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

Supplementary Material

7. Scene-Level Visualizations

Fig. 6 illustrates (i) the reconstructed RGB point cloud, (ii) the semantic point cloud after label lifting, and (iii) the resulting scene-level semantic voxel grid. These visualizations demonstrate the remarkable fidelity of our data generation framework and its ability to propagate sparse 2D annotations to a globally consistent, voxel-level 3D ground-truth.

8. Implementation Details

Data Generation. We generate data on an AMD Ryzen Threadripper PRO 7985WX 64-Cores (allocating 8 cores) with 120 GB of memory.

For **3D reconstruction** (Sec. 3.3.1), we utilize the Agisoft Metashape 2.2.0 photogrammetric reconstruction software. After reconstruction, we ensure high geometric fidelity across the whole scene by removing a small, noisy margin at the border of the scene, which naturally arises from aerial SfM+MVS due to decreased image overlap (shown in Fig. 11). Subsequently, during **semantic annotation** (Sec. 3.3.2), we partition the ground plane into a regular grid of scene-dependent square cells with edge lengths of 23 to 28 meters to determine the subset of frames for manual annotation. Moreover, we apply kNN with $k = 100$ for unlabelled point assignment, and $k = 200$ for subsequent label refinement. Furthermore, for **densification and voxelization** (Sec. 3.3.3) of *instance classes*, DBSCAN [20] clustering is performed with class-wise parameters detailed in Tab. 6, and we set $\alpha = 0.05$ for α-Shapes [19] and use $K = 24$ camera views during silhouette extraction. For *ground classes*, we set Poisson reconstruction parameters [38] to a depth of 8 and a scale of 1.2. Finally, class-wise group assignments, semantic colors, and class frequencies are reported in Tab. 7.

Table 6. Class-wise DBSCAN [20] parameters for instance separation, discussed in Sec. 3.3.3.

Class	ϵ	MinPts
Building	4.0	1000
Roof	1.0	1000
Vehicle	1.0	500
Crane	1.0	500
Bicycle	0.4	80
Person	0.3	10
Truck	1.0	500

Table 7. Semantic class frequencies, group assignments (Sec. 3.3.3), and semantic color table of the OccuFly dataset.

Group	Color	Name	Frequency [%]
Instance		Building	62.1534
		Roof	2.2018
		Vehicle	0.5683
		Crane	0.0059
		Bicycle	0.0035
		Person	0.0001
		Truck	0.1105
Ground		Grass	8.5614
		Vegetation	4.3121
		Water	1.7539
		Walkway	2.0610
		Dirt	2.3364
		Road	1.8099
		Gravel	1.4511
		Parking Lot	2.8415
Others		Tree	7.5479
		Ground Obstacle	1.9605
		Construction	0.2741
		Cable Tower	0.0047
		Rock	0.0402
		Cable	0.0018

Aerial Semantic Scene Completion. We follow the training protocols of Symphonies [36] and DISC [53] using their official implementations from GitHub. We adapt the codebases to OccuFly's voxel grid resolution of $192 \times 128 \times 128$. All experiments are conducted on a single NVIDIA A100 80GB GPU with a batch size of 1.

Metric Monocular Depth Estimation. We employ Map-Anything-v1.1 [39], Metric3D-v2-ViT-L [31], Depth-Anything-v2-ViT-Small [89], and Depth-Anything-v3-Nested-Giant-Large [51] (as this is the only metric variant). For Depth-Anything-v2 specifically, we follow its metric adaptation protocol and fine-tune the affine-invariant model. Technically, we use official implementations for all methods and train on a single NVIDIA A100 GPU with 80 GB of memory.

9. Additional OccuFly Dataset Evaluation

3D Reconstruction. We provide scene-wise reprojection errors in Tab. 8. An average root mean square reprojection error of 1.24 pixels in our geo-referenced images validates the high metric accuracy of the reconstructed point cloud. Scene-wise reconstructed point clouds are shown in Fig. 6.

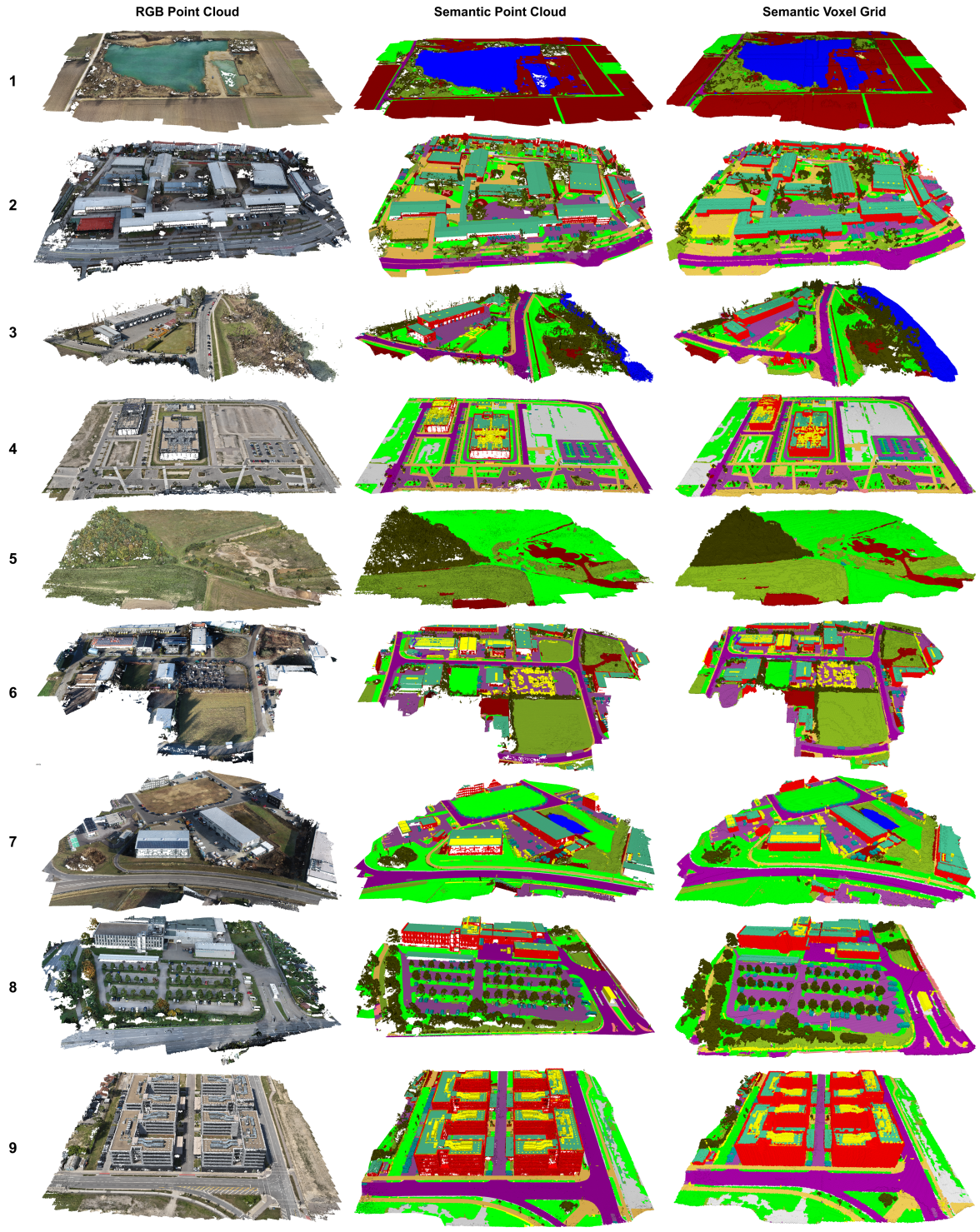

Figure 6. Scene-level outputs of our proposed data generation framework for all scenes 1-9 of the OccuFly dataset. **Left:** RGB pointcloud from 3D reconstruction (Sec. 3.3.1). **Center:** Semantic point cloud from semantic annoation (Sec. 3.3.2). **Right:** Semantic voxel grid from densification and voxelization (Sec. 3.3.3). Zoom in for best view.

Table 8. Scene-wise root mean square (RMS) reprojection error after 3D reconstruction (Sec. 3.3.1).

Scene	RMS Reprojection Error [px]
1	0.469
2	0.474
3	0.388
4	0.451
5	0.422
6	2.13
7	2.04
8	2.61
9	2.22
Average	1.24

Table 9. Scene-wise manual semantic annotation ratios for UAV platforms DJI Phantom 4 RTK (P4) [15] and DJI Mavic 3 Enterprise Series (M3-ES) [16]. Note that the number of acquired images marginally differs from the number of images finally provided in the dataset, as we remove images at the border of each reconstructed scene to ensure high geometric fidelity (see Sec. 8).

Scene	UAV Platform	Aquired Images	Annotated Images	Ratio [%]
1	P4	421	73	17.34
2	P4	338	48	14.20
3	M3-ES	1048	66	6.30
4	M3-ES	1252	102	8.15
5	M3-ES	1082	74	6.84
6	P4	380	52	13.68
7	P4	284	40	14.08
8	P4	251	38	15.14
9	M3-ES	1337	93	6.96
Total		**6393**	**586**	**9.17**

Semantic Annotation. As detailed in Sec. 3.3.2, we manually annotate only a small subset of images to subsequently lift semantic labels to 3D. This annotation costs approximately 29 minutes per image, which results in 1.3 days per scene to annotate >1.5 billion voxels in total. Due to our coverage-aware selection of annotated images, manual effort grows sublinearly with the number of 3D points. In Fig. 7, we present qualitative examples of the manual annotations, which exhibit pixel-accurate delineation. Furthermore, Tab. 9 reports per-scene annotation ratios, achieving an average annotation ratio of 9.17 %, indicating exceptional annotation efficiency. Additionally, Fig. 11 shows scene-wise image overlap during data collection, which exceeds >90 % for all scenes. This substantial overlap ensures accurate 3D reconstruction and semantic label lifting.

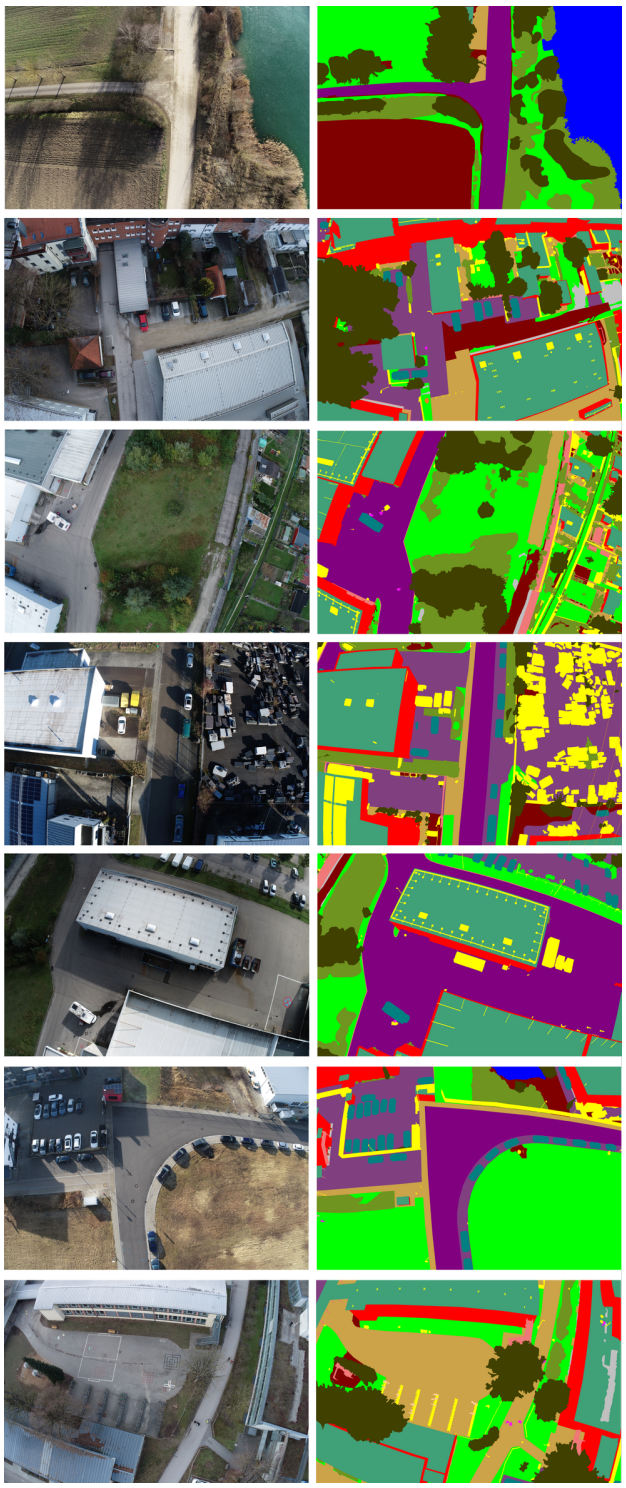

Figure 7. Examples of manual image annotations used for 3D label lifting, showing exceptional pixel-accurate delineation. Zoom in for best view.

Table 10. Semantic class comparison with 2D aerial image datasets.

	UDD [12]	VDD [7]	UAVid [58]	AeroScapes [61]	ICG [34]	SkyScapes [1]	OccuFly (ours)
# Classes	4	7	8	11	20	31 (20)	22

Metric Depth Maps. Table 8 reports a mean reprojection error of 1.24 pixels, indicating high geometric consistency of the reconstruction. Since the metric depth maps are derived from these reconstructed points, a low reprojection error serves as a strong proxy for depth accuracy. Moreover, we provide per-scene depth histograms for all nine scenes to illustrate the dataset's metric depth distributions (see Fig. 8). Most scenes show peaks at 30–50 meters, reflecting the image acquisition altitudes, while certain scenes, such as scene 9, exhibit a more diverse depth range.

Semantic Class Taxonomy. Beyond SSC, Tab. 10 positions OccuFly among established 2D aerial semantic segmentation datasets, where its 21-class taxonomy ranks second. While SkyScapes [1] ranks first, 12 of its 31 classes are lane-markings, effectively reducing its distinct class count to 20. Consequently, OccuFly provides one of the most detailed aerial taxonomies to date, strengthening fine-grained semantic evaluation and enabling seamless comparability with established 2D benchmarks.

Dataset Diversity. The class distributions shown in Fig. 9 highlight distinct semantic characteristics across environments, indicating that OccuFly provides rich domain diversity with distinct spatial layouts, architectural densities, and scale, by varying season, environment, altitude, and disjoint geographic locations across splits.

10. Additional Benchmark Evaluation

10.1. Aerial Semantic Scene Completion

In addition to the evaluation discussed in Secs. 5.2 and 5.3, we provide altitude-wise and class-wise metrics in Tab. 11, along with further qualitative results in Fig. 10. Consistent with the main manuscript, performance remains uniformly low across all altitude ranges, indicating that viewpoint height has a limited impact compared to the overall difficulty of the task. The class-wise analysis further reflects this trend, with only frequent classes being recovered to a limited extent, while rare classes are largely undetected. Qualitative results underscore these observations, showing coarse geometric structure but fragmented semantic predictions. Together, these findings reinforce the challenges of aerial SSC and underline the need for dedicated benchmarks such as OccuFly.

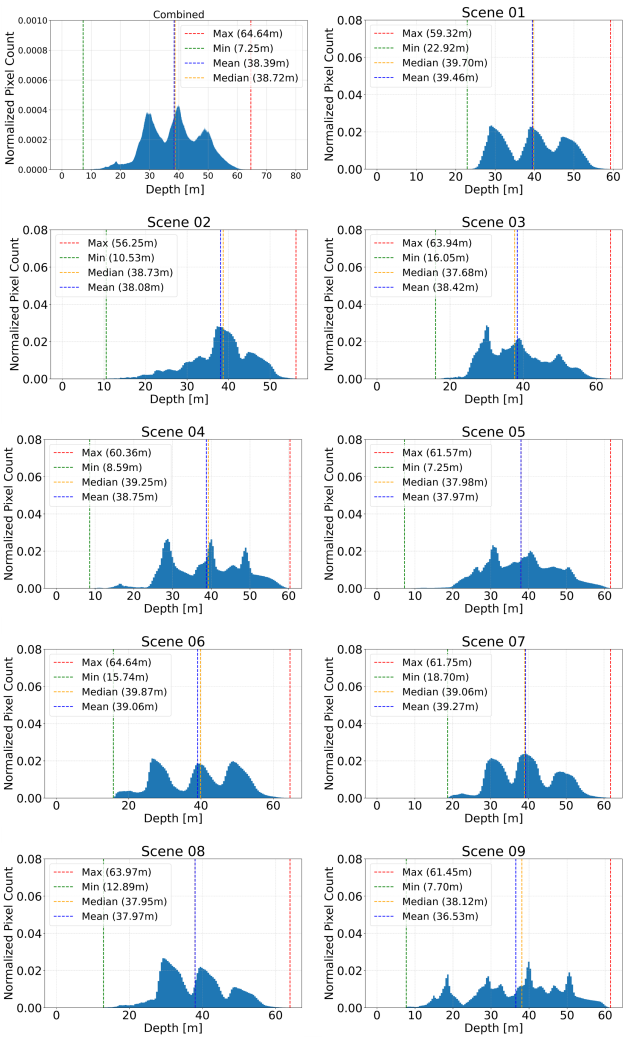

Figure 8. Depth map histograms for each of the 9 scenes in the OccuFly dataset. Zoom in for best view.

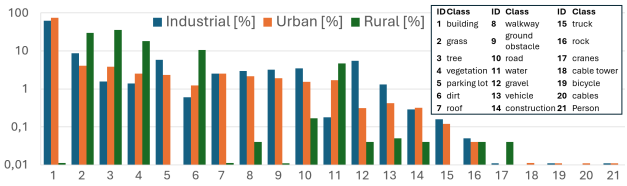

Figure 9. Environment-wise semantic class frequencies.

10.2. Aerial Metric Monocular Depth Estimation

In addition to Tab. 5, we provide a detailed analysis of depth estimation performance across different altitude ranges, presented in Tab. 12. The observed trends are consistent with the findings in the main manuscript: performance degrades with increasing altitude, highlighting the difficulty of metric depth estimation from higher viewpoints, while fine-tuning leads to substantial improvements across all altitudes.

Table 11. Altitude and class-wise SSC performance for Symphonies [36] and DISC [53] on the OccuFly test set in % (**best**).

Altitude [m]	Method	IoU	mIoU	Road (0.9377%)	Walkway (1.0560%)	Dirt (1.2102%)	Gravel (0.7960%)	Rock (0.0210%)	Grass (4.1978%)	Vegetation (2.3234%)	Tree (6.8357%)	Ground-Obs. (1.7566%)	Person (0.0002%)	Bicycle (0.0046%)	Vehicle (0.5195%)	Water (1.1322%)	Building (75.7457%)	Roof (1.7417%)	Cables (0.0017%)	Cable-Tower (0.0040%)	Parking-Lot (1.4349%)	Constructions (0.1689%)	Cranes (0.0052%)	Truck (0.1067%)
50	Symphonies	15.88	0.58	1.22	0.15	**0.19**	**0.54**	0.01	1.64	0.61	**0.72**	0.20	0.00	0.00	0.35	0.00	4.66	0.75	0.00	0.00	0.00	0.00	0.00	0.00
	DISC	**31.10**	**2.20**	**1.24**	**0.74**	0.12	0.24	0.01	**2.62**	**0.80**	0.23	**1.37**	0.00	**0.01**	**4.37**	0.00	**28.58**	**1.38**	0.00	0.00	0.00	0.00	0.00	0.00
40	Symphonies	10.71	0.52	0.40	0.09	0.05	0.02	0.01	1.59	1.08	**1.92**	0.09	0.00	0.00	0.68	0.00	3.38	0.57	0.00	0.00	0.00	0.00	0.00	0.00
	DISC	**27.85**	**1.77**	**0.87**	**0.25**	**0.18**	**0.13**	**0.02**	**1.61**	**1.54**	1.42	**0.28**	0.00	0.00	**4.80**	0.00	**21.46**	**1.12**	0.00	0.00	0.00	0.00	0.00	0.00
30	Symphonies	13.22	0.76	0.56	0.41	0.09	**0.25**	0.00	2.82	1.72	**2.32**	0.13	0.00	0.00	1.19	0.00	3.39	1.60	0.00	0.00	0.00	0.00	0.00	0.00
	DISC	**26.88**	**2.23**	**0.98**	**0.44**	**0.60**	0.10	**0.01**	**3.17**	**2.41**	0.78	**0.32**	0.00	0.00	**5.45**	0.00	**26.31**	**1.78**	0.00	0.00	0.00	0.00	0.00	0.00
all	Symphonies	13.68	0.58	0.88	0.14	0.13	**0.33**	0.01	1.76	0.94	**1.42**	0.15	0.00	0.00	0.51	0.00	4.08	0.74	0.00	0.00	0.00	0.00	0.00	0.00
	DISC	**29.52**	**2.04**	**1.08**	**0.53**	**0.17**	0.18	0.01	**2.32**	**1.33**	0.77	**0.90**	0.00	**0.01**	**4.67**	0.00	**25.48**	**1.30**	0.00	0.00	0.00	0.00	0.00	0.00

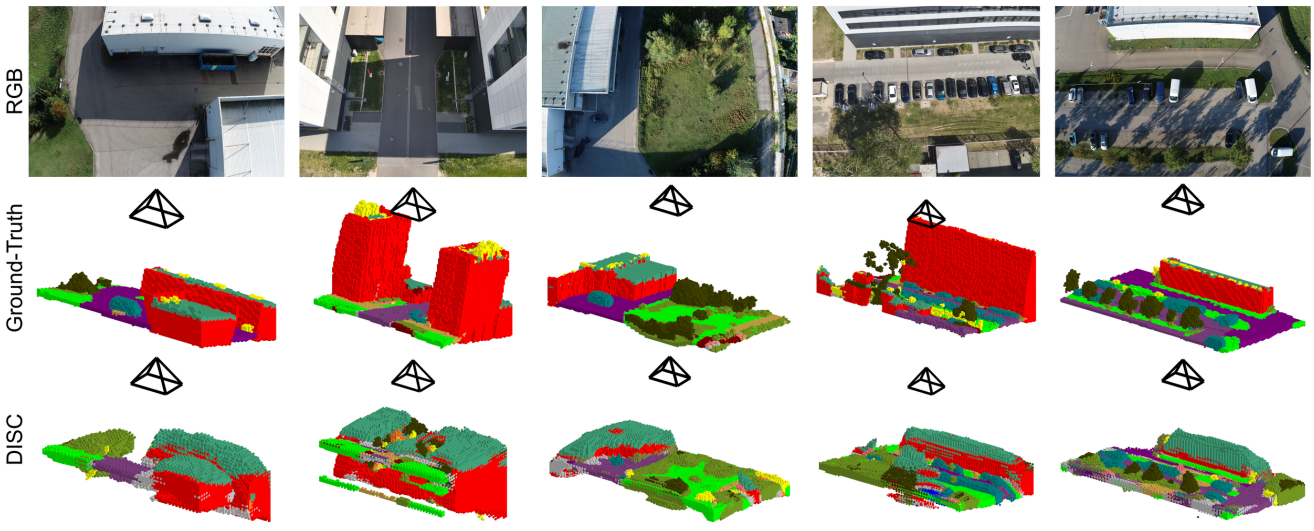

Figure 10. Additional SSC visualizations and qualitative evaluation of DISC [53] on the OccuFly test set.

Table 12. Altitude-wise metric monocular depth estimation performance, comparing zero-shot vs. fine-tuned foundation models on the OccuFly test set.

Altitude [m]	Method	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	AbsRel↓	RMSE↓	MAE↓	SILog↓
50	MapAnythingV1.1 [39]	0.000	0.000	0.000	0.836	36.783	35.935	0.057
	Metric3Dv2 [31]	0.014	0.161	0.363	0.516	24.582	23.250	0.166
	DepthAnything2 [89]	0.000	0.000	0.003	0.767	34.576	33.515	0.191
	DepthAnything3 [51]	0.000	0.000	0.000	0.673	29.703	29.017	**0.030**
	Metric3Dv2-OccuFly	0.283	0.831	0.987	0.379	15.373	14.950	0.099
	DepthAnything2-OccuFly	**0.844**	**0.997**	**1.000**	**0.129**	**5.985**	**5.261**	0.114
40	MapAnythingV1.1 [39]	0.000	0.001	0.005	0.780	26.735	25.906	0.083
	Metric3Dv2 [31]	0.137	0.277	0.485	0.437	17.102	15.808	0.165
	DepthAnything2 [89]	0.002	0.024	0.111	0.693	25.031	23.918	0.208
	DepthAnything3 [51]	0.000	0.001	0.073	0.564	19.450	18.817	**0.051**
	Metric3Dv2-OccuFly	0.209	0.802	0.981	0.395	13.138	12.625	0.096
	DepthAnything2-OccuFly	**0.795**	**0.957**	**0.998**	**0.148**	**4.486**	**3.823**	0.119
30	MapAnythingV1.1 [39]	0.000	0.000	0.002	0.764	23.307	22.952	0.060
	Metric3Dv2 [31]	0.036	0.129	0.589	0.459	14.518	14.084	0.107
	DepthAnything2 [89]	0.005	0.029	0.050	0.737	22.961	22.493	0.147
	DepthAnything3 [51]	0.002	0.026	0.669	0.471	14.494	14.216	**0.052**
	Metric3Dv2-OccuFly	0.460	0.757	0.992	0.347	10.917	10.215	0.101
	DepthAnything2-OccuFly	**0.919**	**0.982**	**0.998**	**0.103**	**3.108**	**2.666**	0.086
All	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

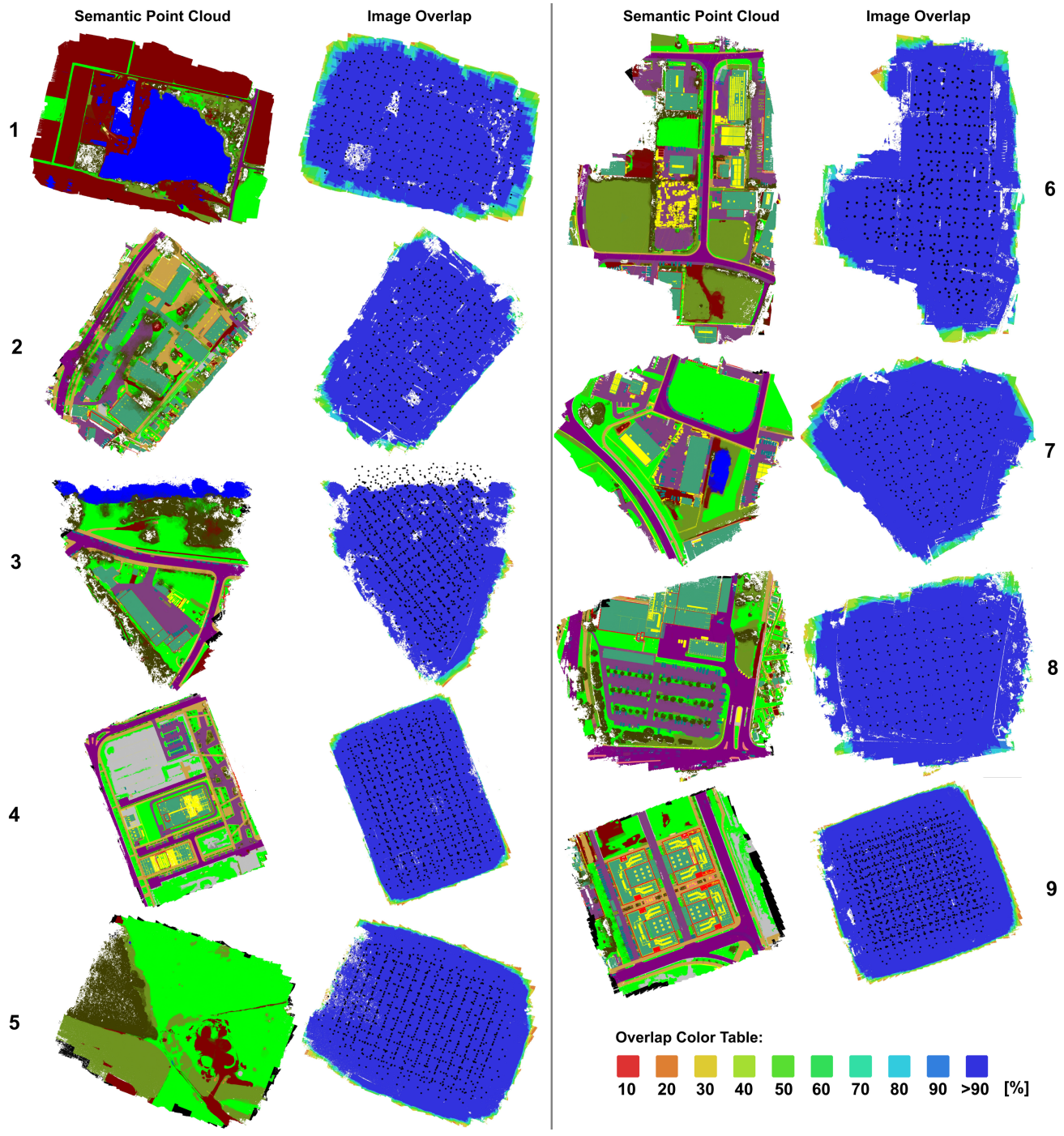

Figure 11. Scene-wise image overlap during data collection for all scenes 1-9 of the OccuFly dataset. **Left:** Top-down view of the semantic point cloud. **Right:** Image overlap with camera centers depicted as black dots. Note that we remove scene borders with $<90\,\%$ overlap to ensure geometric and semantic fidelity, as discussed in Sec. 8. Zoom in for best view.