This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

OccuFly: A 3D Vision Benchmark for Semantic Scene Completion from the Aerial Perspective

Markus Gross[1,2,3,*] Sai B. Matha [1] Aya Fahmy[1]
Rui Song[4] Daniel Cremers[2,3] Henri Meeß[1]
[1]Fraunhofer Institute IVI [2]TU Munich [3]MCML [4]UCLA

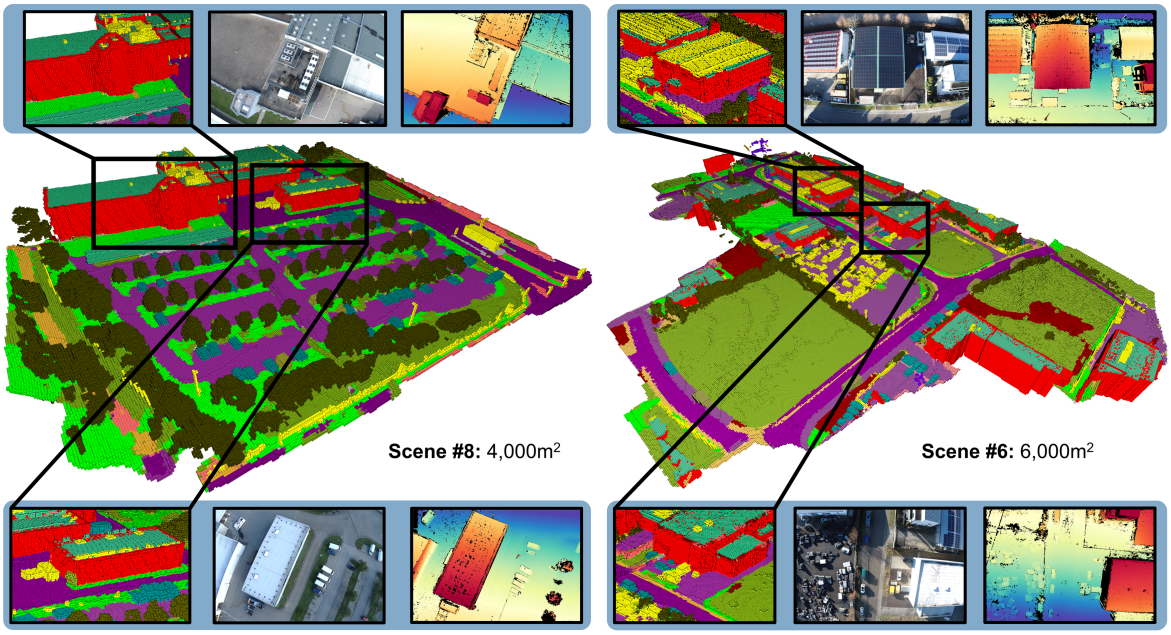


Figure 1. OccuFly introduces the first real-world, aerial 3D SSC benchmark dataset, consisting of 9 scenes with over 20,000 samples of RGB images, semantic occupancy grids, and metric depth maps, including 21 semantic classes. OccuFly covers almost 200,000 m^2 at 50 m, 40 m, and 30 m altitude in urban, industrial, and rural environments during spring, summer, fall, and winter. Zoom in for the best view.

Abstract

Semantic Scene Completion (SSC) is essential for 3D perception in mobile robotics, as it enables holistic scene understanding by jointly estimating dense volumetric occupancy and per-voxel semantics. Although SSC has been widely studied in terrestrial domains such as autonomous driving, aerial settings like autonomous flying remain largely unexplored, thereby limiting progress on downstream applications. Furthermore, LiDAR sensors are the primary modality for SSC data generation, which poses challenges for most uncrewed aerial vehicles (UAVs) due to flight regulations, mass and energy constraints, and the sparsity of LiDAR point clouds from elevated viewpoints. To address these limitations, we propose a LiDAR-free, camera-based data generation framework. By leveraging classical 3D reconstruction, our framework automates semantic label transfer by lifting <10% of annotated images into the reconstructed point cloud, substantially minimizing manual 3D annotation effort. Based on this framework, we introduce OccuFly, the first real-world, camera-based aerial SSC benchmark, captured across multiple altitudes and all seasons. OccuFly provides over 20,000 samples of images, semantic voxel grids, and metric depth maps across 21 semantic classes in urban, industrial, and rural environments, and follows established data organization for seamless integration. We benchmark both SSC and metric monocular depth estimation on OccuFly, revealing fundamental limitations of current vision foundation models in aerial settings and establishing new challenges for robust 3D scene understanding in the aerial domain. Visit https://github.com/markus-42/occufly.

*Corresponding author: markus.gross@tum.de

1. Introduction

Modern approaches in 3D computer vision enable holistic scene understanding for downstream applications such as autonomous navigation, surveillance, and augmented reality [18]. To this end, one essential approach is Semantic Scene Completion (SSC) [8], which jointly infers the complete geometry of a 3D scene from a sparse observation, such as a camera image or LiDAR scan, while simultaneously assigning semantic labels to each element, typically represented in a voxelized 3D occupancy grid [72]. Recently, SSC has been extended to Panoptic Scene Completion by incorporating instance-level awareness [27]. While strong industry funding has led the research community to focus on terrestrial environments such as autonomous driving, low-altitude aerial scene understanding for autonomous flying of uncrewed aerial vehicles (UAVs) is largely restricted to either 2D datasets [1, 7, 11, 12, 25, 34, 41, 58, 59, 61, 64, 73, 76, 81, 83, 87, 94], or 3D mesh and point cloud datasets [3, 14, 46, 52, 56, 75, 77, 90]. Notably, the SSC objective remains largely unexplored from the aerial perspective, as no dedicated real or synthetic datasets exist, thereby confining related downstream applications to terrestrial environments.

Technically, SSC datasets are typically generated by (i) fusing multiple sparse LiDAR sweeps with registered poses to capture occluded regions as a dense point cloud, where (ii) each point is manually annotated with semantic labels and (iii) subsequently voxelized to produce SSC ground-truth [93]. While effective for ground vehicles, such LiDAR-based data generation becomes challenging in aerial environments. First, multi-modal sensor setups for UAVs are not as widespread or advanced as in autonomous driving, since UAV platforms are subject to strict flight regulations, such as US [23] or EU [21] regulations, and they must adhere to stringent mass and energy constraints, which conflict with the heavier and more power-demanding nature of LiDARs compared to cameras. Second, LiDAR sparsity persists and may even worsen from an elevated vantage point, leaving significant areas unobserved and unlabeled, which in turn would yield incomplete or low-quality ground-truth.

To address these limitations, we introduce OccuFly, the first real-world, low-altitude 3D vision benchmark for aerial Semantic Scene Completion. Crucially, we propose a data generation framework that is based on camera modality, which is considered to be ubiquitous on modern UAVs. Our dataset provides over 20,000 samples, resulting in 5× the number of samples and 6× the number of voxels compared to SemanticKITTI [4], which introduced the first SSC dataset for autonomous driving. Furthermore, we evaluate the state-of-the-art on OccuFly, yielding a comprehensive 3D aerial vision benchmark.

Our **contributions** are summarized as follows:

- We propose a novel and scalable data generation framework to construct SSC ground-truth, thereby (i) relying on camera modality to avoid LiDAR-based point cloud sparsity, (ii) avoiding LiDAR hardware to adhere to mass and energy constraints of most UAVs, and (iii) reducing manual semantic labeling from tedious 3D annotation to efficient 2D annotation.
- Based on this framework, we present OccuFly, the first real-world aerial SSC benchmark consisting of 9 scenes that provide over 20,000 samples of nadir and oblique perspective images with corresponding 3D semantic voxel grids, including 21 semantic classes. OccuFly covers almost 200,000 m^2 at 50 m, 40 m, and 30 m altitude in urban, industrial, and rural environments across all seasons.
- In addition to the SSC samples, OccuFly offers more than 20,000 per-frame metric depth maps, on which we further train and release the Depth-Anything-v2 [89] depth estimation method, enabling state-of-the-art SSC.
- We identify a consistent domain gap in which state-of-the-art (i) terrestrial SSC models, (ii) 2D vision foundation models, and (iii) 3D vision foundation models fail to generalize to aerial data, thereby positioning OccuFly as a valuable and robust dataset to address these limitations.

2. Related Work

Datasets and Benchmarks. Apart from indoor SSC [69], the first outdoor SSC dataset was proposed by SemanticKITTI [4] for autonomous driving. Despite its impact, the limited scale and diversity of SemanticKITTI impeded the development of generalizable SSC models and their comprehensive evaluation [48]. To address these limitations, multiple impactful datasets followed, such as nuScenes [6], Waymo [71], and KITTI-360 [50], all relying on LiDAR modality for data generation. As elaborated in Sec. 1, LiDAR-based SSC ground-truth generation faces fundamental challenges stemming from sparse point clouds, as many regions remain unobserved and consequently unlabeled. Occlusions, misalignment from aggregating multiple sweeps, and dynamic objects further exacerbate these gaps, creating inaccuracies in the resulting volumetric labels. Additionally, the process of manually annotating such sparse 3D point clouds is both time-consuming and error-prone, undermining scalability for large-scale 3D reconstruction tasks.

To mitigate these challenges, several benchmarks have emerged. Specifically, Occ3D [74] combines multi-sweep densification with mesh reconstruction and camera filtering to reduce occlusions and label noise. Meanwhile, OpenOccupancy [78] combines pseudo-label generation with extensive human annotation to double the density of occupancy labels and refine boundaries. In parallel, OCFBench [54] synchronizes dynamic objects to address spatiotemporal occlusions, and excludes unknown voxels through ray-casting, thereby mitigating LiDAR sparsity. Similarly, SSCBench [48] ex-

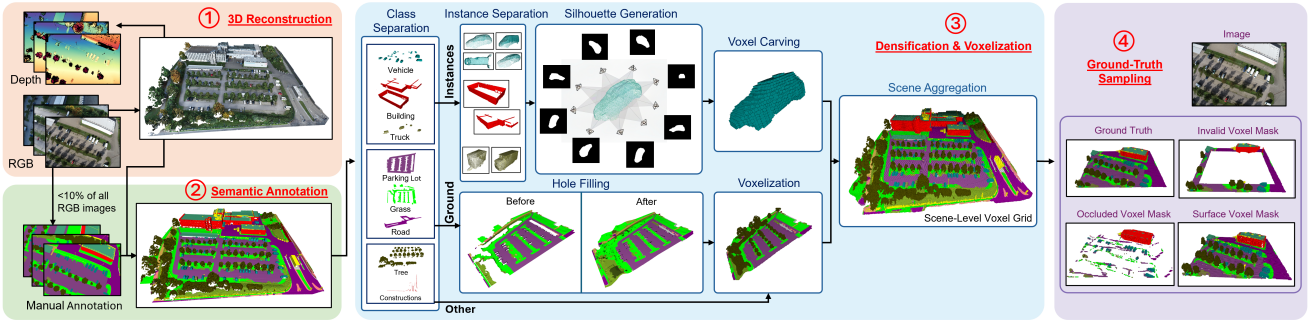


Figure 2. Proposed image-based data generation framework. An overview is provided in Sec. 3.1. Zoom in for the best view.

cludes unobserved voxels from training and evaluation, and integrates data from multiple sources to enhance geographic diversity. Additionally, OpenScene [13] leverages voxel densification and flow integration to fill unobserved regions. Finally, UniOcc [80] unifies real and synthetic (CARLA [67], OpenCOOD [86]) datasets, and provides label-free evaluation metrics to address suboptimal LiDAR coverage. Notably, these benchmarks obtain depth maps implicitly by projecting LiDAR points onto images, relying on precise sensor extrinsics that can introduce geometric inaccuracies.

In contrast, our proposed data generation framework reconstructs metric 3D point clouds from geo-referenced imagery. We unproject manually annotated 2D semantic masks into these reconstructions via 2D–3D correspondences, yielding accurate per-point semantics, which substantially minimizes 3D annotation effort. We further refine and densify the semantic point cloud, voxelize it, and extract consistent per-frame ground-truth voxel grids and metric depth maps.

Methods. Apart from recurrent [79] and multi-view SSC methods [9, 28, 67, 82], single-view SSC was pioneered by MonoScene [8], which bridges 2D and 3D representations through an optics-inspired feature projection and a 3D context relation prior. Recent advancements have explored diverse feature representations to bridge 2D image cues with volumetric representations, broadly categorized into planar and voxel-based approaches [85].

In particular, planar representations, such as bird's-eye-view (BEV) [49] or tri-perspective views (TPV), enable compact feature aggregation and long-range context modeling. TPVFormer [32] introduces the TPV framework with perpendicular planes and a transformer encoder for lifting multi-view features into semantic occupancy grids, while DISC [53] extends BEV with discriminative queries and dual-attention decoders to disentangle instance-scene contexts for category-specific interactions. Building on hybrid designs, CGFormer [91] fuses voxel and TPV spaces via context-aware queries and 3D deformable cross-attention, enhancing depth-aware aggregation for superior fidelity.

In contrast, voxel-based methods directly operate on 3D grids for fine-grained occupancy and semantics. Two-stage transformers like VoxFormer [47] and VisHall3D [57] employ sparse queries from depth priors followed by densification or visibility-aware decoupling to handle visible and occluded regions, respectively. Symphonies [36] refines voxel queries with class-centric instance propagation for dynamic 2D-3D reasoning, whereas SOAP [43] incorporates adaptive decoders drawing from semantic repositories. Notably, many of these methods [36, 47, 53, 91] apply depth estimation to initialize voxel proposals as geometric priors.

Specialized enhancements include tri-axis approaches [2], State Space Models [44, 45], self-supervised methods [30, 35], and Gaussian representations [33]. We refer the interested reader to the surveys of [85, 93] for a comprehensive discussion.

3. Data Generation Framework

3.1. Overview

Our proposed data generation framework consists of four modules, shown in Fig. 2. We utilize geo-referenced images to apply traditional multi-view reconstruction, generating a metric 3D point cloud (Sec. 3.3.1). This approach additionally yields metric depth maps and 2D–3D correspondences, allowing image pixels to be associated with reconstructed 3D points, effectively streamlining the creation of 3D semantic annotations. Technically, we enable highly efficient label transfer by manually annotating only a small subset of the camera images ($<10\,\%$ on average) and lifting the semantic pixels into the point cloud (Sec. 3.3.2). This reduces costly 3D annotation to efficient 2D image labeling, substantially lowering annotation effort. Subsequently, we densify individual objects with our novel densification pipeline and eventually voxelize the semantic point cloud (Sec. 3.3.3). As all previous steps are performed on a global scene level, we finally retrieve per-frame ground-truth grids by frustum-culling the scene voxel grid using geo-referenced camera poses and intrinsics, resulting in one fixed-size semantic voxel grid per camera frame. Similar to SemanticKITTI [5], we additionally construct binary masks to distinguish surface, occluded, and invalid voxels (Sec. 3.3.4).

3.2. Problem Formulation

Given a set of N geo-referenced, calibrated RGB images $\mathcal{I} = \{\mathbf{I}_n \in \mathbb{R}^{H\times W\times 3}\}_{n=1}^{N}$, acquired by a pinhole camera with per-frame intrinsics $\mathcal{K} = \{\mathbf{K}_n \in \mathbb{R}^{3\times 3}\}_{n=1}^{N}$, and world-to-camera poses $\mathcal{T} = \{\mathbf{T}_{c\leftarrow w}^{n} \in \mathrm{SE}(3)\}_{n=1}^{N}$, we adopt a fixed-size metric voxel-grid specification (X, Y, Z, r), where X, Y, Z denote the number of voxels along cartesian axes, with voxel edge length $r > 0$. Moreover, we define a semantic label set $\mathcal{C} = \{1, 2, \ldots, C\}$. Each image $\mathbf{I}_n$ is coupled to a ground-truth semantic voxel grid $\mathbf{Y}_n \in \mathcal{C}^{X\times Y\times Z}$, and a metric depth map $\mathbf{D}_n \in \mathbb{R}^{H\times W}$, yielding the dataset as image–grid-depth samples $\{(\mathbf{I}_n, \mathbf{Y}_n, \mathbf{D}_n)\}_{n=1}^{N}$, organized into multiple scenes with scene-dependent sample counts.

3.3. Method

3.3.1. 3D Reconstruction

From the calibrated camera intrinsics $\mathcal{K}$, and the geo-referenced images $\mathcal{I}$ and their poses $\mathcal{T}$, we obtain a metric scene reconstruction via Structure-from-Motion (SfM) [65] and Multi-View Stereo (MVS) [66], abstracted as

$$(\mathcal{P},\, \mathcal{D},\, \mathcal{A}) \;=\; \Psi_{\text{SfM+MVS}}(\mathcal{I}, \mathcal{K}, \mathcal{T}), \tag{1}$$

where the set $\mathcal{P} = \{\mathbf{x}_m \in \mathbb{R}^3\}_{m=1}^{M}$ represents a dense point cloud with $M = |\mathcal{P}|$ number of points in world coordinates, and $\mathcal{D} = \{\mathbf{D}_n \in \mathbb{R}^{H\times W}\}_{n=1}^{N}$ are per-image metric depth maps. Moreover, $\mathcal{A}_n$ denotes per-image 2D–3D correspondences , which satisfy the projection of a 3D point $\mathbf{x}$ onto a 2D image pixel (u, v), formulated as

$$\mathcal{A}_n = \left\{\, ((u, v), \mathbf{x}) \;\middle|\; (u, v, 1)^\top \sim \mathbf{K}_n[\,\mathbf{R}_n \,|\, \mathbf{t}_n\,][\mathbf{x}^\top\ 1]^\top \right\}, \tag{2}$$

where $\sim$ denotes equality up to a non-zero scalar, and $\mathbf{R}_n$ and $\mathbf{t}_n$ are the rotational and translational components of $\mathbf{T}_{c\leftarrow w}^{n}$, respectively. All correspondences of a scene are given by $\mathcal{A} = \bigcup_{n=1}^{N} \mathcal{A}_n$.

3.3.2. Semantic Annotation

We define a finite, non-empty set of semantic classes $\mathcal{C} = \{1, \ldots, C\}$. To minimize manual annotation effort, we annotate only a small subset $\mathcal{J} \subset \{1, \ldots, N\}$ of images, exploiting the fact that each 3D point is observed by multiple cameras. Note that $\mathcal{J}$ denotes the index set of images, not the images themselves. We select $\mathcal{J}$ via spatially stratified sampling across the scene. More specifically, we partition the ground plane into a regular grid of square cells with scene-dependent edge lengths of 23 to 28 meters. For each cell center, we select the image whose pose is closest to that location. Additionally, for cells at the scene border, we select the pose closest to the border. Using the 2D–3D correspondences $\mathcal{A}$ (see Sec. 3.3.1), we quantify the coverage of reconstructed points by

$$\rho(\mathcal{J}) \;=\; \frac{\left|\{\, \mathbf{x} \in \mathcal{P} \,|\, \exists\, n \in \mathcal{J},\, \exists\, (u, v) \text{ s.t. } ((u, v), \mathbf{x}) \in \mathcal{A}_n \,\}\right|}{|\mathcal{P}|} \tag{3}$$

Empirically, $\rho(\mathcal{J}) > 0.99$, while $|\mathcal{J}|/N < 0.1$ on average, implying that annotating less than 10% of all images covers more than 99% of all 3D points, further detailed in Sec. 4.3.

After manually annotating all images in $\mathcal{J}$, we lift semantic labels from images to points via back-projection. Note that this step is similar to [88], but adapted to a LiDAR-free, camera-only aerial setting based on SfM+MVS reconstruction. As most of the points are observed by multiple cameras, we fuse multi-view evidence to assign robust per-point labels. Technically, this is accomplished by unweighted majority voting, where ties are broken by a fixed class-prior order derived from class frequencies. Furthermore, to annotate unlabeled points, we apply k-nearest-neighbor (kNN) with inverse-distance weights within a fixed neighborhood. As a denoising step, we apply a second iteration of kNN to all points, effectively relabeling every point to the dominant class in its neighborhood, similar to [74]. The resulting semantic point cloud $\mathcal{P}_{\mathcal{S}} = \{(\mathbf{x}_m, c_m)\}_{m=1}^{M}$ consists of 3D points $\mathbf{x}_m$ and corresponding semantic class labels $c_m \in \mathcal{C}$.

3.3.3. Class-Aware Densification and Voxelization

Preliminaries. Given the semantic point cloud $\mathcal{P}_{\mathcal{S}} = \{(\mathbf{x}_m, c_m)\}_{m=1}^{M}$ from Sec. 3.3.2, we first partition the semantic classes into three disjoint groups

$$\mathcal{C}_{\text{inst}} \,\cup\, \mathcal{C}_{\text{gnd}} \,\cup\, \mathcal{C}_{\text{oth}} \;=\; \mathcal{C}, \qquad \mathcal{C}_{\text{inst}} \cap \mathcal{C}_{\text{gnd}} \cap \mathcal{C}_{\text{oth}} = \varnothing, \tag{4}$$

corresponding to instance classes $\mathcal{C}_{\text{inst}}$ that are to be densified object-wise (*e.g.*, vehicles), ground classes $\mathcal{C}_{\text{gnd}}$ that are to be surface-reconstructed (*e.g.*, road), and other classes $\mathcal{C}_{\text{oth}}$ that are directly voxelized (*e.g.*, constructions). Based on this class partition, we apply **group separation** to retrieve group-specific point cloud subsets

$$\mathcal{P}_{\text{inst}} = \{(\mathbf{x}, c) \in \mathcal{P}_{\mathcal{S}} \,|\, c \in \mathcal{C}_{\text{inst}}\}\,, \tag{5}$$

$$\mathcal{P}_{\text{gnd}} = \{(\mathbf{x}, c) \in \mathcal{P}_{\mathcal{S}} \,|\, c \in \mathcal{C}_{\text{gnd}}\}\,, \tag{6}$$

$$\mathcal{P}_{\text{oth}} = \{(\mathbf{x}, c) \in \mathcal{P}_{\mathcal{S}} \,|\, c \in \mathcal{C}_{\text{oth}}\}\,. \tag{7}$$

Furthermore, let $\mathcal{G} = [X] \times [Y] \times [Z]$ be a voxel grid. For any point cloud subset $\mathcal{Q} \subset \mathcal{P}$ and a target voxel resolution $r > 0$, we denote by $\mathrm{Vox}_r(\mathcal{Q}) \subset \mathcal{G}$ the set of occupied voxels obtained by standard binning (point rasterization) [22] at resolution r. For triangle meshes, we use the same notation to indicate triangle-to-voxel scan conversion [37].

Instance classes. We apply **instance separation** by decomposing $\mathcal{P}_{\text{inst}}$ into object instances using Euclidean clustering via DBSCAN [20] with class-specific parameters $(\varepsilon_c, \text{minPts}_c)$. This process yields J object instances

$$\mathbb{S} = \text{DBSCAN}(\mathcal{P}_{\text{inst}}, \varepsilon_c, \text{minPts}_c) = \{\mathcal{S}_j \subset \mathcal{P}_{\text{inst}}\}_{j=1}^{J}, \tag{8}$$

where $\mathcal{S}_j$ represents an instance point cloud.

To create the voxelized visual hull from an instance point cloud $\mathcal{S}$, we perform two major steps: Silhouette extraction

and silhouette-based voxel carving. For each instance $\mathcal{S} \in \mathbb{S}$, let $\mathrm{pos}(\mathcal{S}) = \{\, \mathbf{x} \,|\, (\mathbf{x}, c) \in \mathcal{S} \,\}$ denote its 3D positions.

To **extract silhouettes**, we (i) place K virtual cameras $\mathbb{V} = \{(\mathbf{K}_k, \mathbf{T}^k_{c\leftarrow w})\}_{k=1}^K$ quasi-uniformly distributed on the viewing sphere around $\mathrm{pos}(\mathcal{S})$ (in practice we use K=24); (ii) project $\mathrm{pos}(\mathcal{S})$ to each view to obtain 2D point sets $\mathcal{U}_k = \{\, \pi_k(\mathbf{x}) \,|\, \mathbf{x} \in \mathrm{pos}(\mathcal{S}) \,\}$, where π_k is the pinhole projection induced by $(\mathbf{K}_k, \mathbf{T}^k_{c\leftarrow w})$; and (iii) compute a binary silhouette $\Omega_k \subset \mathbb{R}^2$ via the α-shape boundary [19] of $\mathcal{U}_k$, where Ω_k is the set of pixels that belong to the instance.

For **voxel carving**, we apply multi-view silhouette carving [42]. To this end, we back-project each silhouette to a generalized cone

$$\mathcal{R}_k \;=\; \{\, \mathbf{x} \in \mathbb{R}^3 : \pi_k(\mathbf{x}) \in \Omega_k \,\}. \tag{9}$$

The continuous instance hull is $\mathcal{H}(\mathcal{S}) = \bigcap_{k=1}^K \mathcal{R}_k$. To carve within a finite space around the instance hull, let $\mathcal{B}(\mathcal{S})$ be a tight axis-aligned bounding box of $\mathrm{pos}(\mathcal{S})$, dilated by a small margin. We discretize $\mathcal{B}(\mathcal{S})$ into a 3D grid of voxels indexed by $\mathbf{v} = (i, j, k)$. The carved occupancy set is

$$\mathcal{O}_{\text{inst}}(\mathcal{S}) = \{\mathbf{v} \in \mathcal{G} \,|\, \text{center}(\mathbf{v}) \in \mathcal{B}(\mathcal{S}) \cap \mathcal{H}(\mathcal{S})\}\,. \tag{10}$$

All voxels $\mathbf{v} \in \mathcal{O}_{\text{inst}}(\mathcal{S})$ receive the semantic label of its original semantic instance point cloud $\mathcal{P}_{\text{inst}}$. Finally, we aggregate all instances via $\mathcal{O}_{\text{inst}} = \bigcup_{\mathcal{S}\in\mathbb{S}} \mathcal{O}_{\text{inst}}(\mathcal{S})$.

Ground classes. We densify $\mathcal{P}_{\text{gnd}}$ via Poisson surface reconstruction [38] to obtain a watertight triangle mesh $\mathcal{M}_{\text{gnd}} = \Psi_{\text{Poisson}}(\mathcal{P}_{\text{gnd}})$, which **fills holes** and enforces surface continuity, similar to [82]. We then voxelize the mesh:

$$\mathcal{O}_{\text{gnd}} \;=\; \text{Vox}_r(\mathcal{M}_{\text{gnd}}), \tag{11}$$

assigning to each occupied voxel the majority ground class of contributing mesh samples in its cell.

Other classes. For $\mathcal{P}_{\text{oth}}$, we apply direct voxelization:

$$\mathcal{O}_{\text{oth}} \;=\; \text{Vox}_r(\mathcal{P}_{\text{oth}}), \tag{12}$$

with per-voxel semantics determined by majority voting of points falling into the voxel.

Aggregation. We construct a scene-level semantic voxel grid by combining all groups with a fixed precedence order inst $\succ$ oth $\succ$ gnd to resolve label conflicts:

$$\mathbf{Y}(\mathbf{v}) \;=\; \begin{cases} \text{label from } \mathcal{O}_{\text{inst}}, & \mathbf{v} \in \mathcal{O}_{\text{inst}}, \\ \text{label from } \mathcal{O}_{\text{oth}}, & \mathbf{v} \in \mathcal{O}_{\text{oth}} \setminus \mathcal{O}_{\text{inst}}, \\ \text{label from } \mathcal{O}_{\text{gnd}}, & \mathbf{v} \in \mathcal{O}_{\text{gnd}} \setminus (\mathcal{O}_{\text{inst}} \cup \mathcal{O}_{\text{oth}}), \\ 0, & \text{otherwise}, \end{cases} \tag{13}$$

where 0 denotes empty. The resulting $\mathbf{Y}$ constitutes the semantic voxel grid for the whole scene.

3.3.4. Ground-Truth Sampling

Frustum Culling. Given the scene-level semantic voxel grid $\mathbf{Y} \in (\{0\} \cup \mathcal{C})^{X\times Y\times Z}$ and per-frame camera parameters $(\mathbf{K}_n, \mathbf{T}^n_{c\leftarrow w})$, we construct per-frame ground-truth by frustum-culling and rasterization at a fixed metric specification (X, Y, Z, r). Let π_n denote the pinhole projection induced by $(\mathbf{K}_n, \mathbf{T}^n_{c\leftarrow w})$, and let $[d_{\min}, d_{\max}]$ be near/far clipping distances, respectively. Define the truncated frustum

$$\mathcal{F}_n \;=\; \big\{\, \mathbf{x} \in \mathbb{R}^3 \;:\; \pi_n(\mathbf{x}) \in [0, W)\times[0, H) \big\},$$

with $d_{\min} \le d_n(\mathbf{x}) \le d_{\max}$, where $d_n(\mathbf{x})$ is the camera-centric depth of $\mathbf{x}$. We obtain the per-frame grid $\mathbf{Y}_n \in (\{0\} \cup \mathcal{C})^{X\times Y\times Z}$ by discretizing $\mathcal{F}_n$ at resolution r and sampling $\mathbf{Y}$ at voxel centers:

$$\mathbf{Y}_n(\mathbf{v}) \;=\; \begin{cases} \mathbf{Y}(\mathbf{v}), & \text{if center}(\mathbf{v}) \in \mathcal{F}_n, \\ 0, & \text{otherwise}, \end{cases} \qquad \mathbf{v} \in \mathcal{G}.$$

Binary Masks. In addition, we construct three binary masks, similar to SemanticKITTI [5]: invalid $\mathbf{M}^{\text{inv}}_n$, surface $\mathbf{M}^{\text{surf}}_n$, and occluded $\mathbf{M}^{\text{occ}}_n$, all in $\{0, 1\}^{X\times Y\times Z}$.

Invalid Mask. The invalid mask represents voxels outside the field of view:

$$\mathbf{M}^{\text{inv}}_n(\mathbf{v}) \;=\; \mathbf{1}\big[\,\text{center}(\mathbf{v}) \notin \mathcal{F}_n\,\big], \tag{14}$$

where $\mathbf{1}[\cdot]$ denotes the indicator (Iverson) function.

Surface Mask (view-independent). Let $\mathcal{N}_6(\mathbf{v})$ be the 6-neighborhood in the grid, and let $\mathcal{E}(\mathbf{v}) = \mathbf{1}[\mathbf{Y}_n(\mathbf{v}) \neq 0]$ denote occupancy. We mark geometric boundary voxels by

$$\mathbf{M}^{\text{surf}}_n(\mathbf{v}) \;=\; \mathbf{1}\Big[\,\mathcal{E}(\mathbf{v}) = 1 \;\wedge\; \exists\, \mathbf{u} \in \mathcal{N}_6(\mathbf{v}) \,\Big|\, \mathcal{E}(\mathbf{u}) = 0\,\Big].$$

Occluded Mask (view-dependent). For each pixel $(u, v) \in [0, W)\times[0, H)$, consider the ordered set of frustum voxels $\mathcal{V}_n(u, v) = \langle \mathbf{v}_1, \mathbf{v}_2, \dots \rangle$ traversed by the ray $\rho_n(u, v)$ from near to far. Let $i^\star = \min\{i : \mathcal{E}(\mathbf{v}_i) = 1\}$ if such i exists. Then we assign

$$\mathbf{M}^{\text{occ}}_n(\mathbf{v}_j) \;=\; \begin{cases} 0, & j = i^\star, \\ 1, & j > i^\star \text{ and } \mathcal{E}(\mathbf{v}_j) = 1, \\ 0, & \text{otherwise}, \end{cases}$$

for all $\mathbf{v}_j \in \mathcal{V}_n(u, v)$. We set $\mathbf{M}^{\text{occ}}_n(\mathbf{v}) = 0$ if $\mathbf{v}$ lies outside all rays or is invalid.

Finally, the per-frame ground-truth sample consists of $(\mathbf{I}_n, \mathbf{Y}_n, \mathbf{M}^{\text{inv}}_n, \mathbf{M}^{\text{surf}}_n, \mathbf{M}^{\text{occ}}_n)$, where invalid voxels are excluded from evaluation, surface voxels represent view-independent geometric boundaries, and occluded voxels capture view-dependent occupied regions behind the first visible surface along camera rays.

Table 1. OccuFly dataset statistics, discussed in Sec. 4.2.

Scene	Season	Environment	Area [m^2]	Number of Samples			
				50 m	40 m	30 m	Total
Training							**14,804**
01	Winter	Rural	41,234	406	469	512	1,387
02	Winter	Urban	8,529	277	468	132	877
03	Spring	Urban	25,077	871	1,315	1,427	3,613
04	Spring	Industrial	55,579	1,140	1,258	1,564	3,962
05	Summer	Rural	24,810	1,006	1,682	2,277	4,965
Validation							**1,965**
06	Spring	Urban	5,428	366	266	327	959
07	Spring	Industrial	5,802	279	384	343	1,006
Test							**3,842**
08	Fall	Industrial	4,314	183	316	383	882
09	Spring	Urban	23,165	1,304	1,416	240	2,960
Total			**193,938**	**5,832**	**7,574**	**7,205**	**20,611**

Table 2. Comparison of terrestrial and aerial vision-based SSC benchmark datasets, detailed in Sec. 4.3.

	Camera Views	# of Samples	Depth Maps	Semantic Classes	Grid Resolution $X \times Y \times Z$	r
Terrestrial Benchmarks						
SemanticKITTI [5]	single	4,649	✗	19	$256 \times 256 \times 32$	0.20
OpenOccupancy [78]	multi	34,149	✗	16	$512 \times 512 \times 40$	0.20
SSCBench-Waymo [48]	multi	19,985	✗	14	$256 \times 256 \times 32$	0.20
SSCBench-nuScenes [48]	multi	34,078	✗	16	$256 \times 256 \times 32$	0.20
SSCBench-KITTI-360 [48]	multi	12,865	✗	19	$256 \times 256 \times 32$	0.20
Occ3D-Waymo [74]	multi	200,000	✗	14	$3200 \times 3200 \times 128$	0.05
Occ3D-nuScenes [74]	multi	40,000	✗	16	$200 \times 200 \times 16$	0.40
Aerial Benchmarks						
OccuFly (ours)	single	20,611	✓	21	$192 \times 128 \times 128$	0.50

Table 3. Comparison of OccuFly depth maps to other real-world low-altitude aerial datasets containing metric depth maps (Sec. 4.3).

Dataset	# Depth Maps	Density	Environments	Seasons
WildUAV [24]	∼1,500	Dense	Rural	Summer, Autumn
UseGeo [60]	829	Sparse	Urban	n.a.
OccuFly (ours)	20,611	Dense	Rural, Urban, Industrial	Spring, Summer, Autumn, Winter

4. OccuFly Dataset

4.1. Data Collection

We utilize two DJI UAV platforms for photogrammetric data acquisition [26]: The Phantom 4 RTK (P4) [15] and the DJI Mavic 3 Enterprise Series (M3-ES) [16], capturing images at 5472×3648 and 4000×3000 pixels, respectively. Geo-referenced camera poses and orientations are recorded by each flight controller via onboard sensor fusion (GNSS, IMU, compass, and magnetometer) [17]. We collect data from nine scenes in urban, industrial, and rural environments, spanning spring, summer, fall, and winter. All data were collected within the broader geographic region of Ingolstadt, Germany. For data acquisition, we executed automated double-grid flight patterns at altitudes of 50 m, 40 m, and 30 m. These missions yielded ground sampling distances (GSD) of 1.4, 1.1, and 0.8 cm/pixel with the P4 platform, and 6.7, 5.4, and 4.1 cm/pixel with the M3-ES platform. The double-grid pattern provided, on average, 67 % side and 74 % forward image overlap. For oblique acquisitions, camera tilt angles were set to $-75°$ at 50 m and 40 m, and $-70°$ at 30 m. Additionally, using the M3-ES UAV, we collected nadir ($0°$ tilt) imagery at all altitudes. Further details are provided in the supplementary material.

4.2. Dataset Statistics

We summarize OccuFly in Tab. 1. The dataset comprises 9 scenes and more than 20,000 annotated samples, each including (i) an RGB image, (ii) a semantic voxel grid, and (iii) a metric depth map. Voxel grids are annotated with 21 semantic classes that are detailed in the supplementary, including class frequencies. OccuFly spans approximately 193,938 m^2 across altitudes of 50 m, 40 m, and 30 m, covering urban, industrial, and rural environments in spring, summer, fall, and winter. 3D space is discretized into voxel grids of resolution $192 \times 128 \times 128$ with voxel size $r = 0.5$ m. Ground-truth at 40 m and 30 m is generated via frustum culling (Sec. 3.3.4) of the 50 m scene-level semantic grid. Moreover, the dataset follows the SSCbench [48] data organization to streamline integration for existing methods. Finally, we report semantic group assignments (Sec. 3.3.3) together with scene-wise depth distributions in the supplementary material.

4.3. Dataset Evaluation

SSC Ground-Truth. Since OccuFly provides the first real-world aerial SSC dataset, we compare it against established terrestrial datasets (Tab. 2). Similar to SemanticKITTI [5], which introduced real-world SSC to autonomous driving, OccuFly introduces real-world SSC to the aerial domain, but at a substantially larger scale: the number of samples is more than five times higher, and the total number of labeled voxels is over six times larger than SemanticKITTI. OccuFly further provides the largest class taxonomy while adhering to the SSCBench-style data structure [48] for seamless integration.

Depth Map Ground-Truth. In Tab. 3, we compare OccuFly to other real-world, low-altitude aerial datasets that include metric depth maps. To the best of our knowledge, WildUAV [24] and UseGeo [60] are the only publicly available datasets of this kind. OccuFly is substantially larger, providing more than 13× and more than 24× as many metric depth maps, respectively, while spanning a broader range of environments and seasons. This positions OccuFly as the largest and most diverse publicly available low-altitude metric depth estimation dataset to date, which further enables holistic vision-based 3D scene understanding, such as SSC.

3D Reconstruction. We evaluate the 3D foundation model DepthAnything3 [51] on all scenes of OccuFly, as illustrated in Fig. 3, and observe an average metric scale deviation of 526 %. This result supports our choice of classical SfM+MVS reconstruction and aligns with prior work [84],

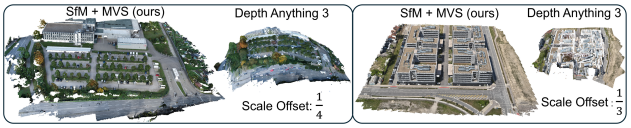


Figure 3. Evaluation of our classical 3D reconstruction compared to DepthAnything3 [51] foundation model, detailed in Sec. 4.3.

indicating that current 3D foundation models struggle to reliably handle aerial imagery. Additionally, we quantify geometric consistency of the SfM+MVS pipeline using the standard RMSE reprojection error. Across all scenes, an average reprojection error of 1.24 pixels demonstrates strong geometric fidelity [29], reflecting precise 2D-3D alignment under high-resolution inputs. Scene-wise reconstructions and reprojection errors are provided in the supplementary.

Semantic Annotation Efficiency. As an alternative to manual annotation, we evaluate the use of 2D pseudo-labels following prior work such as PointSSC [88]. Using the 2D foundation model Semantic Segment Anything [10] to generate pseudo-labels yields poor segmentation quality, which we attribute to its backbone Segment Anything Model (SAM) [40] that underperforms in the aerial domain [62]. To provide a stronger baseline, we train a ConvNeXt [55] model on a consolidation of seven established aerial segmentation datasets [7, 12, 34, 58, 61, 63, 70]. However, evaluation against our manual annotations results in only 17.58 % mIoU, indicating that pseudo-labels remain highly unreliable in the aerial domain and reinforcing the need for accurate manual labeling. To this end, our method leverages multi-view correspondences to lift only a small subset of manually annotated images to 3D. Specifically, annotating fewer than 10% of the images per scene suffices to automatically label over 99% of reconstructed points, as determined by Eq. (3), demonstrating a highly efficient and scalable annotation strategy. We report scene-wise annotation ratios and coverage statistics in the supplementary material.

Geometric Fidelity and Semantic Accuracy. Using RTK-based camera poses, we render the semantic point cloud into 135 additional manual annotations across all scenes and altitudes that were not used during data generation. This evaluation achieves 92 % pixel-wise agreement, confirming high geometric fidelity and semantic accuracy.

5. Benchmark Experiments

5.1. Experimental Setup

Aerial Semantic Scene Completion. We benchmark the state-of-the-art SSC methods Symphonies [36] and DISC [53], using official implementations and evaluation protocols to ensure scientific rigor. As these methods require depth maps to back-project geometric priors (Sec. 2), we replace their terrestrial depth estimator with our DepthAnything2-Occufly model (discussed in the next paragraph). Further details are provided in the supplementary.

Table 4. Altitude-wise quantitative SSC evaluation for DISC [53] and Symphonies [36] on the OccuFly test set, discussed in Sec. 5.2.

Altitude [m]	Method	IoU [%]	mIoU [%]
50	Symphonies [36]	15.88	0.58
	DISC [53]	**31.10**	**2.20**
40	Symphonies [36]	10.71	0.52
	DISC [53]	**27.85**	**1.77**
30	Symphonies [36]	13.22	0.76
	DISC [53]	**26.88**	**2.23**
All	Symphonies [36]	13.68	0.58
	DISC [53]	**29.52**	**2.04**

Table 5. Depth estimation evaluation, comparing zero-shot vs. fine-tuned foundation models on the OccuFly test set (see Sec. 5.2).

Method	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	AbsRel↓	RMSE↓	MAE↓	SILog↓
MapAnythingV1.1 [39]	0.000	0.000	0.003	0.799	30.068	29.309	0.069
Metric3Dv2 [31]	0.073	0.208	0.455	0.471	19.578	18.409	0.156
DepthAnything2 [89]	0.002	0.015	0.059	0.729	28.382	27.392	0.192
DepthAnything3 [51]	0.000	0.005	0.141	0.591	22.615	22.019	**0.043**
Metric3Dv2-OccuFly	0.278	0.806	0.985	0.381	13.643	13.134	0.098
DepthAnything2-OccuFly	**0.834**	**0.976**	**0.999**	**0.134**	**4.844**	**4.193**	0.112

Metric Monocular Depth Estimation. This task is crucial for vision-based SSC, as state-of-the-art methods [36, 47, 53, 92] initialize 3D geometric priors via back-projecting metric depth maps (see Sec. 2). Notably, no established metric mono depth models exist for the aerial domain. We therefore evaluate the potential of OccuFly's metric depth maps by benchmarking foundation models in zero-shot and fine-tune settings for MapAnythingV1.1 [39], Metric3Dv2 [31], and DepthAnything version 2 [89] and version 3 [51], using established metrics and official implementations. Note that MapAnything cannot be fine-tuned, as it is trained for 3D reconstruction and produces depth maps only as a byproduct. Furthermore, DepthAnything3 does not provide train scripts. We therefore zero-shot evaluate it and fine-tune v2 instead.

5.2. Quantitative Results

Aerial Semantic Scene Completion. Quantitative results in Tab. 4 indicate that state-of-the-art models can recover coarse scene geometry, yet struggle to infer semantic structure. Our detailed altitude and class-wise analysis provided in the supplementary material further substantiates these observations. Overall, our findings highlight key challenges specific to aerial SSC and position OccuFly as a robust and valuable benchmark for advancing research in this domain.

Metric Monocular Depth Estimation. Results in Tab. 5 show that foundation models underperform significantly in aerial environments, but improve markedly when fine-tuned on OccuFly depth maps. Notably, the SILog metric, which captures scale-invariant performance (*i.e.*, relative, non-metric accuracy), shows a slight degradation after fine-tuning, indicating a minor trade-off between relative con-

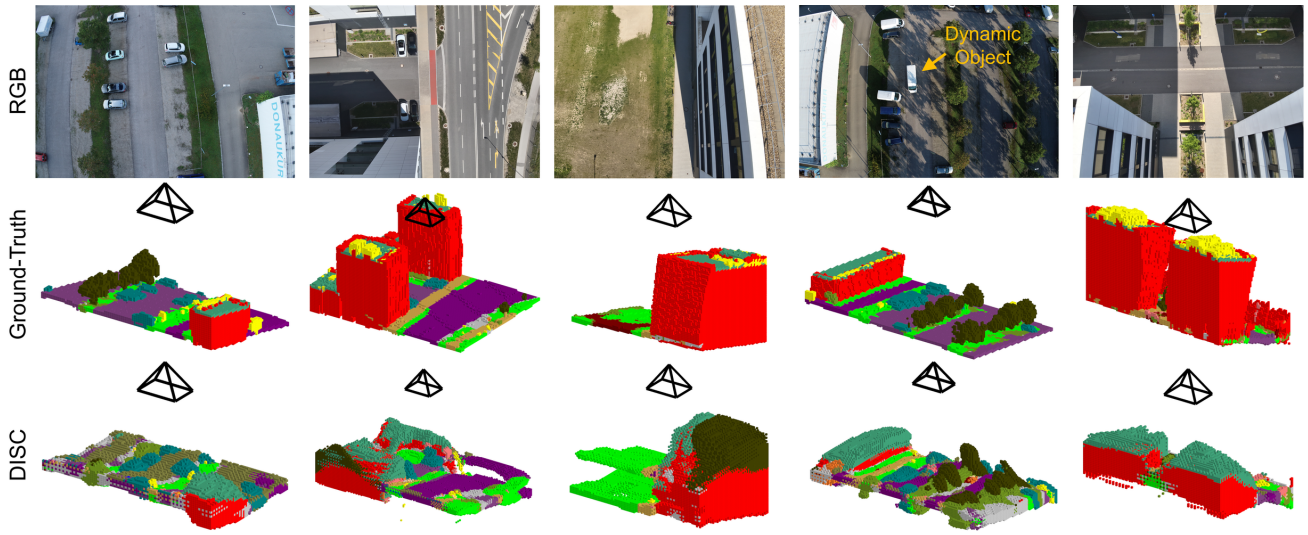


Figure 4. Visual SSC examples and qualitative SSC evaluation for DISC [53] on the OccuFly test set, discussed in Sec. 5.3.

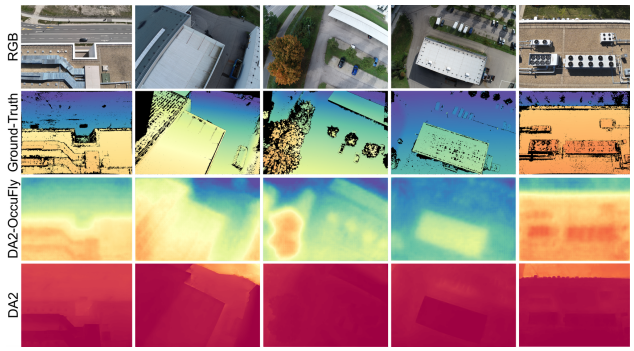


Figure 5. Qualitative evaluation of zero-shot and fine-tuned depth estimation for DepthAnything2 [89] on OccuFly test set (Sec. 5.3).

sistency and metric correctness. Altitude-wise evaluation in the supplementary reveals a steady performance decline with increasing altitude, suggesting that higher viewpoints introduce greater challenges for accurate depth estimation. Taken together, these findings demonstrate that altitude is a critical factor in aerial depth prediction and that targeted adaptation can substantially improve performance, establishing OccuFly as a significant contribution for advancing metric 3D understanding in aerial vision.

5.3. Qualitative Results

Aerial Semantic Scene Completion. Qualitative results in Fig. 4 substantiate the quantitative findings, showing that models capture coarse geometric layout but fail to produce consistent and accurate semantic predictions. These shortcomings are particularly evident in complex aerial viewpoints, reflecting a domain gap for models primarily developed on terrestrial data. By exposing these limitations, OccuFly highlights key challenges in aerial SSC and provides a rigorous testbed to drive progress in this domain.

Metric Monocular Depth Estimation. Qualitative depth visualizations in Fig. 5 utilize a single, metrically consistent colormap across all depth maps, allowing for a direct comparison of absolute ranges. Fine-tuned DepthAnything2-OccuFly (DA2-OccuFly) visually reconstitutes the ground-truth topology with coherent depth gradients and realistic color distributions, indicating well-calibrated metric estimates. In contrast, zero-shot DepthAnything2 (DA2) often shows sharper object boundaries but is dominated by saturated red hues, evidencing a systematic underestimation of distance and poor absolute scaling. This visual mismatch mirrors the quantitative results, underscoring that viewpoint altitude and in-domain fine-tuning are key to achieving accurate metric aerial depth estimation.

6. Conclusion

OccuFly introduces the first real-world aerial 3D SSC benchmark, comprising 9 scenes and over 20 k samples with RGB images, semantic voxel grids, and metric depth maps across 21 semantic classes. Our LiDAR-free, image-based data framework is highly scalable, requiring minimal manual annotation and fostering holistic aerial scene understanding.

Our data generation framework faces challenges that offer research opportunities: (1) It assumes static scenes, suppressing truly dynamic objects. Dynamic-capable reconstruction methods (*e.g.*, 4D Gaussian Splatting [68]) are a promising remedy. (2) Data acquisition may incur temporal inconsistencies across cross-altitude capture, when frustum culling uses images taken later. (3) The labeling process could be fully automated by replacing manually annotated masks with robust 2D pseudo-labels, further scaling data generation.

References

[1] Seyed Majid Azimi, Corentin Henry, Lars Sommer, Arne Schumann, and Eleonora Vig. Skyscapes fine-grained semantic understanding of aerial scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7393–7403, 2019.

[2] Jongseong Bae, Junwoo Ha, and Ha Young Kim. Three cars approaching within 100m! enhancing distant geometry by tri-axis voxel scanning for camera-based semantic scene completion, 2025.

[3] Radu Beche and Sergiu Nedevschi. Claravid: A holistic scene reconstruction benchmark from aerial perspective with delentropy-based complexity profiling, 2025.

[4] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019.

[5] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9297–9307, 2019.

[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020.

[7] Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation. *Journal of Visual Communication and Image Representation*, 109:104429, 2023.

[8] Anh-Quan Cao and Raoul de Charette. MonoScene: Monocular 3D semantic scene completion. In *CVPR*, 2022.

[9] Dubing Chen, Huan Zheng, Jin Fang, Xingping Dong, Xianfei Li, Wenlong Liao, Tao He, Pai Peng, and Jianbing Shen. Rethinking temporal fusion with a unified gradient descent view for 3d semantic occupancy prediction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1505–1515, 2025.

[10] Jiaqi Chen, Zeyu Yang, and Li Zhang. Semantic segment anything. https://github.com/fudan-zvg/Semantic-Segment-Anything, 2023.

[11] Lyujie Chen, Feng Liu, Yan Zhao, Wufan Wang, Xiaming Yuan, and Jihong Zhu. Valid: A comprehensive virtual aerial image dataset. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2009–2016, 2020.

[12] Yu Chen, Yao Wang, Peng Lu, Yisong Chen, and Guoping Wang. Large-scale structure from motion with semantic constraints of aerial images. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 347–359. Springer, 2018.

[13] OpenScene Contributors. Openscene: The largest up-to-date 3d occupancy prediction benchmark in autonomous driving, 2023.

[14] Oussema Dhaouadi, Johannes Meier, Luca Wahl, Jacques Kaiser, Luca Scalerandi, Nick Wandelburg, Zhuolun Zhuo, Nijanthan Berinpanathan, Holger Banzhaf, and Daniel Cremers. Highly accurate and diverse traffic data: The deepscenario open 3d dataset. In *2025 IEEE Intelligent Vehicles Symposium*. IEEE, 2025.

[15] DJI. Phantom 4 rtk. https://www.dji.com/phantom-4-rtk/info, 2016. Accessed: 2025-11-08.

[16] DJI. Mavic 3 enterprise series. https://enterprise.dji.com/mavic-3-enterprise, 2022. Accessed: 2025-11-08.

[17] DJI. Onboard SDK documentation. https://developer.dji.com/onboard-api-reference/group__telem.html, 2025. Accessed: 2025-11-08.

[18] Aloisio Dourado, Teofilo E. De Campos, Hansung Kim, and Adrian Hilton. Edgenet: Semantic scene completion from a single rgb- d image. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 503–510, 2021.

[19] Herbert Edelsbrunner, David Kirkpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559, 1983.

[20] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page 226–231. AAAI Press, 1996.

[21] European Commission. Commission implementing regulation (eu) 2019/947 of 24 may 2019 on the rules and procedures for the operation of unmanned aircraft, 2019. https://eur-lex.europa.eu/eli/reg_impl/2019/947/oj.

[22] Olivier Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, 1993. Section 2.3.2: Voxel Grids.

[23] Federal Aviation Administration. Small unmanned aircraft systems, 2016. https://www.ecfr.gov/current/title-14/chapter-I/subchapter-F/part-107.

[24] Horatiu Florea, Vlad-Cristian Miclea, and Sergiu Nedevschi. Wilduav: Monocular uav dataset for depth estimation tasks. *2021 IEEE 17th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2021.

[25] Michael Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2019.

[26] Markus Gerke and Heinz-Juergen Przybilla. Accuracy analysis of photogrammetric uav image blocks: Influence of onboard rtk-gnss and cross flight patterns. *Photogrammetrie - Fernerkundung - Geoinformation*, 2016(1):17–30, 2016.

[27] Markus Gross, Aya Fahmy, Danit Niwattananan, Dominik Muhle, Rui Song, Daniel Cremers, and Henri Meeß. IP-Former: Visual 3d panoptic scene completion with context-adaptive instance proposals. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[28] Xiyue Guo, Jiarui Hu, Junjie Hu, Hujun Bao, and Guofeng Zhang. Sgformer: Satellite-ground fusion for 3d semantic scene completion. *arXiv preprint arXiv:2503.16825*, 2025.

[29] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[30] Adrian Hayler, Felix Wimbauer, Dominik Muhle, Christian Rupprecht, and Daniel Cremers. S4c: Self-supervised semantic scene completion with neural fields. In *2024 International Conference on 3D Vision (3DV)*, pages 409–420. IEEE, 2024.

[31] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024.

[32] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9223–9232, 2023.

[33] Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. Gaussianformer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27477–27486, 2025.

[34] Institute of Computer Graphics and Vision, Graz University of Technology. Semantic drone dataset, 2019.

[35] Aleksandar Jevtić, Christoph Reich, Felix Wimbauer, Oliver Hahn, Christian Rupprecht, Stefan Roth, and Daniel Cremers. Feed-forward SceneDINO for unsupervised semantic scene completion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[36] Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries. *CVPR*, 2024.

[37] Arie Kaufman and Eyal Shimony. 3d scan-conversion algorithms for voxel-based graphics. In *Proceedings of the 1986 Workshop on Interactive 3D Graphics*, pages 45–75, Chapel Hill, NC, USA, 1986.

[38] Michael M. Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, pages 61–70, Aire-la-Ville, Switzerland, Switzerland, 2006. Eurographics Association.

[39] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision (3DV)*. IEEE, 2026.

[40] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[41] Benedikt Kolbeinsson and Krystian Mikolajczyk. DDOS: The drone depth and obstacle segmentation dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

[42] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, 1994.

[43] Hyo-Jun Lee, Yeong Jun Koh, Hanul Kim, Hyunseop Kim, Yonguk Lee, and Jinu Lee. Soap: Vision-centric 3d semantic scene completion with scene-adaptive decoder and occluded region-aware view projection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17145–17154, 2025.

[44] Heng Li, Yuenan Hou, Xiaohan Xing, Yuexin Ma, Xiao Sun, and Yanyong Zhang. Occmamba: Semantic occupancy prediction with state space models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 11949–11959, 2025.

[45] Shijie Li, Zhongyao Cheng, Rong Li, Shuai Li, Juergen Gall, Xun Xu, and Xulei Yang. Global-aware monocular semantic scene completion with state space models, 2025.

[46] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3205–3215, 2023.

[47] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9087–9098, 2023.

[48] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, Yue Wang, Hang Zhao, Zhiding Yu, and Chen Feng. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[49] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[50] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2023.

[51] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[52] Lin Liqiang, Liu Yilin, Hu Yue, Yan Xingguang, Xie Ke, and Huang Hui. Capturing, reconstructing, and simulating: the urbanscene3d dataset. In *ECCV*, pages 93–109, 2022.

[53] Enyu Liu, En Yu, Sijia Chen, and Wenbing Tao. Disentangling instance and scene contexts for 3d semantic scene completion, 2025.

[54] Xinhao Liu, Moonjun Gong, Qi Fang, Haoyu Xie, Yiming Li, Hang Zhao, and Chen Feng. Lidar-based 4d occupancy

completion and forecasting. *arXiv preprint arXiv:2310.11239*, 2023.

[55] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, 2022.

[56] Rafael Lopez-Campos and Jose Martinez-Carranza. Espada: Extended synthetic and photogrammetric aerial-image dataset. *IEEE Robotics and Automation Letters*, 6(4):7981–7988, 2021.

[57] Haoang Lu, Yuanqi Su, Xiaoning Zhang, Longjun Gao, Yu Xue, and Le Wang. Vishall3d: Monocular semantic scene completion from reconstructing the visible regions to hallucinating the invisible regions, 2025.

[58] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS journal of photogrammetry and remote sensing*, 165:108–119, 2020.

[59] Alina Marcu, Mihai Pirvu, Dragos Costea, Emanuela Haller, Emil Slusanschi, Ahmed Nabil Belbachir, Rahul Sukthankar, and Marius Leordeanu. Self-supervised hypergraphs for learning multiple world interpretations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 983–992, 2023.

[60] F. Nex, E. K. Stathopoulou, F. Remondino, M. Y. Yang, L. Madhuanand, Y. Yogender, B. Alsadik, M. Weinmann, B. Jutzi, and R. Qin. Usegeo - a uav-based multi-sensor dataset for geospatial research. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 13:100070, 2024.

[61] Ishan Nigam, Chen Huang, and Deva Ramanan. Ensemble knowledge transfer for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1499–1508. IEEE, 2018.

[62] Lucas Prado Osco, Qiusheng Wu, Eduardo Lopes de Lemos, Wesley Nunes Gonçalves, Ana Paula Marques Ramos, Jonathan Li, and José Marcato. The segment anything model (sam) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation*, 124:103540, 2023.

[63] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and {Robin Roberson} Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021. Publisher Copyright: © 2013 IEEE.

[64] Giulia Rizzoli, Francesco Barbato, Matteo Caligiuri, and Pietro Zanuttigh. Syndrone-multi-modal uav dataset for urban scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2210–2220, 2023.

[65] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[66] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[67] Rui Song, Chenwei Liang, Hu Cao, Zhiran Yan, Walter Zimmer, Markus Gross, Andreas Festag, and Alois Knoll. Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17996–18006, 2024.

[68] Rui Song, Chenwei Liang, Yan Xia, Walter Zimmer, Hu Cao, Holger Caesar, Andreas Festag, and Alois Knoll. Coda-4dgs: Dynamic gaussian splatting with context and deformation awareness for autonomous driving. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE/CVF, 2025.

[69] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 190–198, 2017.

[70] Simon Speth, Artur Gonçalves, Bastien Rigault, Satoshi Suzuki, Mondher Bouazizi, Yutaka Matsuo, and Helmut Prendinger. Deep learning with RGB and thermal images onboard a drone for monitoring operations. *Journal of Field Robotics*, 39(6):840–868, 2022.

[71] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[72] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic robotics*. MIT Press, Cambridge, Mass., 2005.

[73] Pengju Tian, Zhirui Wang, Peirui Cheng, Yuchao Wang, Zhechao Wang, Liangjin Zhao, Menglong Yan, Xue Yang, and Xian Sun. Ucdnet: Multi-uav collaborative 3-d object detection network by reliable feature mapping. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–16, 2025.

[74] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *arXiv preprint arXiv:2304.14365*, 2023.

[75] Khiem Vuong, Anurag Ghosh, Deva Ramanan, Srinivasa Narasimhan, and Shubham Tulsiani. Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[76] Jiahao Wang, Xiangyu Cao, Jiaru Zhong, Yuner Zhang, Haibao Yu, Lei He, and Shaobing Xu. Griffin: Aerial-ground cooperative detection and tracking dataset and benchmark, 2025.

[77] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[78] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. *arXiv preprint arXiv:2303.03991*, 2023.

[79] Xuzhi Wang, Xinran Wu, Song Wang, Lingdong Kong, and Ziping Zhao. Monocular semantic scene completion via masked recurrent networks. In *Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2025.

[80] Yuping Wang, Xiangyu Huang, Xiaokang Sun, Mingxuan Yan, Shuo Xing, Zhengzhong Tu, and Jiachen Li. Uniocc: A unified benchmark for occupancy forecasting and prediction in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2025.

[81] Z. Wang, P. Cheng, M. Chen, P. Tian, Z. Wang, X. Li, X. Yang, and X. Sun. Drones help drones: A collaborative framework for multi-drone object trajectory prediction and beyond. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[82] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. *arXiv preprint arXiv:2303.09551*, 2023.

[83] Rouwan Wu, Xiaoya Cheng, Juelin Zhu, Xuxiang Liu, Maojun Zhang, and Shen Yan. Uavd4l: A large-scale dataset for uav 6-dof localization. In *International Conference on 3D Vision (3DV)*, 2024.

[84] Xinyi Wu, Steven Landgraf, Markus Ulrich, and Rongjun Qin. An evaluation of dust3r/mast3r/vggt 3d reconstruction on photogrammetric aerial blocks. *Geo-spatial Information Science*, 0(0):1–19, 2025.

[85] Huaiyuan Xu, Junliang Chen, Shiyu Meng, Yi Wang, and Lap-Pui Chau. A survey on occupancy perception for autonomous driving: The information fusion perspective. *Information Fusion*, 114:102671, 2025.

[86] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589, 2022.

[87] Qi Yan, Jianhao Zheng, Simon Reding, Shanci Li, and Iordan Doytchinov. Crossloc: Scalable aerial localization assisted by multimodal synthetic data. *arXiv preprint arXiv:2112.09081*, 2021.

[88] Yuxiang Yan, Boda Liu, Jianfei Ai, Qinbu Li, Ru Wan, and Jian Pu. Pointssc: A cooperative vehicle-infrastructure point cloud benchmark for semantic scene completion. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 17027–17034. IEEE, 2024.

[89] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024.

[90] Hui Ye, Raj Sunderraman, and Shihao Ji. Uav3d: A large-scale 3d perception benchmark for unmanned aerial vehicles. In *The 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[91] Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-Liang Shen. Context and geometry aware voxel transformer for semantic scene completion. In *Advances in Neural Information Processing Systems*, 2024.

[92] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2304.05316*, 2023.

[93] Yanan Zhang, Jinqing Zhang, Zengran Wang, Junhao Xu, and Di Huang. Vision-based 3d occupancy prediction in autonomous driving: a review and outlook. *arXiv preprint arXiv:2405.02595*, 2024.

[94] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.