

UAVScenes: A Multi-Modal Dataset for UAVs

Sijie Wang^{*1} Siqi Li^{*1} Yawei Zhang^{*1} Shangshu Yu^{*2} Shenghai Yuan^{*1} Rui She^{*3}
 Quanjiang Guo⁴ JinXuan Zheng¹ Ong Kang Howe¹ Leonrich Chandra¹ Shrivarshann Srijejan¹
 Aditya Sivasdas¹ Toshan Aggarwal¹ Heyuan Liu¹ Hongming Zhang¹ Chujie Chen¹ Junyu Jiang¹
 Lihua Xie¹ Wee Peng Tay¹

¹Nanyang Technological University

²School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China

³Beihang University

⁴University of Electronic Science and Technology of China

wang1679@e.ntu.edu.sg, shyuan@ntu.edu.sg

Abstract

Multi-modal perception is essential for unmanned aerial vehicle (UAV) operations, as it enables a comprehensive understanding of the UAVs' surrounding environment. However, most existing multi-modal UAV datasets are primarily biased toward localization and 3D reconstruction tasks, or only support map-level semantic segmentation due to the lack of frame-wise annotations for both camera images and LiDAR point clouds. This limitation prevents them from being used for high-level scene understanding tasks. To address this gap and advance multi-modal UAV perception, we introduce UAVScenes, a large-scale dataset designed to benchmark various tasks across both 2D and 3D modalities. Our benchmark dataset is built upon the well-calibrated multi-modal UAV dataset MARS-LVIG, originally developed only for simultaneous localization and mapping (SLAM). We enhance this dataset by providing manually labeled semantic annotations for both frame-wise images and LiDAR point clouds, along with accurate 6-degree-of-freedom (6-DoF) poses. These additions enable a wide range of UAV perception tasks, including segmentation, depth estimation, 6-DoF localization, place recognition, and novel view synthesis (NVS). Our dataset is available at <https://github.com/sijieaaa/UAVScenes>

1. Introduction

With the expansion of the low-altitude aerial economy [44, 54, 70, 83], UAVs have become indispensable for aerial taxi services [15, 17, 18], low-altitude logistics [8], agri-

^{*}The first six authors contribute equally: Sijie Wang, Siqi Li, Yawei Zhang, Shangshu Yu, Shenghai Yuan, and Rui She.

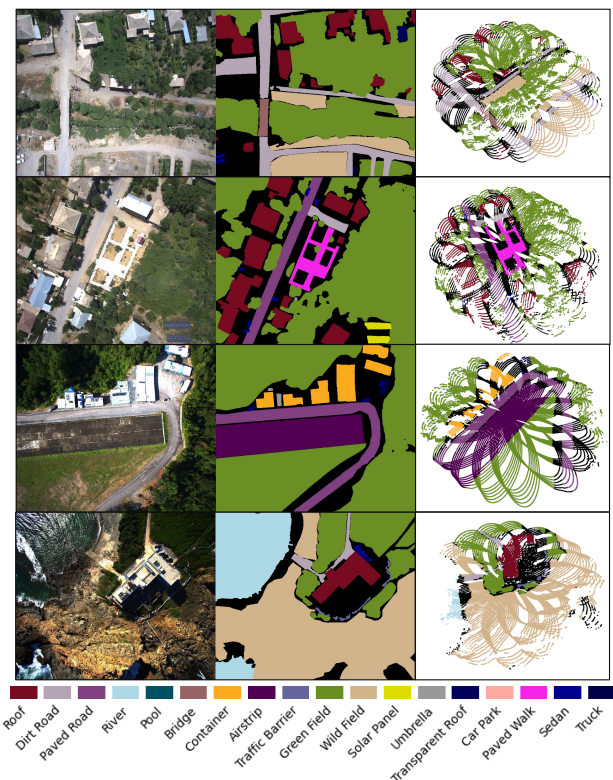


Figure 1. Visualization of frame-wise image and LiDAR point cloud annotations of the proposed UAVScenes dataset.

culture [69], inspection [16, 66, 67, 105], and emergency response [44]. Unlike ground vehicles, UAVs can operate above ground constraints, addressing limitations in current urban systems.

Yet, as UAVs tackle increasingly sophisticated tasks in diverse urban settings, they require training on high-quality datasets for reliable perception. Despite the potential of a

growing low-altitude economy, most existing datasets focus on single-modality camera data [11, 30, 39, 65, 75, 80, 104]. Although cameras provide rich texture information, they cannot capture the vital 3D spatial data required for more comprehensive and robust scene understanding.

Recent perception advances enable UAVs to carry lightweight 3D sensors, such as depth cameras, which work well for close-range tasks but remain limited in broader applications [2, 91, 106]. In contrast, 3D LiDAR offers long-range depth information and, when combined with cameras, provides multi-modal perception that substantially enhances UAV capabilities. While recent datasets [27, 56, 74, 87, 102, 103, 122] integrate cameras and LiDAR for richer 3D data, most focus on SLAM or 3D reconstruction and lack annotations for broader UAV tasks. Other multi-modal datasets [49, 61] only label 3D maps, limiting support for frame-wise tasks like image and LiDAR point cloud semantic segmentation.

In summary, despite advancements in UAV datasets, existing multi-modal UAV datasets either focus on SLAM [73, 112] and 3D reconstruction [25, 26] or only label 3D maps [49, 61]. None provides full semantic annotations for frame-wise geo-referenced images and LiDAR data. This gap limits their utility for real-time aerial scene understanding [110], navigation [31], and precision operations [8, 57].

To address this gap in UAV perception research, we present a large-scale annotated multi-modal benchmark dataset, **UAVScenes**. Built upon the MARS-LVIG dataset [56], originally designed for SLAM, UAVScenes includes semantic annotations for frame-wise camera images and LiDAR point clouds (see Fig. 1), along with accurate 6-degree-of-freedom (6-DoF) poses and reconstructed 3D maps. It supports a wide range of multi-modal UAV perception tasks, including semantic segmentation, depth estimation, localization, and novel view synthesis (NVS). Our main contributions are as follows:

- We present UAVScenes, a comprehensive multi-modal dataset for UAV perception that provides robust semantic scene understanding for both images and LiDAR point clouds, along with accurate 6-DoF poses and reconstructed 3D maps.
- Our dataset features over 120k frames with semantic annotations for images and LiDAR point clouds, surpassing the scale of most existing UAV research datasets.
- We conduct extensive benchmarking and evaluation of state-of-the-art (SOTA) methods on our dataset, establishing it as a wide-ranging UAV perception benchmark that supports at least six distinct tasks.

2. Related Work

In this section, we summarize existing autonomous driving and UAV datasets, discussing their sensor modalities, task coverage, and environmental constraints. We then highlight

their limitations in multi-modality and semantic labeling. By comparison, UAVScenes is designed to address these gaps by offering robust multi-modal coverage and frame-wise semantic annotations.

2.1. UAV Datasets

UAV perception datasets have become increasingly important due to the unique challenges posed by aerial perspectives. Over the years, various UAV datasets [30, 33, 39, 56, 61, 74, 86, 99, 103, 107] have been proposed, each contributing to different aspects of UAV perception.

Synthetic Datasets: Synthetic UAV datasets are primarily generated using simulation tools or platforms like Google Earth and CARLA [29]. Typical datasets like Mid-Air [33], TartanAir [97], University-1652 [118], and SynDrone [81] are used for various tasks. Mid-Air and TartanAir offer large-scale synthetic images and LiDAR-type data in unstructured environments. University-1652 features synthetic aerial images with satellites and ground views, providing the view when flying around the target. SynDrone offers semantic annotations for both synthetic LiDAR and camera. These datasets provide sufficient synthetic drone-view data for localization or scene understanding tasks.

Camera-Only Datasets: Beyond synthetic datasets, early real-world UAV datasets predominantly consist of visual camera-only modality due to the high cost of sensors and the relative immaturity of fusion technologies. These datasets provide visual imagery typically used for camera-based tasks. For example, some datasets [1, 14, 30, 32, 39, 65, 75, 80, 107, 120] include semantic or object annotations, supporting tasks such as semantic segmentation and object detection. Additionally, other datasets [22, 99, 104, 121] provide location data for each image, which can be used to benchmark localization and place recognition models.

However, these datasets lack the 3D LiDAR modality, limiting their application in 3D scene understanding and high-precision multi-modal fusion tasks.

Multi-Modal Datasets: With the advancement of sensor technology, an increasing number of multi-modal UAV datasets [27, 49, 56, 61, 74, 86, 87, 102, 103, 122] have emerged in recent years.

The H3D dataset [49] provides annotations on 3D maps reconstructed by LiDAR and camera data. However, it does not contain frame-wise annotations, which limits its applicability for frame-wise perception evaluation.

The Drone Vehicle dataset [86] enhances drone surveillance with labeled imagery for object detection and tracking. It also features infrared capabilities for visibility in low-light conditions. However, the absence of LiDAR restricts its use in 3D scene understanding and high-precision localization. NTU VIRAL [74] is a dataset designed for UAV SLAM and includes camera and LiDAR data. It enables research in tasks like place recognition, 3D map-

Dataset	Modality	LiDAR Type	6-DoF Pose	#Real Camera Frames with Frame-wise Anno. (type)	#Real LiDAR Frames with Frame-wise Anno.	Multiple Traversals	3D Map
Mid-Air [33]	Simulation	no real LiDAR	✓	no real camera	no real LiDAR	✓	✓ (with anno.)
TartanAir [97]	Simulation	no real LiDAR	✓	no real camera	no real LiDAR	✓	✓ (with anno.)
University-1652 [118]	Google Earth	no real LiDAR	-	no real camera	no real LiDAR	-	-
SynDrone [81]	Simulation	no real LiDAR	-	no real camera	no real LiDAR	-	-
UAVDT [30]	C	-	-	80k (bbox)	-	-	-
VisDrone [120]	C	-	-	40k (bbox)	-	-	-
CARPK [39]	C	-	-	1k (bbox)	-	-	-
Semantic Drone [1]	C	-	-	0.6k (mask)	-	-	-
Aerospaces [75]	C	-	-	3k (mask)	-	-	-
UAVid [65]	C	-	-	0.3k (mask)	-	-	-
FloodNet [80]	C	-	-	9k (mask)	-	-	-
CrossLoc [107]	C	-	✓	7k (mask)	-	✓	✓ (with anno.)
ALTO [22]	C	-	-	-	-	-	-
STPLS3D [19]	C	-	-	-	-	-	✓ (with anno.)
VDD [14]	C	-	-	0.4k (mask)	-	-	-
SUES-200 [121]	C	-	-	-	-	✓	-
UAV-VisLoc [104]	C	-	-	-	-	✓	-
HazyDet [32]	C	-	-	12k (bbox)	-	-	-
UAVD4L [99]	C	-	✓	-	-	-	✓
Hessigheim 3D [49]	C+L	RIEGL VUX-1LR	-	-	-	-	✓ (with anno.)
Drone Vehicle [86]	C+IR	-	-	57k (bbox)	-	-	-
NTU VIRAL [74]	C+L	2×3D-Ouster-16	✓	-	-	✓	✓
UrbanScene3D [61]	C+L	Trimble-X7	✓	-	-	✓	✓ (with anno.)
GrAco [122]	C+L	Velodyne-16	-	-	-	✓	-
GauU-Scene V2 [103]	C+L	DJI-L1*	✓	-	-	✓	✓
FIREStereo [27]	C+L	Velodyne-16	-	-	-	-	-
MUN-FRL [87]	C+L	Velodyne-16	✓	-	-	✓	-
MARS-LVIG [56]	C+L	DJI-L1* + Livox-Avia	-	-	-	✓	✓
UAVScenes (ours)	C+L	Livox-Avia	✓	120k (mask)	120k	✓	✓ (with anno.)

Table 1. Comparison of UAV datasets. Our dataset is the only one offering frame-wise annotations for both LiDAR and camera data on real scenes. We only count frame-wise annotations of real data, excluding synthetic or rendered data. “C” represents visible cameras, “L” represents LiDARs, and “IR” represents infrared cameras. “-” indicates that the dataset does not support this feature. “*” means that the DJI-L1 sensor produces encrypted point clouds, so per-frame LiDAR cannot be accessed. “bbox” denotes bounding boxes, and “mask” denotes semantic or instance masks.

ping, and localization. However, it is collected in indoor and small-scale outdoor environments, limiting its application for large-scale scene understanding. The UrbanScene3D dataset [61] provides high-resolution imagery and LiDAR data from an urban setting, offering capabilities for 3D scene segmentation [115] and localization for UAVs. However, it only offers semantic information on the reconstructed 3D map without considering frame-wise LiDAR point clouds, which prevents benchmarking frame-wise camera [58] and LiDAR [63] scene parsing [13]. GrAco [122] and FIREStereo [27] do not offer 6-DoF poses and focus on 3-DoF localization and stereo estimation [93], respectively. The GauU-Scene datasets [102, 103] collect UAV camera and LiDAR data in various urban environments and provide geo-aligned 3D maps. However, GauU-Scene uses DJI-L1 LiDAR, a closed-source sensor with encrypted point cloud data, hindering frame-wise LiDAR perception. MUM-FRL [87] is equipped only with a short-range LiDAR, resulting in substantial undetected point cloud data on the ground due to the high flying altitude. The MARS-LVIG [56] dataset stands out by providing multi-modal data across diverse scenarios, including multiple traversals through towns, valleys, airports, and islands. Additionally, it features a synchronized camera-LiDAR

suite, ensuring well-aligned images and point clouds.

However, these existing multi-modal UAV datasets lack frame-wise annotations for both images and LiDAR point clouds, limiting their utility for benchmarking advanced multi-modal perception tasks.

The UAVScenes dataset aims to fill this gap by providing comprehensive semantic annotations for both frame-wise images and LiDAR data. Additionally, it includes accurate 6-DoF poses and reconstructed point cloud maps, enabling a wide range of tasks such as segmentation, depth estimation, 6-DoF localization, place recognition, and NVS.

As shown in Tab. 1, UAVScenes is the only dataset that simultaneously offers 6-DoF poses as well as frame-wise image and LiDAR point cloud annotations. By providing both image and LiDAR annotations with precise pose alignment, UAVScenes has the potential to significantly advance research in multi-modal UAV perception.

2.2. Other Annotated Multi-Modal Datasets

Besides UAV research, there are also annotated multi-modal datasets in other domains. In autonomous driving, widely used examples include KITTI [36], KITTI-360 [59], nuScenes [12], Waymo-Open [85], and K-Radar [78]. In robotics, popular multi-modal datasets include Wild-

Scenes [89], 2D-3D-Semantic [6], RELIS-3D [42], EnviroDat [76], and Great Outdoors [43]. For indoor settings, EmbodiedScan [96], ScanNet [24], ScanNet++ [111], and NYU Depth v2 [84] are widely used. However, these datasets are not collected from UAV perspectives, limiting their suitability for evaluating and benchmarking various UAV tasks. To bridge this gap, UAVScenes is designed to provide a comprehensive benchmark tailored for UAV-based research.

3. The UAVScenes Dataset

UAVScenes builds on the MARS-LVIG [56] dataset. Among existing multi-modal UAV datasets [27, 87, 103, 122], MARS-LVIG stands out for its extensive sequential data gathered in diverse, large-scale environments—towns, valleys, airports, and islands—all traversed multiple times. This makes it ideal for benchmarking a variety of perception tasks and the most suitable foundation for our new dataset.

Although MARS-LVIG provides rich data, it primarily targets SLAM, offering only 4-DoF poses and reconstructed 3D point-cloud maps. These constraints limit its applicability to tasks requiring semantic annotations and 6-DoF poses.

To address these gaps, we extend MARS-LVIG with comprehensive camera and LiDAR semantic annotations and reconstruct 6-DoF poses with aligned 3D maps. Additionally, we benchmark six tasks and compare leading SOTA methods.

3.1. Choices for 3D Reconstruction

The MARS-LVIG dataset provides only 4-DoF poses using RTK, which includes a 3-DoF location and a yaw angle. As a result, MARS-LVIG is suitable solely for 4-DoF localization benchmarking as it lacks the necessary 6-DoF poses required for evaluating more fine-grained localization and reconstruction tasks, such as 6-DoF localization and NVS.

Initially, we attempt to use SOTA open-source LiDAR-visual-inertial (LVI) SLAM methods, including FAST-LIVO [117] and R3LIVE [60]. However, ground-facing flight causes LiDAR degeneration [56], leading to unsatisfactory reconstruction results (e.g., missing too many poses, producing distorted 3D maps, and failing in reconstruction).

As an alternative, we use structure-from-motion (SfM) solutions to reconstruct the 6-DoF poses along with the corresponding 3D maps. We have tried various SfM solutions, including COLMAP [82], RealityCapture¹, Metashape², and DJI Terra³. Among them, Terra, which can accept global navigation satellite system (GNSS) coordinates as the pose initializations and is specially designed for UAV scenes, provides relatively better reconstruction results. As shown in Figs. 2 and 3, the rendered image aligns well with

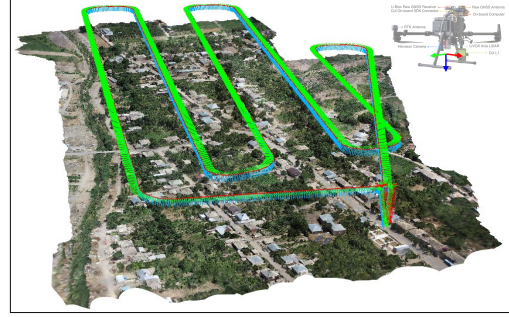


Figure 2. Reconstructed 3D maps and 6-DoF poses using Terra. Poses are downsampled for better visualization.

the real captured images using the reconstructed 3D maps and 6-DoF poses.

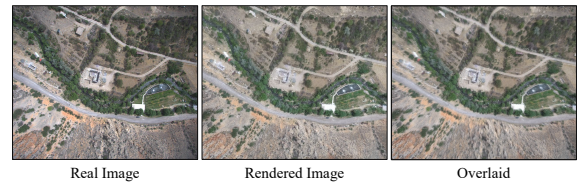


Figure 3. Visualization of the image rendered from the reconstructed 3D map and the 6-DoF camera pose. The rendered image closely aligns with the original image when overlaid.

3.2. Frame-Wise Image Semantic Annotations

3.2.1. Static Class Annotations.

The MARS-LVIG dataset consists of multiple sensor data sequences. We need to ensure that annotations are consistent across consecutive frames. Following the annotation methodology used in SemanticKITTI [7], we divide the entire MARS-LVIG dataset into 8 distinct splits based on environmental and illumination conditions. Each split contains 1-3 sequences collected continuously within the same day, ensuring minimal scene changes except for dynamic objects. Details can be found in the supplement.

We apply Terra SfM to each split, resulting in 8 reconstructed 3D maps and their corresponding poses. The reconstruction for each split usually takes 3-10 hours⁴. For each 3D map, we conduct manually annotating for 16 static scene classes. These annotated 3D maps are then rendered onto the corresponding camera views to produce annotated 2D semantic masks as shown in Fig. 4.

To ensure the quality of the rendered semantic annotations, we manually check for consistency and correct any unsatisfactory annotations. This process ensures that the image semantic annotations are both sequentially consistent and of high quality.

3.2.2. Dynamic Class Annotations.

Since the rendered static scene masks do not account for dynamic objects, we manually annotate instance-wise la-

¹<https://www.capturingreality.com/>

²<https://www.agisoft.com/>

³<https://enterprise.dji.com/dji-terra>

⁴Hardware: i9-13900K + RTX 4090*2.

bels for 2 dynamic object classes (sedan and truck) in each image as shown in Fig. 4. As MARS-LVIG is sequentially captured, manual annotating can be partially accelerated by auto-labeling tracking (tracking is always unstable), followed by human verification and fixing. We use X-AnyLabeling⁵ to achieve tracking. In total, we have manually annotated over 280k dynamic instances in the dataset (see the statistics in Tab. 2). The 2D static semantic annotations and 2D dynamic instance annotations are then combined to produce the final 2D annotations for each image.

This annotation process results in 120k annotated images with 19 classes, including 16 static classes, 2 dynamic classes, and 1 background class. The overall class distribution is shown in Fig. 5.

	Sedan	Truck
Avg. BBox Height (pixel)	72	106
Avg. BBox Width (pixel)	68	106
Avg. BBox Area (pixel)	5001	12411
Avg. Polygon Area (pixel)	3210	6873
Avg. Occupancy Ratio	67%	62%
#Instances	270k	14k

Table 2. Image instance statistics for dynamic object classes.

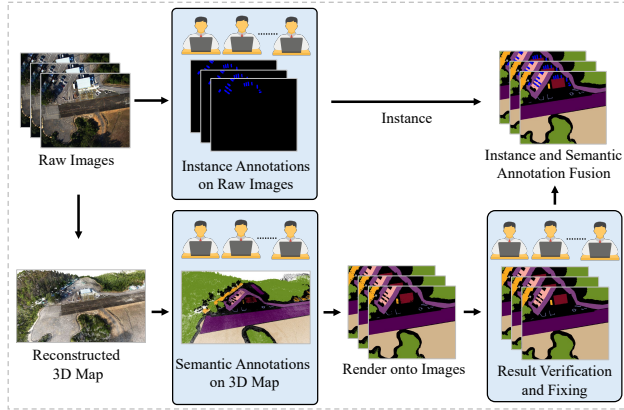


Figure 4. The 2D image annotating pipeline. Manual annotating is conducted at 3D map annotations, instance annotations, and fixing stages.

3.3. Frame-Wise LiDAR Point Cloud Annotations

The LiDAR is a crucial sensor for multi-modal perception. It casts laser beams to capture the spatial information⁶ of the environment.

In the MARS-LVIG dataset, there are two distinct LiDAR sensors: the DJI-L1 and the Livox-Avia. Due to manufacturer-imposed encryption on the DJI-L1’s output data, access to its raw point clouds is restricted, limiting its utility for open research. Therefore, our dataset annotations are focused exclusively on data captured by the open-source Livox-Avia LiDAR, which enables unrestricted ac-

⁵<https://github.com/CVHub520/X-AnyLabeling>

⁶Some LiDARs also provide other information, e.g., intensity.

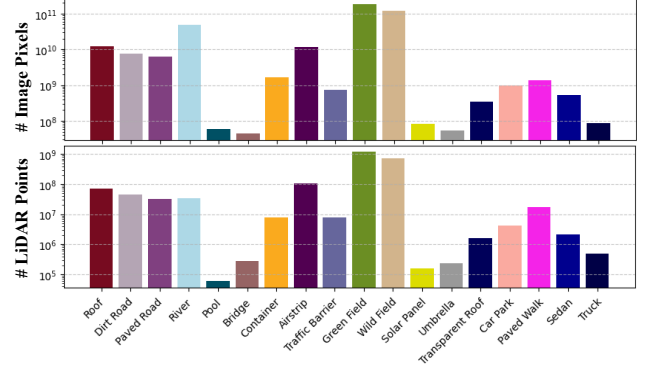


Figure 5. The annotation class distribution visualization. The background class is ignored. Above: image pixel annotations. Below: LiDAR point cloud annotations.

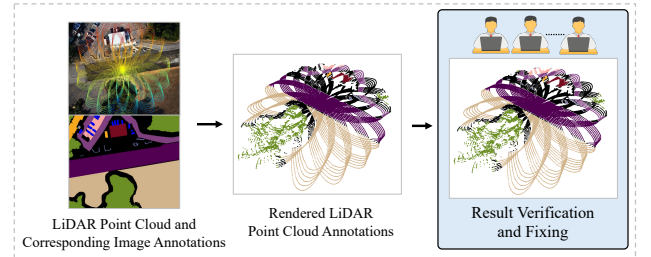


Figure 6. The LiDAR point cloud annotating pipeline. Manual annotating is conducted at fixing stages.

cess to high-quality point cloud data, facilitating broader applicability and reproducibility in academic and industrial research.

The MARS-LVIG dataset is acquired using a hardware-synchronized camera-LiDAR suite, which incorporates a calibrated camera and LiDAR sensors. By leveraging this calibration, we project image annotations onto the corresponding LiDAR point clouds. Following the procedure described in Sec. 3.2, we conduct thorough consistency checks, manually correcting any unsatisfactory annotations within the LiDAR point clouds to ensure high fidelity between each camera-LiDAR frame pair. This workflow is illustrated in Fig. 6. The class distribution for LiDAR point cloud annotations is shown in Fig. 5.

4. Benchmark Experiments

In this section, we establish benchmarks on various perception tasks using the proposed UAVScenes dataset. The existing benchmark tasks include frame-wise image and LiDAR semantic segmentation, place recognition, depth estimation, 6-DoF localization, and NVS.

4.1. Image Semantic Segmentation

Image semantic segmentation is a fundamental task in computer vision and is essential for evaluating the performance of vision models. It involves predicting the class label for each pixel in an input image. We consider several backbone

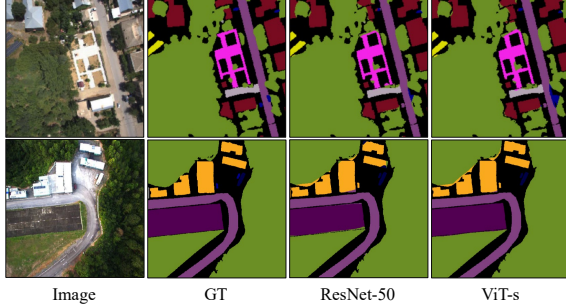


Figure 7. Visualization of Image semantic segmentation results.

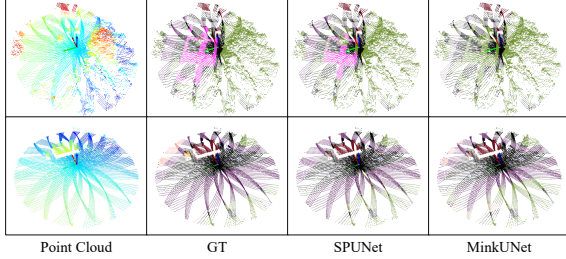


Figure 8. Visualization of LiDAR semantic segmentation results.

architectures, including ResNet [38], ConvNext [62], ConvNextV2 [98], ViT [28], MambaOut [114], and DeiT3 [88]. We use UperNet [101] as the segmentation head, which is widely used in semantic segmentation evaluations. All models are based on the TIMM package⁷.

As shown in Tab. 3, among these backbones, DeiT3 [28] achieves the best performance. Moreover, the Transformer-based models generally outperform convolutional neural networks (CNNs), which demonstrate the effectiveness of Transformers. We visualize some examples in Fig. 7.

4.2. Frame-wise LiDAR Semantic Segmentation

Frame-wise LiDAR point cloud semantic segmentation is a crucial task for 3D scene understanding, involving the prediction of class labels for each point in the LiDAR-generated point cloud. For this task, we run and evaluate three baseline models: MinkUNet [21], SPUNet [23], and PTv2 [100]. All models are based on the Pointcept package⁸.

As shown in Tab. 3, the three networks show comparable performance. PTv2 with the least number of parameters (11M) can surpass MinkUNet (38M), demonstrating the network architecture effectiveness. In addition, the pool class would be a challenging class for LiDAR semantic segmentation as all three networks show 0 class IoU. This would be attributed to the low quantity of annotated LiDAR point clouds as in Fig. 5. We visualize some LiDAR semantic segmentation examples in Fig. 8.

⁷<https://github.com/huggingface/pytorch-image-models>

⁸<https://github.com/Pointcept/Pointcept>

4.3. Place Recognition

Place recognition treats localization as a retrieval problem [5]. In this approach, the place is recognized by matching a query image to a database of images, with the location of the top-matched database image being regarded as the query’s location. In this task, we compare camera-based, LiDAR-based, and fusion-based place recognition methods. The image-based models include GeM [79], RRM [53], ConvAP [4], MixVPR [3], AnyLoc [46], and SALAD [41]. The LiDAR-based models include MinkLoc3D [50], MinkLoc3D V2 [51], and BEVPlace [64]. The fusion-based models include MinkLoc++ [52], AdaFusion [55], LCPR [119], and UMF [35].

In Tab. 4, we compare the recall performance of different models. Generally, fusion-based place recognition models outperform their single-modal counterparts, demonstrating the effectiveness of multi-modal fusion. For camera-based place recognition models, the inclusion of strong foundation backbones like DINO V2 [77] brings significant improvements. Additionally, the projection model BEVPlace performs worse than the voxel model MinkLoc3D, indicating that BEV LiDAR projection may not be a suitable format for place recognition under UAV perspectives.

4.4. Novel View Synthesis

NVS generates new perspectives of a scene from limited image viewpoints, paving the way for realistic and efficient 3D scene generation that captures intricate lighting, textures, and geometric details. NVS is largely driven by neural radiance fields (NeRFs) [68], which model 3D scenes as continuous functions with differentiable rendering, and 3D Gaussians (GS) [48], which represent scenes as learnable 3D Gaussians for rasterized rendering. To evaluate NVS, we introduce NeRFs-based and 3D GS-based baselines: Instant-NGP [71], 3DGS [48], GaussianPro [20], DCGaussian [94], and Pixel-GS [116].

The quantitative results and qualitative visualizations on UAVScenes are presented in Tab. 5 and Fig. 9, respectively. All 3D GS-based methods use the raw point cloud provided by the dataset as the initialization. The NeRF method Instant-NGP performs poorly on large-scale aerial images. The 3D GS methods, 3DGS and Pixel-GS, achieve better rendering performance than others. However, in certain areas, such as adjacent buildings and repetitive forest scenes, the performance of NVS still requires improvement, as highlighted by red boxes.

4.5. 6-DoF Visual Localization

6-DoF visual localization is a fundamental task in computer vision, essential for applications like robotics and augmented reality. Its goal is to estimate the 6-DoF pose of a query image within a pre-existing environment map. Currently, absolute pose regression (APR) and scene co-



















#Params	Arch.	Model	mIoU↑	Per Class IoU↑																	
				Roof	Dirt Road	Paved Road	River	Pool	Bridge	Conta.	Airstrip	Traffic Barrier	Green Field	Wild Field	Solar Panel	Umbre.	Transp. Roof	Car Park	Paved Walk	Sedan	Truck
																					
Image Semantic Segmentation																					
21M	CNN	ResNet-34 [38]	59.9	74.3	53.9	77.4	91.6	25.4	21.4	69.9	88.1	53.8	89.8	87.4	74.7	2.1	64.2	52.0	89.1	22.3	41.5
25M	CNN	ResNet-50 [38]	61.3	76.0	52.6	77.8	88.4	19.3	30.9	69.3	91.9	49.4	90.4	88.6	77.9	8.5	65.8	51.9	94.5	20.1	49.8
44M	CNN	ResNet-101 [38]	60.7	77.0	53.9	78.1	75.9	29.2	33.8	70.3	92.6	54.0	91.0	80.9	80.4	8.3	66.2	50.3	91.1	16.5	43.3
28M	CNN	ConvNext-t [62]	55.3	72.4	46.4	71.1	84.4	18.1	20.7	64.2	82.8	46.2	91.2	84.4	79.1	1.1	61.4	39.9	81.1	10.5	39.8
28M	CNN	ConvNextV2-t [98]	53.1	70.8	44.6	67.4	88.0	21.0	18.0	55.9	80.1	43.6	90.7	86.6	72.3	5.8	56.9	27.6	80.0	7.8	38.5
48M	CNN	MambaOut-s [114]	51.8	65.2	46.6	69.9	56.5	25.4	19.0	58.1	78.0	36.5	82.3	82.7	76.2	2.6	57.8	44.1	79.9	12.0	39.2
26M	CNN	MambaOut-t [114]	50.0	59.0	43.1	63.5	65.6	19.1	20.0	55.9	74.0	34.4	80.0	81.0	76.1	1.3	57.8	40.5	80.0	12.0	37.1
5M	Transf.	ViT-t [28]	62.8	74.3	58.8	76.7	90.9	41.2	52.8	51.3	80.4	39.4	93.6	90.3	88.9	19.3	76.4	62.3	86.4	26.5	20.3
22M	Transf.	ViT-s [28]	63.9	75.0	61.2	77.4	88.7	49.0	54.9	56.5	86.5	51.4	94.3	90.0	89.5	11.2	80.5	52.4	89.7	20.2	21.9
22M	Transf.	DeiT3-s [88]	67.6	76.0	67.0	81.1	91.0	58.1	57.8	62.7	88.0	41.6	91.1	91.7	90.0	24.1	82.9	63.2	93.0	28.1	30.0
38M	Transf.	DeiT3-m [88]	68.3	77.6	66.2	79.3	92.2	52.3	56.6	58.9	88.6	53.2	93.6	92.4	90.1	30.9	83.5	60.4	93.7	27.3	32.5
Frame-wise LiDAR Semantic Segmentation																					
38M	-	MinkUNet [21]	32.7	74.5	43.4	57.6	61.3	0.0	10.3	14.4	47.3	32.3	86.2	81.8	2.3	1.4	31.1	9.9	18.3	13.4	3.1
39M	-	SPUNet [23]	34.4	73.9	38.1	56.3	37.0	0.0	15.1	38.5	65.4	38.8	85.7	78.1	0.0	0.0	23.0	8.4	47.2	13.0	0.0
11M	-	PTv2 [100]	33.2	71.7	38.4	32.7	38.2	0.0	8.6	47.9	34.2	50.1	75.1	55.0	3.0	53.8	41.3	2.0	0.1	27.1	18.4

Table 3. Semantic segmentation results with mIoU (%) and class IoU (%). Above: Camera-Based. Below: LiDAR-Based.

Model	Modality	Recall@1↑	Recall@5↑	Recall@10↑
GeM [79]	C	42.1	55.8	62.0
RRM [53]	C	41.7	53.4	61.1
ConvAP [4]	C	41.1	54.8	63.0
MixVPR [3]	C	34.0	53.0	61.6
AnyLoc [46] (DINO V2-s [77])	C	58.5	74.4	79.1
SALAD [41] (DINO V2-s [77])	C	67.1	76.4	79.8
MinkLoc3D [50]	L	41.9	60.0	66.7
MinkLoc3D V2 [51]	L	42.8	61.5	67.3
BEVPlace [64]	L	32.6	54.6	64.2
MinkLoc++ [52]	C+L	47.1	63.5	69.0
AdaFusion [55]	C+L	46.3	63.4	70.2
LCPR [119]	C+L	42.3	62.3	68.8
UMF [35]	C+L	40.1	53.9	61.0

Table 4. Place recognition performance with Recall@K (K = 1, 5, 10) (%). AnyLoc and SALAD use the visual foundation backbone DINO V2[77], while other models use ResNet-18 as the 2D backbone.

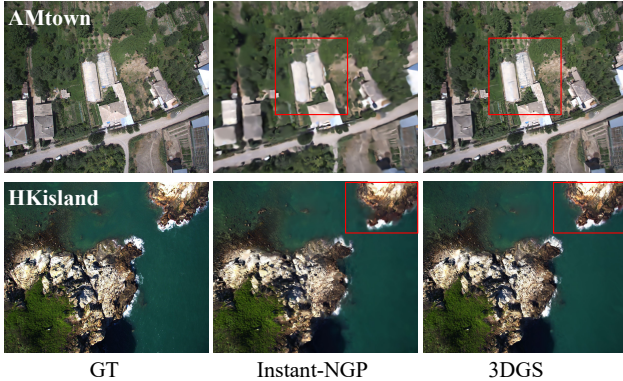


Figure 9. Qualitative evaluation of NVS. The areas outlined in red highlight regions with significant rendering discrepancies.

ordinate regression (SCR) have made significant strides in localization. APR estimates the 6-DoF pose of an input image through direct regression, enabling an end-to-end, highly efficient localization process. SCR localizes by regressing the 3D coordinates of 2D image pixels rather than

directly estimating the camera pose, enabling training via re-projection error. The camera pose is then determined through 2D-3D correspondences. This paper conducts experiments using modern APR baselines, including PoseNet [47], AtLoc [90], and RobustLoc [95], as well as SCR baselines such as ACE [10], GLACE [92], and FocusTune [72].

The localization errors (position and rotation) and qualitative visualizations on UAVScenes are presented in Tab. 6 and Fig. 10, respectively. All APR-based methods demonstrate strong performance, with RobustLoc achieving the best results, significantly outperforming others. SCR-based methods use a frozen pre-trained encoder for faster training, but its ground-urban-scene pretraining limits performance on UAV-view images, leading to higher localization errors.

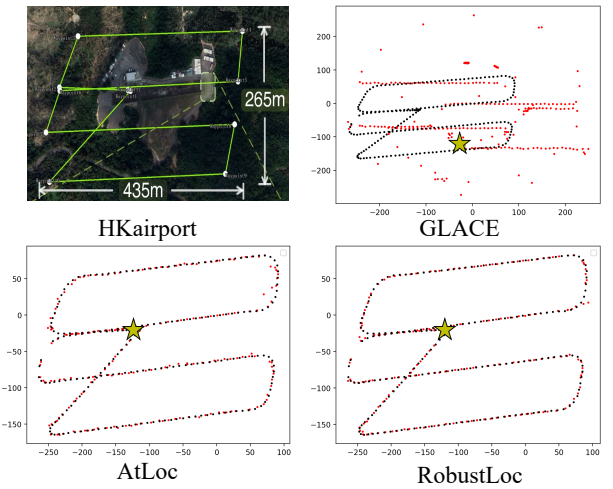


Figure 10. Visualization of 6-DoF localization. The ground truth and prediction are black and red lines, respectively. The star denotes the first frame. Metrics are in meters.

Model	AMtown			AMvalley			HKairport			HKisland		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Instant-NGP [71]	20.09	0.451	0.674	23.20	0.497	0.622	19.39	0.442	0.604	16.81	0.508	0.556
3DGS [48]	22.95	0.547	0.551	25.12	0.576	0.514	20.92	0.523	0.481	17.85	0.553	0.494
GaussianPro [20]	22.83	0.546	0.549	25.09	0.576	0.511	20.86	0.521	0.478	17.82	0.552	0.491
DCGaussian [94]	22.92	0.537	0.556	25.10	0.573	0.516	20.81	0.510	0.495	17.77	0.546	0.503
Pixel-GS [116]	22.95	0.547	0.551	25.11	0.576	0.514	20.93	0.523	0.481	17.84	0.553	0.494

Table 5. Quantitative evaluation of NVS. The evaluated metrics include PSNR, SSIM, and LPIPS.

Model	AMtown	AMvalley	HKairport	HKisland	Average
ACE [10]	180.8 / 1.5	145.8 / 0.6	83.0 / 0.6	121.9 / 0.8	132.9 / 0.9
FocusTune [72]	188.9 / 1.0	156.0 / 0.6	72.1 / 0.6	113.7 / 0.9	132.7 / 0.8
GLACE [92]	90.9 / 0.6	90.1 / 0.4	58.3 / 0.5	102.8 / 0.7	85.5 / 0.6
PoseNet [47]	43.1 / 0.2	22.7 / 0.2	14.9 / 0.2	22.5 / 0.1	25.8 / 0.2
AtLoc [90]	12.7 / 0.2	9.0 / 0.1	6.0 / 0.1	6.7 / 0.1	8.6 / 0.1
RobustLoc [95]	5.9 / 0.1	9.8 / 0.1	4.9 / 0.1	3.6 / 0.1	6.1 / 0.1

Table 6. Quantitative evaluation of visual localization on the UAVScenes dataset. We report median position error (m) and median rotation error (degree).

4.6. Depth Estimation

Depth estimation involves predicting pixel-wise depth values from input images, bridging the gap between 2D imagery and 3D spatial understanding. This task is particularly valuable for evaluating camera-only perception systems that require real-time or lightweight operation (e.g., lightweight cameras-only UAVs). Since the MARS-LVIG dataset does not provide such evaluation (though its calibrated camera-LiDAR suite can support), we add this task to create a more comprehensive benchmark. In this section, we evaluate zero-shot depth estimation models to assess their generalization capabilities in UAV aerial views. We consider both single-step models and diffusion-based multi-step models. The single-step models include ZoeDepth [9], Depth Anything [9], Depth Anything V2 [109], Metric3D [113], and Metric3D V2 [40]. The diffusion models include GeoWizard [34] and Marigold [45]. The ground truth depth for each image is derived from the corresponding LiDAR frame.

As shown in Tab. 7, Metric3D V2 demonstrates the best performance in terms of absolute relative error and square relative error. However, Depth Anything V2 outperforms it in the δ_1 metric. For diffusion-based models, which only support affine-invariant depth maps, the performance is relatively worse compared to their single-step counterparts. We visualize the depth predictions of the single-step models in Fig. 11. Most zero-shot monocular depth estimation schemes lack generalization ability and accuracy in the UAV perspective, underscoring the need for advancements in this area.

5. Limitation and Conclusion

Although UAVScenes has captured large-scale environments, expanding its diversity remains crucial. Future efforts could include complex urban or downtown areas with varied streets, vehicles, high-rise buildings, and pedestrians.

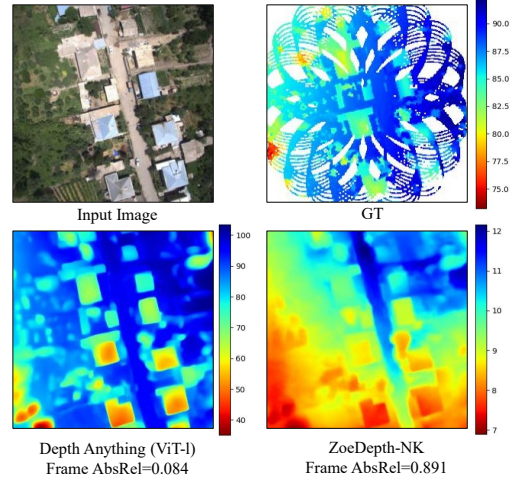


Figure 11. Visualization of the zero-shot depth estimation results. The depth ground truth is from the corresponding LiDAR point cloud. The color bar indicates depth values.

Model	AbsRel↓	SqRel↓	δ_1 ↑
ZoeDepth-K [9]	0.976	81.752	0
ZoeDepth-N [9]	0.975	81.508	0
ZoeDepth-NK [9]	0.894	69.939	0
Depth Anything (ViT-b) [108]	0.707	46.102	0.010
Depth Anything (ViT-l) [108]	0.472	36.029	0.453
Depth Anything V2 (ViT-b) [108]	0.939	76.630	1.670
Depth Anything V2 (ViT-l) [108]	1.517	261.925	0.089
Metric3D (ConvNeXt-t) [113]	0.790	58.411	0.009
Metric3D (ConvNeXt-l) [113]	0.682	53.504	0.160
Metric3D V2 (ViT-s) [40]	0.830	68.084	0.028
Metric3D V2 (ViT-l) [40]	0.540	31.960	0.074
Marigold [45]	0.994	84.409	0
GeoWizard [34]	0.995	84.485	0

Table 7. Zero-shot depth estimation performance on the UAVScenes dataset. "l" denotes large. Above are single-step models, and below Marigold and GeoWizard are diffusion-based models that can only produce affine-invariant depth maps. Evaluation metrics follow MonoDepth2 [37].

UAVScenes is a versatile multi-modal UAV dataset that provides rich semantic annotations for both 2D images and 3D LiDAR point clouds. With precisely aligned 6-DoF poses and associated 3D maps, it accommodates diverse research needs. By introducing a standardized benchmark for UAV perception tasks, UAVScenes offers consistent evaluation and comparison across multiple modalities. It thus serves as a fundamental resource for advancing UAV perception and mapping, driving progress in autonomous navigation, scene comprehension, and cross-modal learning within the UAV field.

References

- [1] Semantic drone dataset. <https://github.com/ayushdabra/drone-images-semantic-segmentation>. 2, 3
- [2] Stanislav Alexovič, Milan Lacko, and Ján Bačík. 3d mapping with a drone equipped with a depth camera in indoor environment. *Acta Electrotechnica et Informatica*, 23(1): 18–24, 2023. 2
- [3] Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. MixVPR: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2998–3007, 2023. 6, 7
- [4] Amar Alibey, Brahim Chaibdraa, and Philippe Giguere. GSV-Cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, 2022. 6, 7
- [5] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Padjla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016. 6
- [6] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 4
- [7] Jens Behley, Martin Garbade, Andres Milioto, Jan Quen- zel, 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*, pages 9297–9307, 2019. 4
- [8] Francesco Betti Sorbelli. Uav-based delivery systems: a systematic review, current trends, and research challenges. *Journal on Autonomous Transportation Systems*, 1(3):1–40, 2024. 1, 2
- [9] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 8
- [10] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses. In *CVPR*, 2023. 7, 8
- [11] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35(10):1157–1163, 2016. 2
- [12] 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*, pages 11621–11631, 2020. 3
- [13] Haoxin Cai, Shenghai Yuan, Xinyi Li, Junfeng Guo, and Jianqi Liu. BEV-LIO(LC): Bev image assisted lidar-inertial odometry with loop closure. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025. 3
- [14] Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation. *arXiv preprint arXiv:2305.13608*, 2023. 2, 3
- [15] Kun Cao, Muqing Cao, Shenghai Yuan, and Lihua Xie. Direct: A differential dynamic programming based framework for trajectory generation. *IEEE Robotics and Automation Letters*, 7(2):2439–2446, 2022. 1
- [16] Muqing Cao, Yang Lyu, Shenghai Yuan, and Lihua Xie. Online trajectory correction and tracking for facade inspection using autonomous uav. In *2020 IEEE 16th International Conference on Control & Automation (ICCA)*, pages 1149–1154. IEEE, 2020. 1
- [17] Muqing Cao, Kun Cao, Shenghai Yuan, Kangcheng Liu, Yan Loi Wong, and Lihua Xie. Path planning for multiple tethered robots using topological braids. In *Proceedings of Robotics: Science and Systems*, 2023. 1
- [18] Muqing Cao, Kun Cao, Shenghai Yuan, Thien-Minh Nguyen, and Lihua Xie. Neptune: nonentangling trajectory planning for multiple tethered unmanned vehicles. *IEEE Transactions on Robotics*, 39(4):2786–2804, 2023. 1
- [19] Meida Chen, Qingyong Hu, Zifan Yu, Hugues Thomas, Andrew Feng, Yu Hou, Kyle McCullough, Fengbo Ren, and Lucio Soibelman. Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. *arXiv preprint arXiv:2203.09065*, 2022. 3
- [20] Kai Cheng, Xiaoxiao Long, Kaizhi Yang, Yao Yao, Wei Yin, Yuexin Ma, Wenping Wang, and Xuejin Chen. Gaussianpro: 3d gaussian splatting with progressive propagation. In *ICML*, 2024. 6, 8
- [21] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 6, 7
- [22] Ivan Cisneros, Peng Yin, Ji Zhang, Howie Choset, and Sebastian Scherer. Alto: A large-scale dataset for uav visual place recognition and localization. *arXiv preprint arXiv:2207.12317*, 2022. 2, 3
- [23] SpConv Contributors. SpConv: Spatially sparse convolution library. <https://github.com/traveller59/spconv>, 2022. 6, 7
- [24] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 4
- [25] Tianchen Deng, Yaohui Chen, Leyan Zhang, Jianfei Yang, Shenghai Yuan, Jiuming Liu, Danwei Wang, Hesheng Wang, and Weidong Chen. Compact 3d gaussian splatting for dense visual slam. *arXiv preprint arXiv:2403.11247*, 2024. 2
- [26] Tianchen Deng, Nailin Wang, Chongdi Wang, Shenghai Yuan, Jingchuan Wang, Danwei Wang, and Weidong Chen. Incremental joint learning of depth, pose and implicit scene representation on monocular camera in large-scale scenes. *arXiv preprint arXiv:2404.06050*, 2024. 2
- [27] Devansh Dhrafani, Yifei Liu, Andrew Jong, Ukcheol Shin, Yao He, Tyler Harp, Yaoyu Hu, Jean Oh, and Sebastian

- Scherer. Firestereo: Forest infrared stereo dataset for uas depth perception in visually degraded environments. *arXiv preprint arXiv:2409.07715*, 2024. 2, 3, 4
- [28] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6, 7
- [29] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 2
- [30] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018. 2, 3
- [31] Meng Joo Er, Shenghai Yuan, and Ning Wang. Development control and navigation of octocopter. In *2013 10th IEEE International Conference on Control and Automation (ICCA)*, pages 1639–1643. IEEE, 2013. 2
- [32] Changfeng Feng, Zhenyuan Chen, Renke Kou, Guangwei Gao, Chunping Wang, Xiang Li, Xiangbo Shu, Yimian Dai, Qiang Fu, and Jian Yang. Hazydet: Open-source benchmark for drone-view object detection with depth-cues in hazy scenes. *arXiv preprint arXiv:2409.19833*, 2024. 2, 3
- [33] Michael Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2, 3
- [34] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2025. 8
- [35] Alberto García-Hernández, Riccardo Giubilato, Klaus H Strobl, Javier Civera, and Rudolph Triebel. Unifying local and global multimodal features for place recognition in aliased and low-texture environments. *arXiv preprint arXiv:2403.13395*, 2024. 6, 7
- [36] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3
- [37] Clément Godard, Oisín Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. 2019. 8
- [38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6, 7
- [39] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, pages 4145–4153, 2017. 2, 3
- [40] 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. *arXiv preprint arXiv:2404.15506*, 2024. 8
- [41] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17658–17668, 2024. 6, 7
- [42] Peng Jiang, Philip Osteen, Maggie Wigness, and Srikanth Saripalli. Rellis-3d dataset: Data, benchmarks and analysis. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 1110–1116. IEEE, 2021. 4
- [43] Peng Jiang, Kasi Viswanath, Akhil Nagariya, George Chustz, Maggie Wigness, Philip Osteen, Timothy Overbye, Christian Ellis, Long Quang, and Srikanth Saripalli. Go: The great outdoors multimodal dataset. *arXiv preprint arXiv:2501.19274*, 2025. 4
- [44] Yihang Jiang, Xiaoyang Li, Guangxu Zhu, Hang Li, Jing Deng, Kaifeng Han, Chao Shen, Qingjiang Shi, and Rui Zhang. 6g non-terrestrial networks enabled low-altitude economy: Opportunities and challenges. *arXiv preprint arXiv:2311.09047*, 2023. 1
- [45] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 8
- [46] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 2023. 6, 7
- [47] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-DoF camera relocalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2938–2946, 2015. 7, 8
- [48] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4):1–14, 2023. 6, 8
- [49] Michael Kölle, Dominik Laupheimer, Stefan Schmohl, Norbert Haala, Franz Rottensteiner, Jan Dirk Wegner, and Hugo Ledoux. 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 Journal of Photogrammetry and Remote Sensing*, 1:100001, 2021. 2, 3
- [50] Jacek Komorowski. Minkloc3D: Point cloud based large-scale place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1790–1799, 2021. 6, 7
- [51] Jacek Komorowski. Improving point cloud based place recognition with ranking-based loss and large batch train-

- ing. In *Proceedings of the International Conference on Pattern Recognition*, pages 3699–3705, 2022. 6, 7
- [52] Jacek Komorowski, Monika Wysoczańska, and Tomasz Trzcinski. MinkLoc++: LiDAR and monocular image fusion for place recognition. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8, 2021. 6, 7
- [53] Giorgos Kordopatis-Zilos, Panagiotis Galopoulos, Symeon Papadopoulos, and Ioannis Kompatsiaris. Leveraging efficientnet and contrastive learning for accurate global-scale location estimation. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 155–163, 2021. 6, 7
- [54] Christos Kyrkou and Theodoris Theodoridis. Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles. In *CVPR workshops*, pages 517–525, 2019. 1
- [55] Haowen Lai, Peng Yin, and Sebastian Scherer. AdaFusion: Visual-LiDAR fusion with adaptive weights for place recognition. *IEEE Robotics and Automation Letters*, 7(4): 12038–12045, 2022. 6, 7
- [56] Haotian Li, Yuying Zou, Nan Chen, Jiarong Lin, Xiyuan Liu, Wei Xu, Chunran Zheng, Rundong Li, Dongjiao He, Fanze Kong, et al. Mars-lvig dataset: A multi-sensor aerial robots slam dataset for lidar-visual-inertial-gnss fusion. *The International Journal of Robotics Research*, page 02783649241227968, 2024. 2, 3, 4
- [57] Xueping Li, Jose Tupayachi, Aliza Sharmin, and Madelaine Martinez Ferguson. Drone-aided delivery methods, challenge, and the future: A methodological review. *Drones*, 7(3):191, 2023. 2
- [58] Xiaowei Li, Kuan Xu, Fen Liu, Ruofei Bai, Shenghai Yuan, and Lihua Xie. Airswarm: Enabling cost-effective multi-uav research with cots drones. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025. 3
- [59] 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, 2022. 3
- [60] Jiarong Lin and Fu Zhang. R 3 live: A robust, real-time, rgb-colored, lidar-inertial-visual tightly-coupled state estimation and mapping package. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10672–10678. IEEE, 2022. 4
- [61] Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. Capturing, reconstructing, and simulating: the urbanscene3d dataset. In *European Conference on Computer Vision*, pages 93–109. Springer, 2022. 2, 3
- [62] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 6, 7
- [63] Boyang Lou, Shenghai Yuan, Jianfei Yang, Wenju Su, Yingjian Zhang, and Enwen Hu. Qlio: Quantized lidar-inertial odometry. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025. 3
- [64] Lun Luo, Shuhang Zheng, Yixuan Li, Yongzhi Fan, Beinan Yu, Si-Yuan Cao, Junwei Li, and Hui-Liang Shen. BEV-Place: Learning LiDAR-based place recognition using bird’s eye view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8700–8709, 2023. 6, 7
- [65] 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. 2, 3
- [66] Yang Lyu, Shenghai Yuan, and Lihua Xie. Structure priors aided visual-inertial navigation in building inspection tasks with auxiliary line features. *IEEE Transactions on Aerospace and Electronic Systems*, 58(4):3037–3048, 2022. 1
- [67] Yang Lyu, Muqing Cao, Shenghai Yuan, and Lihua Xie. Vision-based plane estimation and following for building inspection with autonomous uav. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023. 1
- [68] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 6
- [69] UM Rao Mogili and BBVL Deepak. Review on application of drone systems in precision agriculture. *Procedia computer science*, 133:502–509, 2018. 1
- [70] Syed Agha Hassnain Mohsan, Muhammad Asghar Khan, Fazal Noor, Insaf Ullah, and Mohammed H Alsharif. Towards the unmanned aerial vehicles (uavs): A comprehensive review. *Drones*, 6(6):147, 2022. 1
- [71] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 6, 8
- [72] Son Tung Nguyen, Alejandro Fontan, Michael Milford, and Tobias Fischer. Focustune: Tuning visual localization through focus-guided sampling. In *WACV*, pages 3594–3603, 2024. 7, 8
- [73] Thien-Minh Nguyen, Muqing Cao, Shenghai Yuan, Yang Lyu, Thien Hoang Nguyen, and Lihua Xie. Viral-fusion: A visual-inertial-ranging-lidar sensor fusion approach. *IEEE Transactions on Robotics*, 38(2):958–977, 2021. 2
- [74] Thien-Minh Nguyen, Shenghai Yuan, Muqing Cao, Yang Lyu, Thien H Nguyen, and Lihua Xie. Ntu viral: A visual-inertial-ranging-lidar dataset, from an aerial vehicle viewpoint. *The International Journal of Robotics Research*, 41(3):270–280, 2022. 2, 3
- [75] 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. 2, 3
- [76] Linus Nwankwo, Bjoern Ellensohn, Vedant Dave, Peter Hofer, Jan Forstner, Marlene Villneuve, Robert Galler, and Elmar Rueckert. Envodat: A large-scale multisensory dataset for robotic spatial awareness and semantic

- reasoning in heterogeneous environments. *arXiv preprint arXiv:2410.22200*, 2024. 4
- [77] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6, 7
- [78] Dong-Hee Paek, Seung-Hyun Kong, and Kevin Tirta Wijaya. K-radar: 4d radar object detection for autonomous driving in various weather conditions. *Advances in Neural Information Processing Systems*, 35:3819–3829, 2022. 3
- [79] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2018. 6, 7
- [80] Maryam Rahnemounfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Robertson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9: 89644–89654, 2021. 2, 3
- [81] 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. 2, 3
- [82] 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. 4
- [83] Hazim Shakhatreh, Ahmad H Sawalmeh, Ala Al-Fuqaha, Zuocho Dou, Eyad Almaita, Issa Khalil, Noor Shamsiah Othman, Abdallah Khreishah, and Mohsen Guizani. Unmanned aerial vehicles (uavs): A survey on civil applications and key research challenges. *Ieee Access*, 7:48572–48634, 2019. 1
- [84] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 4
- [85] Pei Sun, Henrik Kretschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 3
- [86] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022. 2, 3
- [87] Ravindu G Thalagala, Oscar De Silva, Awantha Jayasiri, Arthur Gubbels, George KI Mann, and Raymond G Gosine. Mun-fri: A visual-inertial-lidar dataset for aerial autonomous navigation and mapping. *The International Journal of Robotics Research*, page 02783649241238358, 2024. 2, 3, 4
- [88] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European conference on computer vision*, pages 516–533. Springer, 2022. 6, 7
- [89] Kavisha Vidanapathirana, Joshua Knights, Stephen Hausler, Mark Cox, Milad Ramezani, Jason Jooste, Ethan Griffiths, Shaheer Mohamed, Sridha Sridharan, Clinton Fookes, et al. Wildscenes: A benchmark for 2d and 3d semantic segmentation in large-scale natural environments. *The International Journal of Robotics Research*, page 02783649241278369, 2024. 4
- [90] Bing Wang, Changhao Chen, Chris Xiaoxuan Lu, Peijun Zhao, Niki Trigoni, and Andrew Markham. AtLoc: Attention guided camera localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10393–10401, 2020. 7, 8
- [91] Dashuai Wang, Wei Li, Xiaoguang Liu, Nan Li, and Chunlong Zhang. Uav environmental perception and autonomous obstacle avoidance: A deep learning and depth camera combined solution. *Computers and Electronics in Agriculture*, 175:105523, 2020. 2
- [92] Fangjinhua Wang, Xudong Jiang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Glace: Global local accelerated coordinate encoding. In *CVPR*, 2024. 7, 8
- [93] Han Wang, Shenghai Yuan, and Keyu Wu. Heterogeneous stereo: A human vision inspired method for general robotics sensing. In *TENCON 2017-2017 IEEE region 10 conference*, pages 793–798. IEEE, 2017. 3
- [94] Linhan Wang, Kai Cheng, Shuo Lei, Shengkun Wang, Wei Yin, Chenyang Lei, Xiaoxiao Long, and Chang-Tien Lu. Dc-gaussian: Improving 3d gaussian splatting for reflective dash cam videos. In *NeurIPS 2024*, 2024. 6, 8
- [95] Sijie Wang, Qiyu Kang, Rui She, Wee Peng Tay, Andreas Hartmannsgruber, and Diego Navarro Navarro. RobustLoc: Robust camera pose regression in challenging driving environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6209–6216, 2023. 7, 8
- [96] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767, 2024. 4
- [97] 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)*, pages 4909–4916. IEEE, 2020. 2, 3
- [98] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023. 6, 7
- [99] Rouwan Wu, Xiaoya Cheng, Juelin Zhu, Yuxiang Liu, Maojun Zhang, and Shen Yan. Uavd4l: A large-scale

- dataset for uav 6-dof localization. In *2024 International Conference on 3D Vision (3DV)*, pages 1574–1583. IEEE, 2024. 2, 3
- [100] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022. 6, 7
- [101] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 6
- [102] Butian Xiong, Zhuo Li, and Zhen Li. Gauu-scene: A scene reconstruction benchmark on large scale 3d reconstruction dataset using gaussian splatting. *arXiv preprint arXiv:2401.14032*, 2024. 2, 3
- [103] Butian Xiong, Nanjun Zheng, Junhua Liu, and Zhen Li. Gauu-scene v2: Assessing the reliability of image-based metrics with expansive lidar image dataset using 3dgs and nerf. *CoRR*, 2024. 2, 3, 4
- [104] Wenjia Xu, Yaxuan Yao, Jiaqi Cao, Zhiwei Wei, Chunbo Liu, Jiuniu Wang, and Mugen Peng. Uav-visloc: A large-scale dataset for uav visual localization. *arXiv preprint arXiv:2405.11936*, 2024. 2, 3
- [105] Xinhang Xu, Muqing Cao, Shenghai Yuan, Thien Hoang Nguyen, Thien-Minh Nguyen, and Lihua Xie. A cost-effective cooperative exploration and inspection strategy for heterogeneous aerial system. In *Proceedings of the 2024 International Conference on Control, Automation, and Systems (ICCA)*, 2024. 1
- [106] Zhefan Xu, Xiaoyang Zhan, Baihan Chen, Yumeng Xiu, Chenhao Yang, and Kenji Shimada. A real-time dynamic obstacle tracking and mapping system for uav navigation and collision avoidance with an rgb-d camera. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10645–10651. IEEE, 2023. 2
- [107] Qi Yan, Jianhao Zheng, Simon Reding, Shanci Li, and Jordan Doytchinov. Crossloc: Scalable aerial localization assisted by multimodal synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17358–17368, 2022. 2, 3
- [108] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 8
- [109] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 8
- [110] Yizhuo Yang, Shenghai Yuan, and Lihua Xie. Overcoming catastrophic forgetting for semantic segmentation via incremental learning. In *2022 17th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 299–304. IEEE, 2022. 2
- [111] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 4
- [112] Pengyu Yin, Haozhi Cao, Thien-Minh Nguyen, Shenghai Yuan, Shuyang Zhang, Kangcheng Liu, and Lihua Xie. Outram: One-shot global localization via triangulated scene graph and global outlier pruning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13717–13723. IEEE, 2024. 2
- [113] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 8
- [114] Weihao Yu and Xinchao Wang. Mambaout: Do we really need mamba for vision? *arXiv preprint arXiv:2405.07992*, 2024. 6, 7
- [115] Shenghai Yuan and Han Wang. Autonomous object level segmentation. In *2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*, pages 33–37. IEEE, 2014. 3
- [116] Zheng Zhang, Wenbo Hu, Yixing Lao, Tong He, and Hengshuang Zhao. Pixel-gs: Density control with pixel-aware gradient for 3d gaussian splatting. In *ECCV*, 2024. 6, 8
- [117] Chunran Zheng, Qingyan Zhu, Wei Xu, Xiyuan Liu, Qizhi Guo, and Fu Zhang. Fast-livo: Fast and tightly-coupled sparse-direct lidar-inertial-visual odometry. In *2022 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4003–4009. IEEE, 2022. 4
- [118] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 1395–1403, 2020. 2, 3
- [119] Zijie Zhou, Jingyi Xu, Guangming Xiong, and Junyi Ma. LCPR: A multi-scale attention-based LiDAR-camera fusion network for place recognition. *IEEE Robotics and Automation Letters*, 2024. 6, 7
- [120] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018. 2, 3
- [121] Runzhe Zhu, Ling Yin, Mingze Yang, Fei Wu, Yuncheng Yang, and Wenbo Hu. Sues-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4825–4839, 2023. 2, 3
- [122] Yilin Zhu, Yang Kong, Yingrui Jie, Shiyong Xu, and Hui Cheng. Graco: A multimodal dataset for ground and aerial cooperative localization and mapping. *IEEE Robotics and Automation Letters*, 8(2):966–973, 2023. 2, 3, 4