arXiv:2603.19609v1 [cs.CV] 20 Mar 2026

LoD-Loc v3: Generalized Aerial Localization in Dense Cities using Instance Silhouette Alignment

Shuaibang Peng[1], Juelin Zhu[1†], Xia Li[1], Kun Yang[2], Maojun Zhang[1], Yu Liu[1], Shen Yan[1†]
[1]National University of Defense Technology, [2]Northwestern Polytechnical University
{psb24,zhujuelin,lixia24,mjzhang,jasonyuliu,yanshen12}@nudt.edu.cn

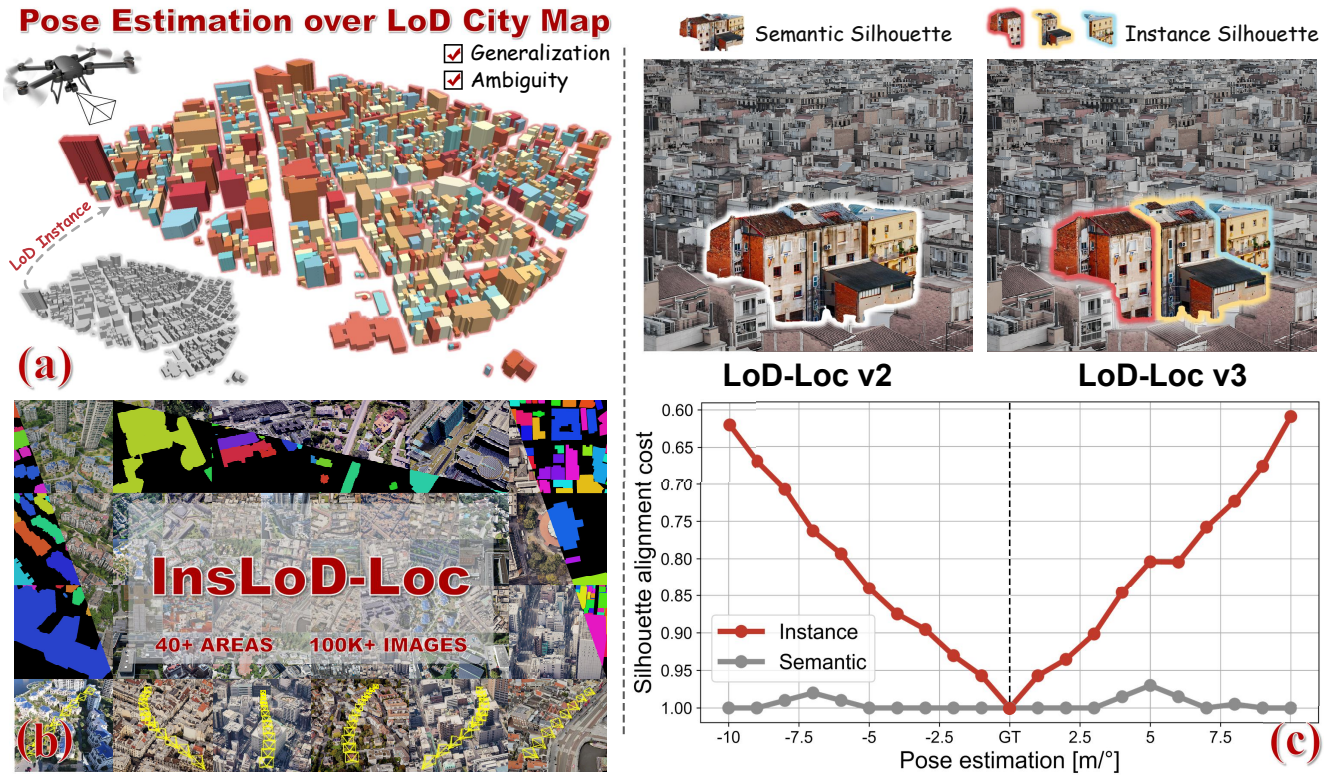

Figure 1. In this paper, (a) we introduce LoD-Loc v3 to address two critical challenges in aerial localization over LoD city models: cross-scene generalization and the ambiguity problem in dense urban scenes. Our solutions are twofold: (b) we construct **InsLoD -Loc**, a large-scale synthetic dataset covering 40 distinct areas for model zero-shot training, and (c) we reformulate the localization paradigm by shifting from semantic to instance silhouette alignment, which provides superior convergence.

Abstract

*We present LoD-Loc v3, a novel method for generalized aerial visual localization in dense urban environments. While prior work LoD-Loc v2 [94] achieves localization through semantic building silhouette alignment with low-detail city models, it suffers from two key limitations: poor cross-scene generalization and frequent failure in dense building scenes. Our method addresses these challenges through two key innovations. First, we develop a new synthetic data generation pipeline that produces **InsLoD-Loc** - the largest instance segmentation dataset for aerial imagery to date, comprising 100k images with precise instance building annotations. This enables trained models to exhibit remarkable zero-shot generalization capability. Second, we reformulate the localization paradigm by shifting from semantic to instance silhouette alignment, which significantly reduces pose estimation ambiguity in dense scenes. Extensive experiments demonstrate that LoD-Loc v3 outper-*

†Corresponding author

forms existing state-of-the-art (SOTA) baselines, achieving superior performance in both cross-scene and dense urban scenarios with a large margin. The project is available at https://nudt-sawlab.github.io/LoD-Locv3/.

1. Introduction

Visual localization for unmanned aerial vehicle (UAV) estimates a camera's pose by matching visual inputs against a georeferenced map, which is a core technology for autonomous systems, supporting applications such as precision navigation [50], cargo transport [81], and emergency response [7, 87]. The dominant paradigm [12, 87, 89] relies on high-fidelity 3D reconstructions, typically generated using Structure-from-Motion (SfM) [71] or photogrammetry [60]. While effective, these high-fidelity models are expensive to create and maintain, and their large data volume raises privacy and protection concerns.

To address these problems, localization methods based on Level-of-Detail (LoD) city models have emerged as a promising alternative [93, 94]. The LoD model is a 3D representation focused on architectural structures, following the CityGML standard [28, 44, 88]. The construction of LoD models has become a global trend, with nationwide productions in the USA [6], China [56, 78], Switzerland [22, 57], Japan [55], Singapore [3, 75], Germany [72], and the Netherlands [18, 86]. Academia and industry are also accelerating this, with open-source global datasets [95] and AI-driven commercial mapping [20]. As a result, using LoD models for localization has the potential to support global UAV navigation in urban environments.

The early LoD-Loc [93] pioneered this direction by aligning wireframes from high-detail LoD models. LoD-Loc v2 [94] extended this to ubiquitous low-detail LoD1 models by aligning building silhouettes. However, LoD-Loc v2 [94] suffers from two critical drawbacks: it usually fails to generalize to unseen scenes, and it completely fails to localize in dense scenes because of the ambiguity of the semantic silhouette, as illustrated in Fig. 1.

To overcome the generalization and ambiguity challenges, we propose LoD-Loc v3. For generalization, we developed a novel synthetic data pipeline to construct **InsLoD-Loc**, a 100k-image dataset with precise instance-level building annotation. To our knowledge, this is the largest aerial instance segmentation dataset, and our pipeline is designed for easy extension. Specifically, this pipeline involves two stages: First, we render the photorealistic RGB images within the Unreal Engine 5 (UE5) [25] platform, which uses the Cesium for Unreal plugin [8] to stream Google Earth Photorealistic 3D Tileset data [35] and the AirSim plugin [65] for capture. Second, we source the corresponding LoD models from [18, 54, 57, 91, 94], align their coordinate systems with the 3D tileset data, and employ OpenSceneGraph (OSG) [58] with our model instancing method (Sec. 3.2) to render precise instance masks using the identical camera poses with RGB-based rendering.

To address the ambiguity problem, we shift the core paradigm from semantic silhouettes to instance silhouettes alignment. Our motivation is that localization in cities should be viewed as an instance alignment process, which, as shown in Fig. 1, is particularly critical in dense environments. Specifically, our method involves three stages: First, a novel 'LoD model instancing' preprocessing step assigns unique identifiers to each building for instance-aware rendering. Second, we use our constructed **InsLoD-Loc** dataset to fine-tune a SAM-based model to extract building instance silhouettes from the query image. Finally, we estimate the pose by aligning the query and rendered instance silhouettes, similar to the coarse-to-fine localization framework in LoD-Loc v2 [94].

On public benchmarks, our method exceeds current SOTA baselines, notably without being trained in-distribution. Besides, we constructed a dataset containing multiple dense urban scenes. On this challenging dataset, our localization results demonstrate a **2000%** improvement over SOTA methods at the $(2\text{m}, 2^\circ)$ accuracy.

In summary, our main contributions are as follows:

1. We propose a synthetic data pipeline and construct a large-scale dataset, **InsLoD-Loc**, which addresses poor cross-scene generalization.
2. We introduce a novel localization method based on instance silhouettes, which resolves the ambiguity problem in dense scenes.
3. We demonstrate through extensive experiments that our method achieves SOTA performance on two public benchmarks and our own dense-scene dataset.

2. Related Works

Localization over 3D Models. Mainstream high-precision visual localization [60, 66, 67, 77, 87] typically relies on information-rich 3D models like SfM point clouds [71] or textured meshes [60]. The core of these methods is to establish 2D-to-3D correspondences, often involving image retrieval [2, 27] followed by local feature matching [19, 38, 49, 67, 77]. A 6-DoF pose is then solved via a PnP-RANSAC algorithm [4, 16, 17, 23, 30, 43, 46, 90]. While accurate, these methods depend on detailed maps that are costly to create and maintain, computationally intensive, raise privacy and protection concerns.

Localization over LoD Models. Early LoD localization was limited to ground-level views [63, 64, 70]. LoD-Loc [93] pioneered aerial localization by aligning neural wireframes with high-detail LoD2/3 models. LoD-Loc v2 [94] adapted this to more common low-detail LoD1 models by aligning building silhouettes. However, its re-

liance on semantic segmentation led to two fundamental challenges: poor generalization due to limited training data, and catastrophic localization failure in dense urban scenes because of ambiguity. To address these, LoD-Loc v3 enhances generalization and ambiguity by introducing a large-scale synthetic dataset and shifting the paradigm from semantic to instance-level alignment.

Instance Segmentation. The goal of instance segmentation is to identify pixel-level and instance-level silhouettes for every object in an image. Traditional methods are two-stage [10, 31, 33, 41, 82] or single-stage [5, 9, 14, 15, 45, 83]. Vision foundation models like SAM [42] have introduced a paradigm shift. Their zero-shot capabilities spurred follow-up works on efficiency [92], quality [39], and domain adaptation [11, 13]. However, the success of these data-driven methods is highly dependent on large-scale, high-quality annotated datasets. While many benchmarks exist for general [29, 47, 73], medical [36, 74], or remote sensing [76, 85] domains, they have a significant domain gap with our task. Even the WHU Building dataset [37] consists of nadir-view satellite imagery, unlike the low-altitude, oblique UAV views in urban canyons. While previous works have demonstrated the immense value of synthetic training data for vision tasks [53] and the feasibility of simulating city-scale aerial imagery for feature evaluation [1, 26], there remains a lack of dedicated instance-level building datasets for localization. To fill this gap, we constructed our 100k-image synthetic dataset to support research in building instance segmentation and localization from low-altitude aerial perspectives.

3. Method

Problem Formulation. Given a LoD model $\mathcal{M}$, a query image I_q, and its prior pose estimate $\boldsymbol{\xi}_q^p$, the objective is to estimate the absolute pose $\boldsymbol{\xi}_q^*$ of I_q.

Overview. In the following sections, we first briefly review the localization framework of LoD-Loc v2 [94] as a preliminary. We then introduce our core contributions: a large-scale dataset constructed via our proposed synthetic data generation pipeline (detailed in Sec. 4.1) to address the problem of poor cross-scene generalization, and a localization algorithm based on instance silhouette alignment to resolve the issue of ambiguity in dense scenes.

3.1. The Coarse-to-Fine Localization Framework

UAV inertial sensors reliably provide gravity direction, reducing the problem to a 4-DoF pose $\boldsymbol{\xi}_q^* = (x, y, z, \theta)$ (3D translation, 1D yaw) [24, 51, 68, 69, 90]. LoD-Loc v2 [94] solves this by aligning a rendered silhouette M_{hyp} from the LoD model $\mathcal{M}$ with a predicted image silhouette M_q, quantified by Intersection over Union (IoU):

$$\mathrm{c}(I_q, I_{hyp}) = \frac{|M_q \cap M_{hyp}|}{|M_q \cup M_{hyp}|} \tag{1}$$

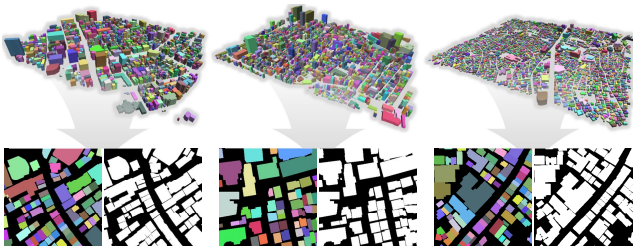

Figure 2. **Visualization of instanced LoD models and labels.** The top row displays the instanced LoD models. The bottom row illustrates the rendered instance labels rendered from the models, alongside their corresponding semantic labels.

The **coarse stage** finds an initial pose $\boldsymbol{\xi}_q^c$ by uniformly sampling the 4-DoF search space (x, y, z, θ) and maximizing c. The **fine stage** then uses a particle filter to refine this estimate, iteratively maximizing the cost c to find the final pose $\boldsymbol{\xi}_q^*$ [34, 38, 40, 48, 52, 62, 80]. Detailed descriptions of both stages can be found in [94].

3.2. Scaling to Open-World Scenes

Although LoD-Loc v2 [94] performs well in specific scenes, its performance degrades significantly or fails entirely in cross-scene or dense urban environments. To address these issues, we propose a new methodology designed to enhance model generalization and ambiguity.

To improve generalization, we constructed **InsLoD-Loc**, a large-scale synthetic dataset (Sec. 4). To address ambiguity in dense scenes, we replace the semantic silhouette cost (Eq. (1)) with an instance silhouette alignment.

Our approach has three stages: First, an instance method for LoD models, assigning a unique ID to each building. Second, an instance segmentation model to extract individual building silhouettes from the query image. Finally, a novel alignment and evaluation mechanism that leverages these instance silhouettes.

LoD Model Instancing. We assign a unique color to each building in $\mathcal{M}$ via topological structure analysis. By parsing the untextured $\mathcal{M}$ as a graph $G = (V, E, F)$, each building B_i forms a connected component G_i, formulating instancing as a graph partition:

$$\mathcal{M} = \bigcup_{i=1}^{M} B_i, \quad \text{s.t.} \quad B_i \cap B_j = \emptyset \quad \forall i \neq j \tag{2}$$

where M is the number of buildings. Each B_i receives a unique 24-bit integer ID mapped to an RGB color (details in Appendix). Assigning this color to all faces of B_i yields the instanced LoD model $\mathcal{M}_{ins}$ as in Fig. 2. Rendering $\mathcal{M}_{ins}$ from any pose $\boldsymbol{\xi}_{hyp}$ produces an Instance Map I_{hyp}, where colors uniquely encode building identities.

Building Silhouette Instance Segmentation. This stage segments each building silhouette from I_q. We fine-tune

a SAM-based model [42] on our **InsLoD-Loc** dataset, inspired by prompt learning [11]. The network, composed of a SAM image encoder, a learnable Prompter Module, and a SAM mask decoder, is adapted to extract precise building silhouettes from aerial images. Specifically, for a given input query image $I_q \in \mathbb{R}^{3\times H\times W}$, the SAM encoder extracts an image embedding F_{embed}. The Prompter Module predicts prompt embeddings from F_{embed}. The SAM decoder takes both embeddings to generate instance masks $\mathcal{S}_q = \{M_q^j\}_{j=1}^N$. This transforms SAM into an automatic, task-specific segmentation pipeline, providing unambiguous geometric constraints and resolving ambiguity from merged silhouettes. Further details are in the Appendix.

Pose Evaluation via Instance Alignment. The core of our localization framework is a novel cost function, c_{ins}, used to evaluate any given pose hypothesis $\boldsymbol{\xi}_{hyp}$. For each hypothesis, we first project the instanced LoD model $\mathcal{M}_{ins}$ using OSG [58] real-time rendering to obtain its set of instance masks, $\mathcal{S}_{hyp} = \{M_{hyp}^k\}_{k=1}^K$. We then compute the alignment cost between this set and the query instance set $\mathcal{S}_q = \{M_q^j\}_{j=1}^N$ using an asymmetric matching strategy. For each predicted instance M_q^j, we find its best match in $\mathcal{S}_{hyp}$ (highest Dice coefficient) d_j^*:

$$d_j^* = \max_{k\in\{1,\dots,K\}} \left(\frac{2\sum_p (M_q^j(p)\cdot M_{hyp}^k(p))}{\sum_p M_q^j(p) + \sum_p M_{hyp}^k(p) + \epsilon} \right) \quad (3)$$

where p is a pixel location, and ϵ is a small constant to prevent division by zero. The final cost c_{ins} is a weighted sum of d_j^*. We explore two weighting strategies:

- **Confidence-based Weighting:** This strategy uses the instance's confidence score s_j as weight $w_j^{(\text{conf})}$.

$$\mathrm{c}_{ins}^{(\text{conf})} = \sum_{j=1}^{N} \frac{s_j}{\sum_{i=1}^N s_i} \cdot d_j^* \quad (4)$$

- **Area-based Weighting:** This strategy uses the instance's bounding box area A_j as weight $w_j^{(\text{area})}$.

$$\mathrm{c}_{ins}^{(\text{area})} = \sum_{j=1}^{N} \frac{A_j}{\sum_{i=1}^N A_i} \cdot d_j^* \quad (5)$$

We adopt the coarse-to-fine localization framework from LoD-Loc v2 [94], using c_{ins} as the core metric. In the coarse stage, we construct the cost volume $\mathcal{C}_{Ins}$ by computing this value over a uniformly sampled grid to select an initial coarse pose $\boldsymbol{\xi}_q^c$. Subsequently, in the fine stage, this cost is used to evaluate particle weights, guiding the optimization process to converge to the final, precise pose $\boldsymbol{\xi}_q^*$.

Supervision. Our instance segmentation network is trained in a supervised manner using pairs of query images I_q, ground truth masks G, and corresponding bounding boxes B. During training, we update the learnable Prompter Module and the SAM mask decoder, while applying LoRA [32] for parameter-efficient fine-tuning to the otherwise frozen SAM vision encoder, significantly reducing computational overhead. The model's overall loss function L is a multi-task loss, composed of a Region Proposal Network (RPN) loss L_{rpn} and a Region of Interest (RoI) head loss L_{roi}:

$$L = L_{\text{rpn}} + L_{\text{roi}} \quad (6)$$

where L_{rpn} supervises candidate box generation within the Prompter Module, and L_{roi} supervises the final classification, regression, and the SAM decoder's mask predictions.

4. Dataset

We construct the **InsLoD-Loc** dataset, a large-scale synthetic dataset for building instance segmentation and 6-DoF localization. It comprises 108,109 RGB images with pixel-accurate instance annotations, covering 40 UAV flight areas across six countries (Japan, Switzerland, China, France, Italy, Netherlands). Fig. 3 shows the geographic distribution of these areas. This section details the construction pipeline and composition of the **InsLoD-Loc** dataset.

4.1. Simulation Environment and Data Generation

Photorealistic RGB Image Generation. To generate high-quality query images, we employed a comprehensive simulation framework based on UE5 version 5.1.1 [25]. Urban scenes are built by streaming Google Photorealistic 3D Tileset data [35] via the Cesium for Unreal plugin [8]. We then utilized the integrated AirSim plugin from Microsoft Research version 2.1.0 [65] to render these scenes from multiple angles, at various altitudes, and across diverse locations. Further details are provided in the Appendix.

Automated Instance Mask Generation. To generate instance masks with pixel-perfect alignment to the RGB images, we first source the corresponding geo-referenced LoD models [18, 54, 57, 91, 94]. These models are converted to the OBJ mesh format. We unify their coordinate systems to precisely match the reference system of the Google 3D Tileset data [35]. This enables us to render a corresponding instance mask via the real-time rendering engine OSG [58] using the *exact same* camera intrinsic and extrinsic parameters as each rendered RGB frame. Finally, we apply Connected Component Analysis to these rendered instance maps to generate the precise instance mask annotations. Further details are provided in the Appendix.

4.2. Data Acquisition and Composition

The dataset was collected from 40 geographically distinct areas using three camera configurations and the corresponding sampling methodologies to maximize diversity in viewpoint, altitude, and environment, as detailed in Tab. 1.

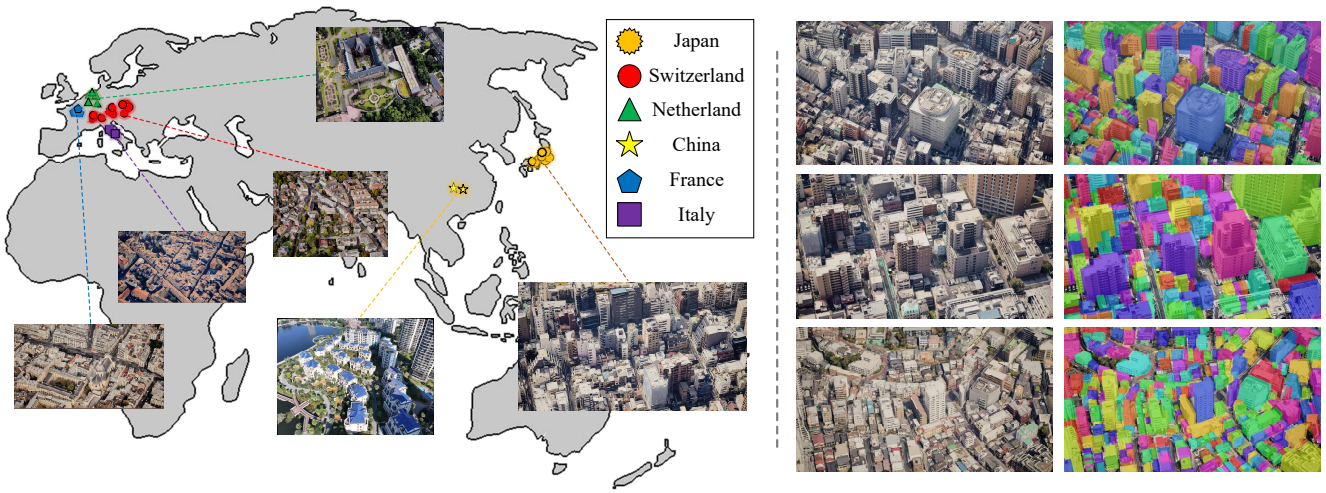

Figure 3. **Overview of the InsLoD-Loc dataset.** The left panel illustrates the geographic distribution of the 40 flight areas across Europe and Asia. The right panel showcases representative samples from the dataset, each displaying (from left to right): a photorealistic RGB query image and its corresponding pixel-accurate instance label, where each color represents a unique building.

Camera	Camera Configuration			Sampling Methodology			
	Resolution[px]	Sensor Size	FOV	Strategy	Path	View Angle	Height[m]
Camera 1	1600×1200	16mm×12mm	45°	Sequence	Irregular	[0°, 70°]	[200, 500]
Camera 2	1920×1080	23.76mm×13.365mm	25°	Grid	Uniform	0°, 45°	[200, 500]
Camera 3	1920×1080	23.76mm×13.365mm	25°	Sequence	Regular	0°, 45°	[400, 500]

Table 1. Camera configurations and sampling methodologies used for data acquisition.

Specifically, the **InsLoD-Loc** dataset encompasses diverse land-use categories, including commercial, industrial, and residential, as well as education, medical, and suburban zones, to ensure comprehensive environmental representation. It is partitioned into geographically independent training, validation, and test sets. For completeness, we note that the RGB images for the China region are available in LoD-Loc v2 [94]. Further details are provided in the Appendix.

5. Experiments

In this section, we first introduce datasets and baseline methods, as well as evaluation metrics in Sec. 5.1, followed by implementation details of our method in Sec. 5.2. Experimental results and ablation studies are detailed in Sec. 5.3 and Sec. 5.4, respectively.

5.1. Experiment Settings and Baselines

Datasets. We evaluate our method on three datasets: UAVD4L-LoDv2, Swiss-EPFLv2, and Tokyo-LoDv3. The UAVD4L-LoDv2 and Swiss-EPFLv2 datasets cover areas of 2.5 km^2 and 8.2 km^2. Details of these two public benchmarks can be found in LoD-Loc v2 [94]. Tokyo-LoDv3 serves as the test set from the **InsLoD-Loc** dataset. It covers six land-use categories and features five dense urban scenes along with corresponding instanced LoD1 models. Further details are provided in the Appendix.

Baselines. We compare against three baseline categories: those that perform localization via feature matching, namely CAD-Loc [61]; those that perform localization via feature alignment, namely MC-Loc [80]; and those that operate by aligning geometric primitives from LoD models, namely LoD-Loc [93] and LoD-Loc v2 [94]. Specifically, for the feature-matching category, we evaluate CAD-Loc with a suite of keypoint-based strategies: 1) the classic SIFT [49] descriptor with Nearest Neighbor (NN) matching; 2) the learning-based SuperPoint (SPP) [19] extractor with the SuperGlue (SPG) [67] matcher; 3) the detector-free matcher LoFTR [77] and its efficient variant 4) e-LoFTR [84]; and 5) the dense feature matcher RoMa [21]. For the feature-alignment category, we evaluate MC-Loc using two distinct feature extraction backbones: 6) a pre-trained DINOv2 [59] encoder and 7) the RoMa [21] dense feature extractor. For the LoD-based alignment category, 8) LoD-Loc [93] is implemented with its default parameters, and 9) we evaluate the three reported variants of LoD-Loc v2 [94]. Implementation details for the baseline exper-

Method		*in-Traj.*				*out-of-Traj.*			
		2m-2°	3m-3°	5m-5°	T.e./R.e.	2m-2°	3m-3°	5m-5°	T.e./R.e.
Prior		0	0	4.30	6.48/1.63	0	0	0.36	11.10/0.92
CAD-Loc [61]	SIFT+NN	0	0	0	-	0	0	0	-
	SPP+SPG	0	0	0	-	0	0	0	-
	LoFTR	0	0	0	-	0	0	0	-
	e-LoFTR	0	0	0	-	0	0	0	-
	RoMA	0	0	0	-	0	0	0	-
MC-Loc [80]	DINOv2	1.20	4.10	17.40	8.29/2.58	2.40	7.40	26.10	7.02/2.29
	RoMa	0.10	0.60	3.30	10.6/8.60	0.20	0.90	3.30	16.9/3.88
LoD-Loc† [93]	-	49.56	71.82	89.09	3.32/1.48	54.20	75.05	89.51	3.33/1.18
LoD-Loc v2† [94]	no refine	0	0	23.38	6.19/0.67	11.68	29.88	51.14	4.78/0.92
	no select	93.50	98.40	99.50	0.74/0.17	90.50	94.80	96.90	0.77/0.16
	Full	93.70	98.40	99.50	0.72/0.15	**97.90**	**99.80**	**100.00**	0.71/0.14
LoD-Loc v3	no refine	0	0	24.10	6.09/0.67	11.70	30.80	52.00	4.58/0.91
	no select	97.40	99.10	99.80	0.50/0.12	96.10	97.80	98.30	0.61/0.12
	Full	**97.60**	**98.90**	**99.70**	**0.49/0.13**	97.40	99.00	99.40	**0.60/0.12**

Table 2. **Quantitative comparison results of different methods over UAVD4L-LoDv2 dataset.** T.e. and R.e. denote median translation error (m) and median rotation error (°). † indicates models trained in-distribution on this dataset. Our method utilizes area-based weighting.

Method		*in-Place.*				*out-of-Place.*			
		2m-2°	3m-3°	5m-5°	T.e./R.e.	2m-2°	3m-3°	5m-5°	T.e./R.e.
Prior		0	0	0.56	17.6/3.87	0	0	1.06	17.9/3.94
CAD-Loc	*same**	**0**	**0**	**0**	**-**	**0**	**0**	**0**	**-**
MC-Loc	DINOv2	0.90	4.40	17.50	8.18/2.54	2.90	9.00	30.20	6.23/1.97
	RoMa	0.20	1.20	4.80	9.80/2.65	0.70	2.10	11.5	10.3/3.97
LoD-Loc†	-	36.79	50.56	69.77	2.87/1.78	14.24	31.39	59.89	8.73/2.78
LoD-Loc v2†	no refine	0.56	3.73	20.79	7.37/3.76	0.53	2.11	11.35	8.92/3.90
	no select	52.10	72.10	88.30	1.90/0.89	31.10	55.90	81.30	2.73/0.73
	Full	54.20	74.60	92.00	1.83/0.85	31.40	58.53	86.30	2.64/0.73
LoD-Loc v3	no refine	0.60	2.90	16.20	8.20/3.79	0.50	1.80	11.10	9.02/3.90
	no select	58.60	77.50	90.60	1.68/0.80	34.60	57.50	80.20	2.65/0.88
	Full	**58.60**	**79.90**	**95.40**	**1.61/0.77**	**36.90**	**64.10**	**88.90**	**2.43/0.83**

Table 3. **Quantitative comparison results of different methods over Swiss-EPFLv2 dataset.** The *same** indicates that the statistics are identical to those in Tab. 2. †, T.e., and R.e. have the same meanings as those in Tab. 2. Our method utilizes area-based weighting.

iments are provided in the Appendix.

Evaluation Metrics. Following standard localization evaluation protocols [79], we report the percentage of queries localized within $(2\text{m}, 2°)$, $(3\text{m}, 3°)$, and $(5\text{m}, 5°)$ thresholds, consistent with LoD-Loc [93].

5.2. Implementation Details

Training. Our segmentation module is trained on the 88,493-image training split of **InsLoD-Loc**. The input image size is set to $(512, 512)$. To optimize the training process, we use the AdamW optimizer with a base learning rate of 2×10^{-4} and a weight decay of 0.05. We employ a cosine annealing schedule and train for 20 epochs. The SAM encoder is based on the pre-trained `ViT-Huge` model and is fine-tuned using the parameter-efficient LoRA technique.

Inference. Inference is performed on test sets differing in region and viewpoint from the training set. Input size remains (512, 512). For the coarse pose selection stage, the query and rendered candidate mask sizes are adapted to each dataset (e.g., 640×360) with a 10m sampling interval. In the fine pose estimation stage, the mask sizes are identical. Consistent with LoD-Loc v2 [94], we set the number of iterations to 40, using 2 beams and an angular perturbation of $2°$. The translational perturbation follows a Gaussian dis-

Method		*Grid-Traj.*				*Sequence-Traj.*			
		2m-2°	3m-3°	5m-5°	T.e./R.e.	2m-2°	3m-3°	5m-5°	T.e./R.e.
Prior		0.10	1.70	8.90	8.03/1.78	0.30	0.80	8.80	8.19//1.76
CAD-Loc	*same**	**0**	**0**	**0**	**-**	**0**	**0**	**0**	**-**
MC-Loc	DINOv2	0	0.26	1.61	31.95/1.88	3.42	10.50	29.81	5.16/0.49
	RoMa	0	0.14	0.39	32.36/1.89	3.58	10.72	31.12	5.16/0.48
LoD-Loc	-	10.52	20.95	33.80	8.03/1.78	0	0	5.06	8.19/1.76
LoD-Loc v2	no refine	4.00	12.50	39.70	5.91/1.15	9.00	23.10	57.40	4.56/0.91
	no select	2.50	7.50	23.50	8.04/1.56	2.90	7.30	24.60	8.53/1.43
	Full	2.70	8.10	22.70	7.86/1.48	2.30	6.80	24.00	8.75/1.52
LoD-Loc v3$_c$	no refine	7.90	21.60	51.70	4.89/0.98	18.60	38.20	75.60	3.58/0.72
	no select	35.00	62.00	83.40	2.53/0.30	41.60	69.10	89.80	2.24/0.28
	Full	**39.30**	**68.00**	**89.90**	**2.29/0.27**	**50.30**	**79.90**	**97.30**	**1.98/0.23**
LoD-Loc v3$_a$	no refine	7.90	21.60	51.70	4.89/0.98	18.60	38.20	75.60	3.58/0.72
	no select	35.50	61.70	85.10	2.49/0.29	42.30	71.70	92.70	2.21/0.27
	Full	**38.10**	**65.40**	**86.40**	**2.42/0.27**	**49.80**	**79.90**	**95.80**	**2.00/0.23**

Table 4. **Quantitative comparison results of different methods over Tokyo-LoDv3 dataset.** The *same** indicates that the statistics are identical to those in Tab. 2. T.e. and R.e. have the same meanings as in Tab. 2. LoD-Loc v3$_c$ method utilizes confidence-based weighting, and LoD-Loc v3$_a$ method utilizes area-based weighting.

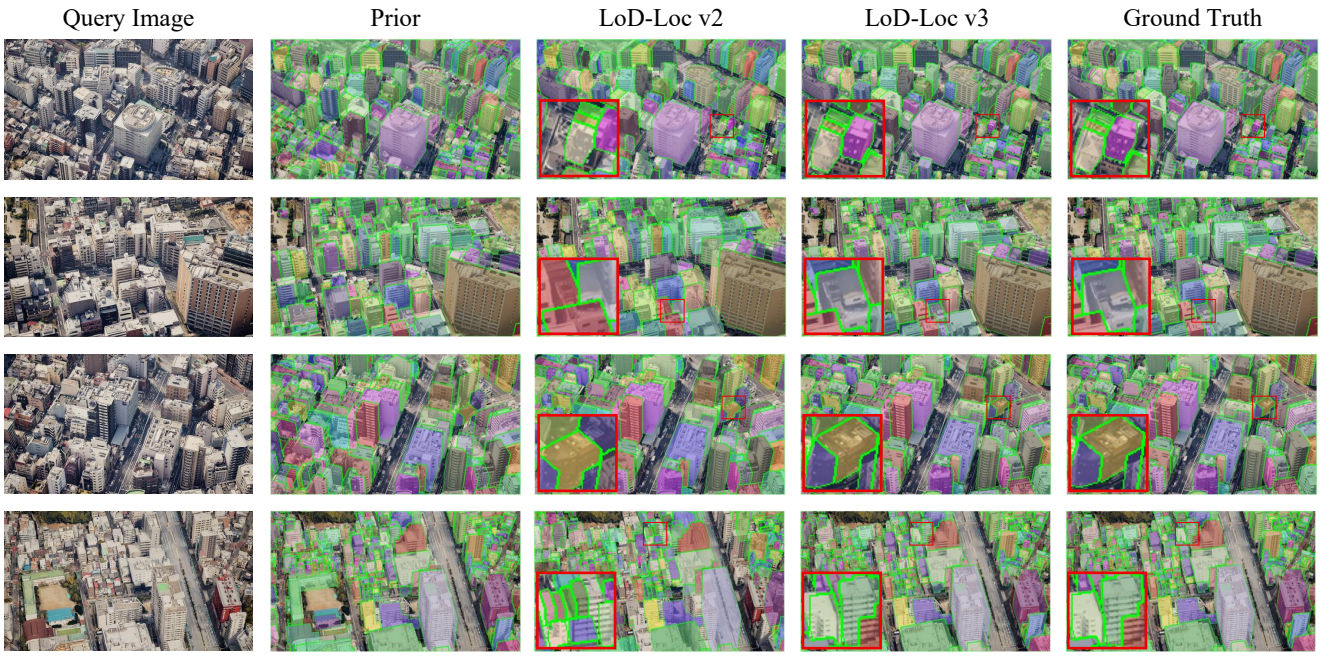

Figure 4. **Visualization of localization results on the Tokyo-LoDv3 dataset.** Superimposed instance masks rendered at estimated poses demonstrate that our method effectively resolves ambiguity in dense urban scenes. The columns from left to right show: query image, prior pose, LoD-Loc v2, LoD-Loc v3 (Ours), and ground truth.

tribution $X \sim \mathcal{N}(0, \sigma^2)$ with a mean of 0 and $\sigma = 1.5$. The decay factor is set to $\gamma = 0.3$, and $n = 52$ candidates are generated per iteration. All training and testing were performed on two NVIDIA RTX 4090 GPUs.

5.3. Evaluation Results

In this section, we compare our results with SOTA visual localization frameworks to validate the superiority of our approach. We also provide visualizations to highlight key aspects and present ablation studies.

Dataset	Method	*in (Grid)-Traj. (Place.)*				*out-of (Sequence)-Traj. (Place.)*			
		2m-2°	3m-3°	5m-5°	T.e./R.e.	2m-2°	3m-3°	5m-5°	T.e./R.e.
UAVD4L-LoDv2	LoD-Loc v2	71.40	91.20	96.10	1.66/0.34	70.00	89.80	96.10	1.62/0.39
	LoD-Loc v3	**97.60**	**98.90**	**99.70**	**0.49/0.13**	**97.40**	**99.00**	**99.40**	**0.60/0.12**
Swiss-EPFLv2	LoD-Loc v2	11.90	28.10	52.00	4.79/1.42	24.00	49.30	77.80	3.02/1.12
	LoD-Loc v3	**58.60**	**79.90**	**95.40**	**1.61/0.77**	**36.90**	**64.10**	**88.90**	**2.43/0.83**
Tokyo-LoDv3	LoD-Loc v2	22.70	43.80	74.70	2.91/0.32	35.60	65.70	92.00	2.41/0.27
	LoD-Loc v3	**39.30**	**68.00**	**89.90**	**2.29/0.27**	**50.30**	**79.90**	**97.30**	**1.98/0.23**

Table 5. **Ablation on Silhouette Representation.** Quantitative comparison of LoD-Loc v2 (semantic) and LoD-Loc v3 (instance) trained on the same InsLoD-Loc dataset.

Alignment	*Grid-Traj.*	*Sequence-Traj*
	2m-2°/3m-3°/3m-3°	2m-2°/3m-3°/3m-3°
LoD-Loc v2	19.60/39.40/72.10	21.50/47.80/89.00
LoD-Loc v3	38.10/65.40/86.40	49.80/79.90/95.80

Table 6. **Ablation on Instance Alignment.** Quantitative comparison of localization performance using separated instance masks versus artificially merged semantic masks.

Evaluation over UAVD4L-LoDv2 dataset. As shown in Tab. 2, our method performs excellently in both *in-Traj.* and *out-of-Traj.* queries, achieving highly competitive performance compared to LoD-Loc v2 [94]. Crucially, LoD-Loc v2 was trained in-distribution, whereas LoD-Loc v3 was trained *only* on our synthetic **InsLoD-Loc** dataset, demonstrating strong cross-domain generalization.

Evaluation over Swiss-EPFLv2 dataset. Results on Swiss-EPFLv2, as shown in Tab. 3, further confirm this. Our method significantly surpasses all baselines, including the in-distribution-trained LoD-Loc v2, across all metrics, despite our model relying solely on synthetic data.

Evaluation over Tokyo-LoDv3 dataset. The Tokyo-LoDv3 dataset features five challenging dense urban scenes. As shown in Tab. 4, LoD-Loc v3 achieves SOTA accuracy in both *Grid-Traj.* and *Sequence-Traj.* queries. In contrast, LoD-Loc v2, which relies on semantic segmentation, fails completely in these scenes due to ambiguity. This validates our instance-alignment strategy for dense scenes, as qualitatively demonstrated in Fig. 4. We note a slight misalignment in the source LoD models for this region, which affects absolute accuracy but not the comparative result.

5.4. Ablation Study

We conducted ablation studies on the UAVD4L-LoDv2, Swiss-EPFLv2, and Tokyo-LoDv3 datasets, focusing on the silhouette representation and alignment methods.

Silhouette Representation. To isolate the performance gains brought by our instance-level representation from the benefits of the new dataset, we retrained the LoD-Loc v2 semantic segmentation model on our **InsLoD-Loc** dataset. Evaluated on the three aforementioned datasets, the retrained LoD-Loc v2 remains significantly inferior to LoD-Loc v3, as shown in Tab. 5. This persistent performance gap confirms that our superior generalization stems fundamentally from the paradigm shift to instance-level representation, rather than merely from scaling up the training data.

Instance Alignment. To explicitly validate the efficacy of the instance alignment mechanism, we conducted an ablation where our predicted instance masks were forcefully merged into a single semantic mask during the pose evaluation stage. As shown in Tab. 6, we validate the superiority of instance masks in Tokyo-LoDv3 datasets, demonstrating that instance alignment is crucial for resolving semantic ambiguities. Further details on the visualization of results and additional ablation studies are provided in the Appendix.

6. Conclusion

This paper presents LoD-Loc v3, a novel aerial localization framework that effectively tackles the challenges of cross-scene generalization and localization ambiguity in dense urban environments. Our solution is twofold: a large-scale synthetic dataset, InsLoD-Loc, enables the learning of domain-invariant features enhancing model generalization, while a novel paradigm, shifting from semantic to instance-level silhouette alignment, effectively resolves ambiguities caused by merged structures in dense cities. Extensive experiments demonstrate that our method achieves SOTA cross-domain performance and succeeds in complex scenes where prior methods fail, significantly advancing the practicality of global-scale aerial localization.

Limitations. Our method's performance is partially dependent on the accuracy of the instance segmentation. Segmentation failures, particularly under extreme adverse weather conditions, may lead to degraded localization accuracy. Further details on the localization results under extreme weather are provided in the Appendix.

Acknowledgments

The authors would like to acknowledge the support from the National Natural Science Foundation of China, Grant No. 62406331.

References

[1] Dhruv Agarwal, Taci Kucukpinar, Joshua Fraser, Jeffrey Kerley, Andrew R Buck, Derek T Anderson, and Kannappan Palaniappan. Simulating city-scale aerial data collection using Unreal Engine. In *Proceedings of the IEEE Applied Imagery Pattern Recognition Workshop*, pages 1–9, 2023. 3

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, 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. 2

[3] Vivian Balakrishnan. Speech by dr vivian balakrishnan, minister-in-charge of the smart nation initiative, at geo connect asia 2021, 2021. https://www.smartnation.gov.sg/media-hub/speeches/geo-connect-asia-2021/. 2

[4] Daniel Barath, Jiri Matas, and Jana Noskova. MAGSAC: marginalizing sample consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10197–10205, 2019. 2

[5] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT: Real-time instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9157–9166, 2019. 3

[6] BuildZero. Open City Model, 2025. https://aws.amazon.com/marketplace/pp/prodview-q5gmfev7upw7e#resources. 2

[7] Yiming Cai, Yao Zhou, Hongwen Zhang, Yuli Xia, Peng Qiao, and Junsuo Zhao. Review of target geo-location algorithms for aerial remote sensing cameras without control points. *Applied Sciences*, 12(24):12689, 2022. 2

[8] CesiumGS. Cesium GS, 2024. https://cesium.com/platform. 2, 4

[9] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. BlendMask: Top-down meets bottom-up for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8573–8581, 2020. 3

[10] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019. 3

[11] Keyan Chen, Chenyang Liu, Hao Chen, Haotian Zhang, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024. 3, 4

[12] Shuxiao Chen, Xiangyu Wu, Mark W Mueller, and Koushil Sreenath. Real-time geo-localization using satellite imagery and topography for unmanned aerial vehicles. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2275–2281, 2021. 2

[13] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. SAM fails to segment anything?–SAM-Adapter: Adapting SAM in underperformed scenes: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2304.09148*, 2023. 3

[14] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. TensorMask: A foundation for dense object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2061–2069, 2019. 3

[15] Xi Chen, Lin Gao, Maojun Zhang, Chen Chen, and Shen Yan. Spectral–spatial adversarial multidomain synthesis network for cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 3

[16] Ondřej Chum and Jiří Matas. Optimal randomized RANSAC. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1472–1482, 2008. 2

[17] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized RANSAC. In *Proceedings of the DAGM Symposium*, pages 236–243, 2003. 2

[18] Delft University of Technology. 3D BAG, 2025. https://3d.bk.tudelft.nl/projects/3dbag/. 2, 4

[19] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018. 2, 5

[20] Ecopia AI. Vividfeatures, 2013. https://www.ecopiatech.com/. 2

[21] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 5

[22] Federal Office of Topography swisstopo. Vision and strategic fields of action 2025, 2021. https://www.swisstopo.admin.ch/en/vision-and-strategic-fields-of-action-2025. 2

[23] Martin A Fischler and Robert C Bolles. Random Sample Consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[24] Victor Fragoso, Joseph DeGol, and Gang Hua. gDLS*: Generalized pose-and-scale estimation given scale and gravity priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2210–2219, 2020. 3

[25] Epic Games. Unreal Engine, 2025. https://www.unrealengine.com/en-US. 2, 4

[26] Ke Gao, Hadi Aliakbarpour, Joshua Fraser, Koundinya Nouduri, Filiz Bunyak, Ricky Massaro, Guna Seetharaman, and Kannappan Palaniappan. Local feature performance evaluation for structure-from-motion and multi-view stereo

using simulated city-scale aerial imagery. *IEEE Sensors Journal*, 21(10):11615–11627, 2020. 3

[27] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *Proceedings of the European Conference on Computer Vision*, pages 369–386, 2020. 2

[28] Gerhard Gröger, Thomas H Kolbe, Claus Nagel, and Karl-Heinz Häfele. OGC city geography markup language (CityGML) encoding standard. *Open Geospatial Consortium: Wayland*, 2012. 2

[29] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3

[30] Bert M Haralick, Chung-Nan Lee, Karsten Ottenberg, and Michael Nölle. Review and analysis of solutions of the three point perspective pose estimation problem. *International Journal of Computer Vision*, 13:331–356, 1994. 2

[31] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2961–2969, 2017. 3

[32] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, 2022. 4

[33] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring R-CNN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6409–6418, 2019. 3

[34] Martin Humenberger, Yohann Cabon, Noé Pion, Philippe Weinzaepfel, Donghwan Lee, Nicolas Guérin, Torsten Sattler, and Gabriela Csurka. Investigating the role of image retrieval for visual localization: An exhaustive benchmark. *International Journal of Computer Vision*, 130(7):1811–1836, 2022. 3

[35] Google Inc. GoogleEarth, 2024. https://earth.google.com/web/. 2, 4

[36] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *Proceedings of the International Conference on MultiMedia Modeling*, pages 451–462, 2020. 3

[37] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2018. 3

[38] Peter Karkus, David Hsu, and Wee Sun Lee. Particle filter networks with application to visual localization. In *Proceedings of the Conference on Robot Learning*, pages 169–178, 2018. 2, 3

[39] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. In *Advances in Neural Information Processing Systems*, 2024. 3

[40] 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. 3

[41] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9799–9808, 2020. 3

[42] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3, 4

[43] Laurent Kneip, Davide Scaramuzza, and Roland Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2976, 2011. 2

[44] Tatjana Kutzner, Carl Smyth, Claus Nagel, Volker Coors, Diego Vinasco-Alvarez, Nobuhiro Ishimaru, Zhihang Yao, Charles Heazel, and Thomas H Kolbe. OGC city geography markup language (CityGML) version 3.0 part 2: GML encoding standard, 2023. https://docs.ogc.org/is/21-006r2/21-006r2.html. 2

[45] Youngwan Lee and Jongyoul Park. CenterMask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2020. 3

[46] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An accurate O(n) solution to the PnP problem. *International Journal of Computer Vision*, 81(2):155–166, 2009. 2

[47] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2015. 3

[48] Yunzhi Lin, Thomas Müller, Jonathan Tremblay, Bowen Wen, Stephen Tyree, Alex Evans, Patricio A Vela, and Stan Birchfield. Parallel inversion of neural radiance fields for robust pose estimation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 9377–9384, 2023. 3

[49] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. 2, 5

[50] Yuncheng Lu, Zhucun Xue, Gui-Song Xia, and Liangpei Zhang. A survey on vision-based UAV navigation. *GeoSpatial Information Science*, 21(1):21–32, 2018. 2

[51] Simon Lynen, Bernhard Zeisl, Dror Aiger, Michael Bosse, Joel Hesch, Marc Pollefeys, Roland Siegwart, and Torsten Sattler. Large-scale, real-time visual–inertial localization revisited. *The International Journal of Robotics Research*, 39 (9):1061–1084, 2020. 3

[52] Dominic Maggio, Marcus Abate, Jingnan Shi, Courtney Mario, and Luca Carlone. Loc-NeRF: Monte Carlo local-

ization using neural radiance fields. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 4018–4025, 2023. 3

[53] Nikolaus Mayer, Eddy Ilg, Philipp Fischer, Caner Hazirbas, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. What makes good synthetic training data for learning disparity and optical flow estimation? *International Journal of Computer Vision*, 126(9):942–960, 2018. 3

[54] Transport Ministry of Land, Infrastructure and Urban Policy Division Tourism. PLATEAU (3D City Model), 2020. https://www.geospatial.jp/ckan/dataset/plateau. 2, 4

[55] Transport Ministry of Land, Infrastructure and Urban Policy Division Tourism. Preparation of 3D City Models: Introduction of Disaster-Related Initiatives in the 'Project PLATEAU' Open Data and Utilization Project, 2025. https://www.bousai.go.jp/kohou/kouhoubousai/r03/102/news_05.html. 2

[56] Ministry of Natural Resources of the People's Republic of China. Notice on Fully Advancing the Constrtuction of Real Scene 3D China, 2022. https://www.csgpc.org/detail/9631.html. 2

[57] Swiss Federal Office of Topography. swissBUILDINGS3D, 2025. https://www.swisstopo.admin.ch/en/landscape-model-swissbuildings3d-2-0. 2, 4

[58] OpenSceneGraph Project. OpenSceneGraph, 1999. http://www.openscenegraph.com/. 2, 4

[59] 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. 5

[60] Vojtech Panek, Zuzana Kukelova, and Torsten Sattler. MeshLoc: Mesh-based visual localization. In *Proceedings of the European Conference on Computer Vision*, pages 589–609, 2022. 2

[61] Vojtech Panek, Zuzana Kukelova, and Torsten Sattler. Visual localization using imperfect 3D models from the internet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13175–13186, 2023. 5, 6

[62] Maxime Pietrantoni, Martin Humenberger, Torsten Sattler, and Gabriela Csurka. SegLoc: Learning segmentation-based representations for privacy-preserving visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15380–15391, 2023. 3

[63] Srikumar Ramalingam, Sofien Bouaziz, Peter Sturm, and Matthew Brand. Skyline2GPS: Localization in urban canyons using omni-skylines. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3816–3823, 2010. 2

[64] Srikumar Ramalingam, Sofien Bouaziz, and Peter Sturm. Pose estimation using both points and lines for geo-localization. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 4716–4723, 2011. 2

[65] Microsoft Research. Microsoft Research, 2025. https://www.microsoft.com/en-us/research/. 2, 4

[66] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 2

[67] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020. 2, 5

[68] Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulo, Richard Newcombe, Peter Kontschieder, and Vasileios Balntas. OrienterNet: Visual localization in 2D public maps with neural matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21632–21642, 2023. 3

[69] Paul-Edouard Sarlin, Eduard Trulls, Marc Pollefeys, Jan Hosang, and Simon Lynen. Snap: Self-supervised neural maps for visual positioning and semantic understanding. In *Advances in Neural Information Processing Systems*, 2024. 3

[70] Olivier Saurer, Georges Baatz, Kevin Köser, L'ubor Ladickỳ, and Marc Pollefeys. Image based geo-localization in the Alps. *International Journal of Computer Vision*, 116(3): 213–225, 2016. 2

[71] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-Motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 2

[72] Senatsverwaltung für Stadtentwicklung, Bauen und Wohnen. Das 3D-Modell, 2023. https://www.berlin.de/sen/sbw/stadtdaten/stadtwissen/digitale-innenstadt/3d-modell/. 2

[73] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Jing Li, Xiangyu Zhang, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8425–8434, 2019. 3

[74] Korsuk Sirinukunwattana, David RJ Snead, and Nasir M Rajpoot. A stochastic polygons model for glandular structures in colon histology images. *IEEE Transactions on Medical Imaging*, 34(11):2366–2378, 2015. 3

[75] Smart Nation and Digital Government Office, Singapore. Integrated environmental modeller, 2019. https://www.smartnation.gov.sg/initiatives/integrated-environmental-modeller/. 2

[76] Hao Su, Shunjun Wei, Min Yan, Chen Wang, Jun Shi, and Xiaoling Zhang. Object detection and instance segmentation in remote sensing imagery based on precise Mask R-CNN. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, pages 1454–1457, 2019. 3

[77] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 2, 5

[78] Xian Sun, Xingliang Huang, Yongqiang Mao, Taowei Sheng, Jihao Li, Zhirui Wang, Xue Lu, Xiaoliang Ma, Deke Tang, and Kaiqiang Chen. GABLE: A first fine-grained 3D building model of China on a national scale from very high resolution satellite imagery. *Remote Sensing of Environment*, 305:114057, 2024. 2

[79] Carl Toft, Will Maddern, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, et al. Long-term visual localization revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2074–2088, 2020. 6

[80] Gabriele Trivigno, Carlo Masone, Barbara Caputo, and Torsten Sattler. The unreasonable effectiveness of pretrained features for camera pose refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12798, 2024. 3, 5, 6

[81] Daniel KD Villa, Alexandre S Brandao, and Mário Sarcinelli-Filho. A survey on load transportation using multirotor UAVs. *Journal of Intelligent & Robotic Systems*, 98: 267–296, 2020. 2

[82] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. PANet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9197–9206, 2019. 3

[83] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. SOLOv2: Dynamic and fast instance segmentation. In *Advances in Neural Information Processing Systems*, pages 17721–17732, 2020. 3

[84] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21666–21675, 2024. 5

[85] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. iSAID: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019. 3

[86] Geospatial World. 3D-Evolution of the Dutch City of Rotterdam, The Netherlands, 2025. https://geospatialworld.net/prime/case-study/aec/3d-evolution-of-the-dutch-city-of-rotterdam-the-netherlands/. 2

[87] Rouwan Wu, Xiaoya Cheng, Juelin Zhu, Yuxiang Liu, Maojun Zhang, and Shen Yan. UAVD4L: A large-scale dataset for UAV 6-DoF localization. In *Proceedings of the International Conference on 3D Vision*, pages 1574–1583, 2024. 2

[88] Olaf Wysocki, Benedikt Schwab, Christof Beil, Christoph Holst, and Thomas H Kolbe. Reviewing open data semantic 3D city models to develop novel 3D reconstruction methods. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48:493–500, 2024. 2

[89] Shen Yan, Xiaoya Cheng, Yuxiang Liu, Juelin Zhu, Rouwan Wu, Yu Liu, and Maojun Zhang. Render-and-compare: Cross-view 6-DoF localization from noisy prior. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 2171–2176, 2023. 2

[90] Shen Yan, Yu Liu, Long Wang, Zehong Shen, Zhen Peng, Haomin Liu, Maojun Zhang, Guofeng Zhang, and Xiaowei Zhou. Long-term visual localization with mobile sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17245–17255, 2023. 2, 3

[91] Amir R Zamir, Tilman Wekel, Pulkit Agrawal, Colin Wei, Jitendra Malik, and Silvio Savarese. Generic 3D representation via pose estimation and matching. In *Proceedings of the European Conference on Computer Vision*, pages 535–553, 2016. 2, 4

[92] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023. 3

[93] Juelin Zhu, Shen Yan, Long Wang, Shengyue Zhang, Yu Liu, and Maojun Zhang. LoD-Loc: Aerial visual localization using LoD 3D map with neural wireframe alignment. In *Advances in Neural Information Processing Systems*, pages 119063–119098, 2024. 2, 5, 6

[94] Juelin Zhu, Shuaibang Peng, Long Wang, Hanlin Tan, Yu Liu, Maojun Zhang, and Shen Yan. LoD-Loc v2: Aerial visual localization over low level-of-detail city models using explicit silhouette alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 26610–26621, 2025. 1, 2, 3, 4, 5, 6, 8

[95] Xiao Xiang Zhu, Sining Chen, Fahong Zhang, Yilei Shi, and Yuanyuan Wang. GlobalBuildingAtlas: An open global and complete dataset of building polygons, heights and LoD1 3D models. *arXiv preprint arXiv:2506.04106*, 2025. 2