, A.D.; Coogan, S.; Egerstedt, M.; Notomista, G.; Sreenath, K.; Tabuada, P. Control Barrier Functions: Theory and Applications. In Proceedings of the European Control Conference, Naples, Italy, 25–28 June 2019; pp. 3420–3431. [\[CrossRef\]](#)
174. Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; Darrell, T. Tent: Fully Test-Time Adaptation by Entropy Minimization. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021; pp. 2928–2942. [\[CrossRef\]](#)
175. Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; Tan, M. Efficient test-time model adaptation without forgetting. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 16888–16905. [\[CrossRef\]](#)
176. Xu, R.; Xiang, H.; Tu, Z.; Xia, X.; Yang, M.H.; Ma, J. V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 107–124. [\[CrossRef\]](#)
177. Xu, R.; Tu, Z.; Xiang, H.; Shao, W.; Zhou, B.; Ma, J. CoBEVT: Cooperative Bird's Eye View Semantic Segmentation with Sparse Transformers. In Proceedings of the Conference on Robot Learning, Atlanta, GA, USA, 6–9 November 2023; pp. 989–1000. [\[CrossRef\]](#)

178. Hu, Y.; Fang, S.; Lei, Z.; Zhong, Y.; Chen, S. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 4874–4886. [[CrossRef](#)]
179. Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; Abbeel, P. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Vancouver, BC, Canada, 24–28 September 2017; pp. 23–30. [[CrossRef](#)]
180. Akkaya, I.; Andrychowicz, M.; Chociej, M.; Litwin, M.; McGrew, B.; Petron, A.; Paino, A.; Plappert, M.; Powell, G.; Ribas, R.; et al. Solving Rubik’s Cube with a Robot Hand. *arXiv* **2019**, arXiv:1910.07113. [[CrossRef](#)]
181. Ma, X.; Xie, J.; Shao, D.; Yao, A.; Dong, C. RA3T: An Innovative Region-Aligned 3D Transformer for Self-Supervised Sim-to-Real Adaptation in Low-Altitude UAV Vision. *Electronics* **2025**, *14*, 2797. [[CrossRef](#)]
182. Zhang, Z. A Flexible New Technique for Camera Calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334. [[CrossRef](#)]
183. Rehder, J.; Siegwart, R.; Furgale, P. A General Approach to Spatiotemporal Calibration in Multisensor Systems. *IEEE Trans. Robot.* **2016**, *32*, 383–398. [[CrossRef](#)]
184. Forster, C.; Carlone, L.; Dellaert, F.; Scaramuzza, D. On-Manifold Preintegration for Real-Time Visual–Inertial Odometry. *IEEE Trans. Robot.* **2017**, *33*, 1–21. [[CrossRef](#)]
185. Xu, W.; Cai, Y.; He, D.; Lin, J.; Zhang, F. FAST-LIO2: Fast Direct LiDAR-Inertial Odometry. *IEEE Trans. Robot.* **2022**, *38*, 2053–2073. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.