

SemCityLoc: Aerial 6DoF Localization Using Semantic 3D City Models

Jingfeng Mao^{1,2} , Xuyang Chen¹ , Qilin Zhang^{1,2,4,5} , Oussema Dhaouadi^{1,2,4,6} , Guangming Wang² , Brian Sheil² , Daniel Cremers^{1,4} , Yan Xia^{3*} , and Olaf Wysocki² 

¹ Technical University of Munich, Munich, Germany

² CV4DT, University of Cambridge, Cambridge, United Kingdom

³ University of Science and Technology of China, Hefei, China

⁴ Munich Center for Machine Learning, Munich, Germany

⁵ Karlsruhe Institute of Technology, Karlsruhe, Germany

⁶ ETH Zurich, Zurich, Switzerland

jingfeng.mao@tum.de yan.xia@ustc.edu.cn

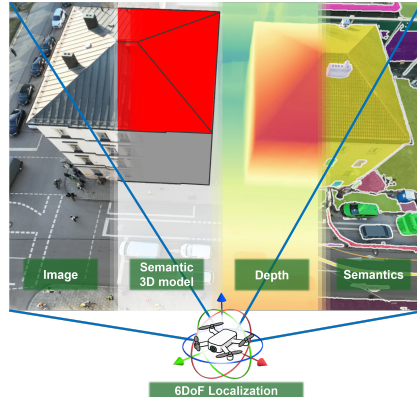


Fig. 1: SemCityLoc. We estimate 6DoF UAV poses by aligning foundation-model semantics and depth with lightweight, semantically structured 3D city models—without radiometric scene reconstructions. We further introduce **SemCityLockeD**, a centimeter-accurate benchmark combining standardized LoD city models with challenging low-altitude UAV imagery.

Abstract. Aerial 6DoF localization typically relies on precise GNSS signals or radiometrically rich 3D reconstructions, limiting scalability and on-board deployment. We propose **SemCityLoc**, a semantic–geometric alignment system that reframes aerial pose estimation as structured surface registration between foundation-model-derived visual priors and standardized LoD-compliant 3D city models. Instead of matching sparse contours or dense texture, our method aligns semantic surfaces and monocular depth with lightweight semantic 3D building models, increasing pose

* Corresponding author

discriminability in repetitive and occluded urban environments. To enable accurate evaluation, we introduce **SemCityLocked**, the first real-world benchmark combining centimeter-accurate UAV poses with standardized LoD1–LoD3 semantic city models and challenging low-altitude imagery. Experiments demonstrate substantial improvements over existing map-based approaches, improving recall by up to 36% and reducing mean positional error from 9.89 m to 2.62 m in challenging urban canyons. Our results indicate that semantically structured geometry provides sufficient and scalable constraints for high-precision aerial localization without radiometric scene reconstructions. The code and data are available at <https://albertchen98.github.io/SemCityLoc>.

Keywords: 6DoF Aerial Pose Estimation · Structured 3D City Models · Map-based Visual Localization System

1 Introduction

Aerial localization for uncrewed aerial vehicles (UAVs) is essential for urban navigation, inspection, mapping, and digital twinning [3, 7, 10, 21, 48]. Accurate six degrees-of-freedom (6DoF) pose estimation enables operation in global navigation satellite systems (GNSS)-denied environments and supports downstream tasks such as reconstruction and change detection [32, 42]. Compared to ground-level localization, aerial imagery exhibits larger baselines, reduced structural detail, and dominant vanishing directions, resulting in weaker geometric constraints [4, 10, 28, 33, 43, 45].

Conventional aerial localization approaches combine real-time kinematic (RTK)-supported sensing with dense textured 3D meshes [21, 22]. While accurate, such radiometrically rich models are memory-intensive, computationally demanding, and difficult to scale, and may raise privacy concerns in urban environments [45, 47, 50]. Lightweight alternatives are therefore desirable.

Recent work has explored reducing reliance on dense radiometry. OrthoLoC [7] aligns UAV imagery with orthophotos or digital surface models (DSMs) but remains 2D texture-dependent. LoD-Loc v1 and v2 [49, 50] demonstrate localization using abstract building models, yet rely primarily on contour-based alignment and scene-specific training, limiting robustness in close-range urban scenarios [49, 50].

Progress is further constrained by benchmark limitations: Many datasets focus on high-altitude flights, provide inconsistent pose accuracy, partially rely on synthetic imagery, or lack standardized level of detail (LoD) definitions despite established CityGML standards [7, 12–14, 19, 39, 44]. As such, systematic evaluation in realistic low-altitude urban canyon settings remains limited.

We argue that semantically structured 3D city models provide sufficient and scalable constraints for high-precision aerial localization when combined with modern vision priors. We present **SemCityLoc**, a semantic-geometric alignment system for aerial 6DoF pose estimation. Our method leverages foundation-model-based semantic segmentation and monocular depth and aligns these cues with lightweight, standardized semantic 3D city models (Fig. 1). By integrating semantic surface classes with geometric structure, SemCityLoc avoids dense radiometric

reconstructions while maintaining strong pose observability, enabling scalable and privacy-preserving localization in GNSS-denied urban environments.

To enable rigorous evaluation in this setting, we introduce **SemCityLockeD**, a real-world benchmark pairing centimeter-accurate UAV poses with standardized LoD semantic 3D city models and challenging low-altitude imagery. Unlike prior datasets, it enables accurate evaluation across LoD-compliant definitions and close-range urban canyon scenarios. Our contributions are as follows:

- **Semantic–geometric aerial localization.** We propose SemCityLoc, an approach that reformulates aerial localization as structured semantic–geometric surface alignment between foundation-model predictions and lightweight LoD city models.
- **Real-world benchmark.** We present SemCityLockeD, a challenging dataset combining centimeter-accurate UAV poses, close-range urban imagery, and standardized LoD1–LoD3 semantic city models for accurate evaluation of map-based aerial localization.
- **Extensive evaluation.** We demonstrate consistent improvements over state-of-the-art map-based localization approaches across our and other benchmarks and challenging urban scenarios.

2 Related Work

Standardized LoD Definitions and Semantic 3D City Models. Level-of-Detail (LoD) definitions in semantic 3D city models follow the international CityGML 3.0 standard [5, 9, 14], illustrated in Figure 2. CityGML encodes hierarchical geometric and semantic information using surface boundary representations, where surfaces (e.g., roof, wall) are explicitly classified. LoD1–LoD3 progressively refine structural detail while preserving standardized georeferencing in global coordinate systems. Governmental CityGML datasets adhere to strict quality

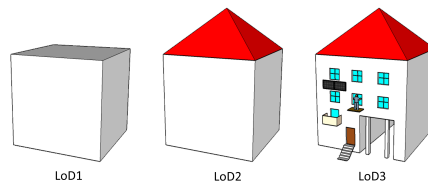


Fig. 2: CityGML Levels of Detail (LoDs). Hierarchical semantic 3D building representations following the CityGML standard [5, 9, 14, 40]. LoD1 models represents simplified building volumes, LoD2 adds detailed roof structures, and LoD3 further incorporates facade elements. Surfaces are semantically classified, enabling structured semantic–geometric reasoning.

standards and often provide centimeter-level geometric accuracy [6, 27]. Recent surveys report more than 216 million openly available semantic 3D building

models worldwide [40, 41]. Their standardized semantics, structured geometry, and global georeferencing make them a scalable and privacy-preserving alternative to radiometric scene reconstructions.

Aerial Localization Benchmarks. Existing UAV localization benchmarks are limited in combining real-world imagery, accurate ground-truth poses, semantic 3D models, and standardized LoD definitions (Tab. 1). Many rely partially or entirely on synthetic data, lack centimeter-accurate poses, or focus on higher-altitude acquisitions (more than 75 m). OrthoLoC [7] provides real-world imagery

Table 1: Comparison of urban UAV benchmark datasets for 6DoF aerial localization. Our SemCityLockeD dataset is the first to combine real-world imagery, accurate RTK-supported poses, close-range challenges, semantic 3D city models, and standardized LoD representations in a single benchmark. ✓ = available/supported; ✗ = not available; ~ = partially available.

Name	Year	View	Real world?	Accurate poses? (GCP/RTK-supported)	Close-range challenge?	Semantic 3D city models?	Standardized LoD?
CrossLoc [44]	2022	oblique + nadir	mixed	✗	✗	✗	✗
MatrixCity [19]	2023	oblique	synthetic	~	~	✗	✗
UAVD4L [39]	2024	oblique + nadir	mixed	✗	✗	✗	✗
Swiss-EPFL [50]	2024	oblique + nadir	mixed	✗	✗	✓	✗
UAVD4L-LoD [50]	2024	oblique + nadir	mixed	✗	✗	✓	✗
GTA-UAV [12]	2025	nadir	synthetic	~	~	✗	✗
OrthoLoC [7]	2025	oblique + nadir	real	✓	~	✗	✗
SemCityLockeD (ours)	2025	oblique + nadir	real	✓	✓	✓	✓

with RTK-supported poses but does not include semantic 3D models. Swiss-EPFL and UAVD4L-LoD [50] introduce image-LoD pairs, yet their LoD representations do not strictly follow established CityGML standards and pose inconsistencies have been reported in prior works [7]. While updated versions have been announced [49], they are not publicly available at the time of writing. Furthermore, most datasets underrepresent close-range urban canyon environments. In contrast, our SemCityLockeD provides exclusively real-world low-altitude urban canyon scenarios with cm-grade standard-compliant 3D LoD models and pose pairings, enabling accurate testing in challenging scenarios.

Map-Based and Semantic Aerial Localization. Classical visual localization relies on dense SfM reconstructions [30], followed by feature matching and PnP-based pose estimation [2, 28]. While accurate, these approaches depend on radiometrically rich and memory-intensive scene models that are difficult to scale in urban environments. Reconstruction-free methods leverage publicly available map data. OrienterNet [29] and MapLocNet [38] use navigation maps but are limited to 3-DoF. OrthoLoC [7] achieves 6-DoF via cross-modal matching with orthophotos and DSMs. On the other hand, lightweight LoD models provide a promising alternative. LoD-Loc [50] aligns neural wireframes with projected building edges, and LoD-Loc v2 [49] extends this via silhouette alignment. However, these approaches rely primarily on contour cues without explicitly incorporating surface semantics or depth geometry. In contrast, we jointly leverage foundation models of semantic segmentation and monocular depth within standardized semantic 3D city models, enabling robust semantic-geometric alignment beyond

contour-based matching. Recent foundation models have significantly improved segmentation [31] and monocular depth estimation [36]. In our framework, these components serve as modular perceptual priors for the localization.

3 Method

As illustrated in Figure 3, SemCityLoc follows a coarse-to-fine semantic-geometric alignment strategy. Given a query image and an initial pose prior, we first perform a 4D semantic cost-volume search to obtain a coarse pose estimate. We then refine this estimate via joint semantic and depth alignment using a particle-filter optimization scheme to recover the final 6DoF camera pose. **Problem Reformulation.** Our goal is to estimate the 6DoF camera pose

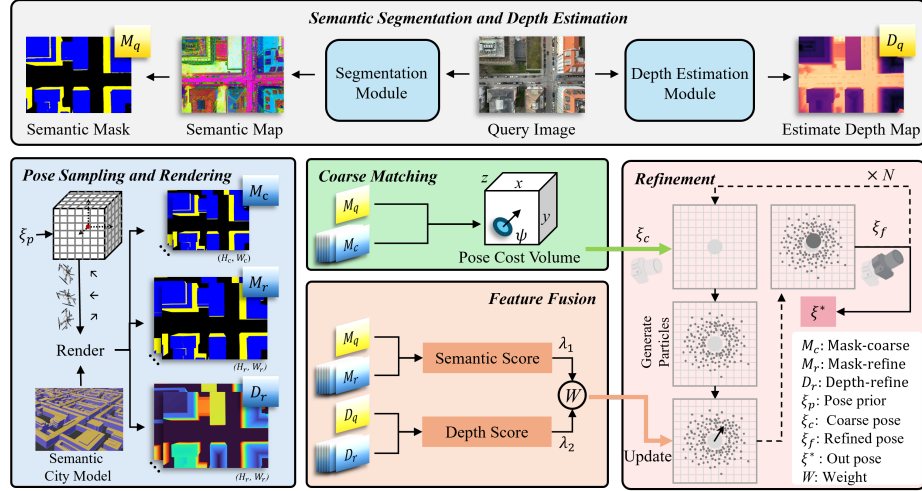


Fig. 3: Overview of SemCityLoc. SemCityLoc estimates the 6DoF camera pose ξ^* of a UAV query image via coarse-to-fine semantic-geometric alignment against a semantic LoD city model. **Inputs** are the query image I_q , a prior pose ξ_p , and the semantic 3D city model. *(Top) Semantic Segmentation and Depth Estimation:* a DINOv3 backbone with a lightweight DPT decoder predicts the query semantic mask M_q , while MoGe-2 predicts the query depth map D_q . *(Bottom-left) Pose Sampling and Rendering:* we uniformly sample candidate poses around the prior ξ_p and render, from the city model, the coarse semantic mask M_c , the fine semantic mask M_r , and the depth map D_r (coarse masks at resolution (H_c, W_c) , fine masks at twice the resolution (H_r, W_r)). *(Coarse Matching):* comparing each rendered mask M_c with M_q through a per-class IoU semantic cost C^s builds a 4D cost volume over (x, y, z, ψ) , whose maximum yields the coarse pose ξ_c . *(Refinement):* starting from ξ_c , a particle filter generates perturbed pose hypotheses; each is scored by *Feature Fusion*, which combines a semantic score (M_q vs. M_r , cost C^s) and a depth score (D_q vs. D_r , cost C^d) into a single weight $W = \lambda_1 C^s + \lambda_2 C^d$. Particles are updated in a weighted manner, and the process repeats for N iterations, refining $\xi_c \rightarrow \xi_f$ and producing the final pose ξ^* .

$\xi^* = (x, y, z, \phi, \theta, \psi)$ where (x, y, z) denote translation and (ϕ, θ, ψ) denote roll, pitch, and yaw. Since roll and pitch can be reliably estimated from gravity measurements, we keep (ϕ, θ) fixed and perform search only over (x, y, z, ψ) . The problem is formulated as follows (Fig. 3): Given a semantic LoD city model $\mathcal{M}$ of a target urban area, a UAV-acquired image I , and an initial coarse camera pose ξ_{pri} , the proposed method aims to determine the accurate 6DoF camera pose ξ^* by jointly leveraging the semantic and geometric features extracted from the image and the 3D semantic city model.

Unlike contour-based alignment, which relies on sparse edge correspondences, semantic surface partitions introduce region-level structural constraints that increase pose observability. By aligning semantically labeled roof and facade surfaces rather than isolated edges, the search space is constrained by both geometric consistency and class-specific spatial arrangement. This reduces ambiguity in repetitive architectural patterns and improves robustness under occlusion and oblique viewing conditions. In essence, semantic alignment transforms the localization problem from edge matching to structured surface registration.

3.1 Coarse Pose Selection

Semantic Segmentation. DINOv3 represents a recently introduced vision foundation model that learns powerful and generalizable global and local representations via self-supervised learning [31]. In semantic segmentation, remarkable performance can be achieved by freezing the pretrained backbone and training a lightweight decoder tailored for the task [11, 46]. Specifically, we employ a pretrained DINOv3 Vision Transformer [31] as a backbone together with a lightweight DPT decoder [25] to predict semantic segmentation masks for the query image I_q :

$$M_q = D_{\theta_d}(F_{\theta_e}(I_q)), \quad (1)$$

$$M_q \in \mathbb{R}^{H \times W \times N_c}, \quad (2)$$

where F_{θ_e} denotes the encoder, D_{θ_d} the decoder, and N_c the number of semantic classes. The segmentation model is treated as a fixed perceptual prior without task-specific architectural modifications.

4D Cost Volume-based Pose Search. We sample the candidate pose ξ^{samp} around the pose prior ξ_{pri} to construct 4D cost volume $\mathcal{C}$. Within this search space, we deploy a 4D search algorithm. Namely, for each sampled camera pose ξ_i^{samp} , we render the corresponding semantic mask M_r^i from LoD city model and compare it with the segmentation mask of the input image M_q . The candidate pose achieving the highest similarity cost is selected as the coarse pose estimate ξ_c . The pose vector ξ_{pri} is represented by a combination of translation and Euler angles:

$$\xi_{pri} = (x_{pri}, y_{pri}, z_{pri}, \phi_{pri}, \theta_{pri}, \psi_{pri}), \quad (3)$$

where, $(x_{pri}, y_{pri}, z_{pri})$ denotes the camera's translation in 3D space, while $(\phi_{pri}, \theta_{pri}, \psi_{pri})$ correspond to the roll, pitch, and yaw angles, respectively. Since

we keep (ϕ, θ) fixed during search, uniform sampling is performed in the 4D space $[x, y, z, \psi]$ centered at the initial pose ξ_{pri} . The sampling range is defined as $\mathbf{e}_c = [e_{cx}, e_{cy}, e_{cz}, e_{c\psi}]$, and the numbers of samples in each dimension are $\mathbf{n}_c = [n_{cx}, n_{cy}, n_{cz}, n_{c\psi}]$, where $\mathbf{e}_c$ denotes the sampling radius in each dimension, and $\mathbf{n}_c$ represents the corresponding number of samples. As a result, a $n_{cx} \times n_{cy} \times n_{cz} \times n_{c\psi}$ discrete 4D pose hypothesis set is generated, which will be used for subsequent cost volume construction and optimal pose selection. The n_{cd} sampled poses along direction $d \in \{x, y, z, \psi\}$ are expressed as:

$$\left\{ \xi_{pri,d} - \frac{e_{cd}}{2}, \dots, \xi_{pri,d} + \frac{e_{cd}}{2} \right\}, \quad (4)$$

For each sampled pose $\xi_i^{\text{samp}} = (\mathbf{R}_i^{\text{samp}}, \mathbf{t}_i^{\text{samp}})$, we obtained the rendered building mask $M_r(\xi_i^{\text{samp}})$ and the semantic building mask M_q , and the semantic cost C_i^s is defined following IoU (Intersection over Union) as:

$$C_i^s = \frac{1}{N_c} \sum_{k=1}^{N_c} \frac{|M_r^k(\xi_i^{\text{samp}}) \cap M_q^k|}{|M_r^k(\xi_i^{\text{samp}}) \cup M_q^k|}. \quad (5)$$

where N_c denotes the number of semantic classes, $\cap$ denotes the pixel-wise intersection of the two binary masks, $\cup$ denotes the pixel-wise union of the two binary masks, and $|\cdot|$ denotes the number of foreground pixels in masks.

Based on semantic cost of each sampled pose ξ_i^{samp} , we construct a four-dimensional cost volume and search the coarse pose ξ_c with maximum value of semantic cost.

3.2 Pose Refinement

Joint Semantic–Depth Alignment. In the fine stage, we employ the model introduced in Section 3.1 to generate semantic masks at twice the spatial resolution. The aforementioned segmentation mask (Sec. 3.1) is extracted from the whole image, which struggles to capture local geometric features. To further improve the pose alignment accuracy, we introduce depth matching using the monocular depth estimator MoGe-2 [36].

Specifically, we utilize a pretrained MoGe-2 model to generate per-pixel depth map D_q for the input UAV image I_q . The estimated depth values must be numerically aligned with the rendered depth map to ensure consistency in depth scale and range. For the depth map $D_r(\xi_i^{\text{samp}})$ rendered at the candidate position ξ_i^{samp} , and estimated depth map D_q , two global scale and shift parameters are estimated for each match, to calculate a robust depth cost:

$$C_i^d = \sum_{p \in \mathcal{M}_{valid}} \frac{1}{D_r(p)} \|s^* D_q(p) + t^* - D_r(p)\|_1, \quad (6)$$

where $\mathcal{M}_{valid}$ is the valid mask of rendered depth map, $D_r(p)$ denotes the rendered depth value at pixel p , and s^* and t^* are the optimal global scale and shift alignment estimated by least squares estimation for each match. The cost

is inversely weighted by rendered depth to reduce the influence of distant, less reliable depth estimates. The final alignment cost combines semantic and depth terms:

$$C_i^{final} = \lambda_1 C_i^s + \lambda_2 C_i^d \quad (7)$$

Particle Filter Optimization. Inspired by other visual localization works [18, 20, 35], we design a particle-filter-based method for camera pose refinement. The core idea is to iteratively refine the initial hypothesis by introducing perturbations to the state, evaluating the resulting cost function, and updating the estimate toward convergence. This stochastic refinement enables local exploration around the coarse hypothesis while avoiding exhaustive high-dimensional grid search, leading to stable and efficient convergence.

Specifically, having the coarse pose ξ_c selected from Sec. 3.1, we introduce small perturbations δ to its translation and rotation parameters, resulting in new candidate poses ξ_{p_i}

$$\{\xi_{p_i}\}_{i=1}^{N_f} = \xi_c + \{\delta\xi_i\}_{i=1}^{N_f}, \quad (8)$$

where $\delta\xi_i$ denotes the perturbation applied to the i -th particle in the pose space. Each perturbed pose ξ_{p_i} is then evaluated by rendering a semantic or geometric observation under that hypothesis and comparing it with the corresponding measurement from the query image using the predefined loss function derived from Eq.(7). The camera pose is then updated in a weighted manner according to the probability scores obtained from the loss function. After performing N_{iter} iterations, we obtain the final optimized pose estimate ξ^* .

4 SemCityLockeD Benchmark Dataset

We introduce **SemCityLockeD**, a novel real-world dataset to address the limitations of existing UAV localization benchmarks that lack accurate ground-truth poses, standardized semantic 3D city models, and realistic close-range urban canyon scenarios (Fig. 4). SemCityLockeD captures a densely built urban environment characterized by closely spaced buildings, narrow urban canyons, and complex structural layouts spanning diverse architecture (19th-21st century) and building functions (residential, offices, shops, landmarks). Data acquisition was performed using a DJI Matrice 350 RTK equipped with a Zenmuse L2 (4/3 CMOS) camera, flown at approximately 75 m altitude in both nadir and oblique configurations. The dataset comprises 962 images (train/val/test: 586/188/188) with a resolution of $5,280 \times 3,956$ pixels, resulting in an average ground sampling distance (GSD) of 1.6 cm. For privacy compliance, license plates and identifiable faces were removed from the imagery. Camera poses were obtained via RTK-GNSS, inertial measurement unit (IMU) integration, and surveyed ground control points (GCPs), yielding georeferenced poses in EPSG:32632 with absolute accuracy of approximately 2 cm. This high-accuracy georeferencing enables reliable pairing between UAV imagery and semantic 3D city models.

The corresponding LoD2 semantic 3D building models were acquired from the open data Bavarian state geoportal [16], providing high absolute georeferenced

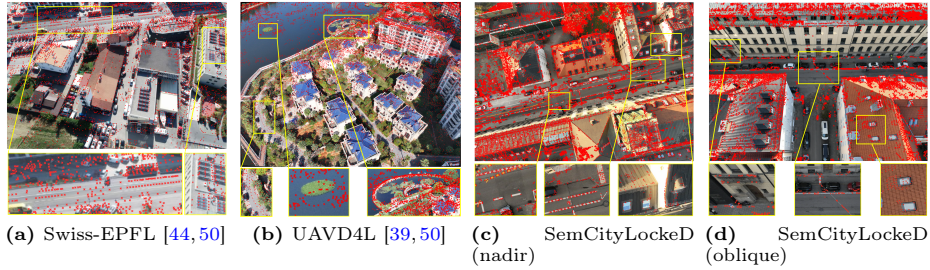


Fig. 4: Pose quality comparison. Georeferenced keypoints are derived from corner detections on high-precision DOPs, lifted to 3D using DSMs, and then projected onto the UAV image using the ground-truth camera poses and intrinsics. In (c) and (d), corresponding to SemCityLockeD, the projected keypoints align closely with image structures, indicating high pose consistency. Note that (a) and (b) correspond to acquisitions at larger flight altitudes (above 75m) and in less dense urban settings, whereas (c) and (d) depict closer-range imagery in narrow urban canyon environments, representing substantially more challenging geometric conditions.

footprint accuracy (approximately 2 cm) [1, 3]. In addition to LoD2, we provide automatically generated LoD1 models and manually refined LoD3 models, enabling systematic evaluation across different LoDs and analysis of localization performance with respect to geometric abstraction, whereby both LoD1 and LoD3 were generated from LoD2 footprints to maintain the high accuracy and consistency. Facade details were modeled based on additional street-level scanning point clouds which are not part of this dataset.

Beyond semantic LoD models, we further provide high-resolution textured mesh reconstructions generated from laser scanning and photogrammetric point clouds of the same UAV acquisition. These meshes complement the dataset, allowing for rendering additional synthetic images, following the protocols of synthetic Swiss-EPFL [44, 50] and UAVD4L [39, 50].

Leveraging the precise pairing between UAV images and 3D city models, we additionally generate automatic semantic labels for the acquired imagery by projecting model semantics into image space. This enables large-scale, geometry-consistent annotation without manual labeling, facilitates supervised learning experiments, and provides a scalable annotation pipeline that can be readily adopted for future UAV datasets equipped with georeferenced 3D models.

In addition to the dataset, we also release a software suite for rendering semantic 3D city models into image and depth maps and pairing them with UAV observations, facilitating reproducibility and standardized evaluation.

5 Experimental Setup

Baselines. We compare our proposed method with re-implemented visual localization baselines using semantic 3D building models with publicly available code: LoD-Loc [50], MC-Loc [35], and CAD-Loc [23] (LoD-Loc v2 implementation unavailable at the time of writing [49]). All baseline methods, except for LoD-Loc,

rely on feature-based extraction and matching approaches, in which the feature extraction algorithms include RoMa [8] or e-LoFTR [37].

Data. Besides evaluating on our introduced challenging scenario of the SemCityLockeD benchmark, we also evaluate on the openly available benchmarks that comprise 3D building models: UAVD4L-LoD and Swiss-EPFL [50] (the LoD-Loc v2 data is unavailable at the time of writing [49]).

Metrics. The evaluation metrics we follow are from the standard visual localization metrics [34] and we set the thresholds at $(2m, 2^\circ)$, $(3m, 3^\circ)$, $(5m, 5^\circ)$, same as the LoD-Loc baseline [50]. In addition, we also compute the average translation and Rotation errors.

6 Results and Discussion

Performance in Challenging Scenarios. We evaluate SemCityLoc on the proposed SemCityLockeD benchmark, which represents densely built urban canyon environments with strong geometric ambiguities. As shown in Table 2, our method consistently outperforms all baselines across all recall thresholds and error metrics. At the strict $2m-2^\circ$ threshold, SemCityLoc achieves 69.15% recall compared to 35.11% for LoD-Loc [50], nearly doubling performance. Recall further improves to 84.04% and 89.36% at $3m-3^\circ$ and $5m-5^\circ$, respectively. Mean rotation error decreases from 1.78° to 0.42° , and translation error from 9.89m to 2.62m. These results demonstrate that integrating semantic segmentation and monocular depth provides substantially stronger geometric constraints than wireframe-based alignment alone, particularly in close-range urban scenes with facade repetitions and limited visibility. Ablation results confirm the importance

Table 2: Camera pose evaluation on the SemCityLockeD dataset. Pose recall is reported in %, rotation error in degrees, and positional errors in meters. Our SemCityLoc significantly outperforms all baselines across all metrics. **Bold** = best result; underlined = second best.

Method		Recall (%)			Mean Error	
Metrics		2m-2°	3m-3°	5m-5°	Rotation (°)	XYZ (m)
CAD-Loc <i>LoD model</i>	e-LoFTR	0	0	0	–	–
	RoMa	0	0	0	–	–
MC-Loc <i>LoD model</i>	DINOv2	5.32	8.51	18.09	5.90	10.71
	RoMa	0	1.06	3.19	6.67	18.61
LoD-Loc <i>LoD model</i>	LoD-Loc	35.11	47.87	53.19	1.78	9.89
SemCityLoc <i>LoD model</i>	<i>No Coarse Selection</i>	41.49	56.38	72.34	1.11	5.94
	<i>No Refinement</i>	<u>56.38</u>	<u>70.21</u>	<u>81.91</u>	1.04	3.20
	<i>Full model</i>	69.15	84.04	89.36	0.42	2.62

of both modules (Tab. 2 and Fig. 5). Coarse semantic selection increases recall at $2m-2^\circ$ from 41.49% to 56.38%, while semantic–depth refinement further improves it to 69.15%. Both stages are therefore complementary and jointly critical for stable pose estimation.

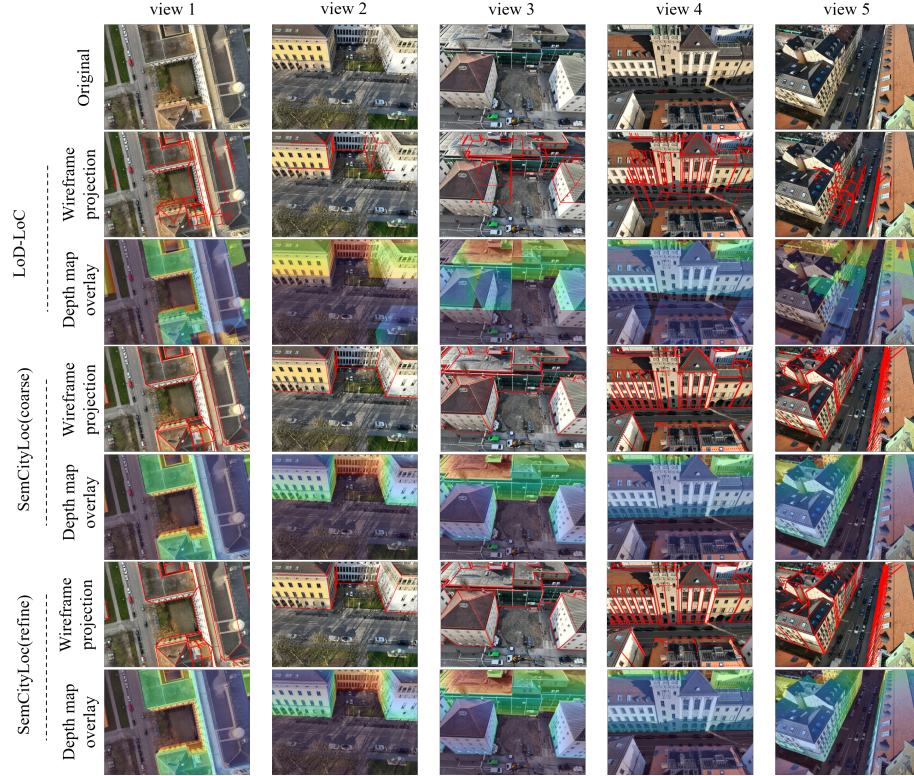


Fig. 5: Qualitative comparison between LoD-Loc [50] and SemCityLoc. Projected LoD geometry using the estimated poses reveals noticeable misalignment for LoD-Loc, especially in oblique views and facade-occluded urban canyon scenes. In contrast, SemCityLoc produces geometrically consistent overlays and accurate depth alignment, demonstrating improved robustness to perspective distortion and structural repetition (especially apparent in view 3).

Qualitative results in Figure 5 further corroborate the improvements. In particular, View 3 highlights a challenging oblique urban canyon configuration with strong perspective distortion and facade self-occlusion, where the baseline fails to produce a consistent geometric alignment. In contrast, SemCityLoc achieves precise overlay between projected LoD geometry and image structures, demonstrating robust semantic-geometric alignment even under severe oblique viewing conditions and repetitive facade patterns.

Efficiency. SemCityLoc achieves markedly improved training efficiency compared with LoD-Loc [50]. As shown in Figure 6, SemCityLoc converges within the first few epochs and stabilizes after approximately 15 epochs, whereas LoD-Loc exhibits slower convergence and pronounced loss oscillations. This behavior can be attributed to two factors. First, SemCityLoc optimizes only a lightweight segmentation head on top of pretrained foundation models, effectively operating

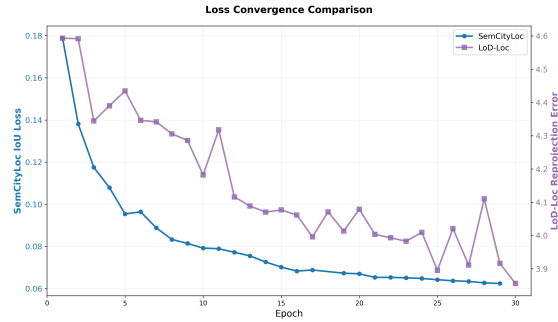


Fig. 6: Training convergence on SemCityLocked. SemCityLoc exhibits faster and more stable optimization compared to LoD-Loc, reflecting the benefit of fine-tuning a lightweight head on pretrained features with IoU-based alignment.

in a well-conditioned representation space. In contrast, LoD-Loc is trained from scratch, requiring simultaneous feature learning and geometric alignment, which increases optimization complexity and susceptibility to local minima. Second, our IoU-based semantic alignment objective induces smoother and more globally structured gradients than reprojection-based losses. Semantic masks provide dense, region-level supervision that is less sensitive to small geometric perturbations, whereas reprojection losses rely on sparse edge correspondences and exhibit higher non-convexity with respect to pose parameters. As a result, IoU-driven optimization yields more stable gradient behavior and faster convergence.

Generalization to Suburban and High-Altitude Scenarios. We further evaluate SemCityLoc on the UAVD4L-LoD and Swiss-EPFL datasets comprising mixed urban scenarios to assess generalization beyond challenging close-range urban canyons. As reported in [50], LoD-Loc achieves strong performance on these benchmarks and is therefore adopted as our primary reference baseline.

As shown in Table 3, SemCityLoc achieves competitive or superior performance compared to LoD-Loc in both in-Traj and out-of-Traj settings on UAVD4L-LoD. In the in-Traj scenario, our method improves most recall thresholds and reduces translation error, while maintaining comparable rotation accuracy. In the out-of-Traj setting, SemCityLoc outperforms the baseline on nearly all metrics, demonstrating improved robustness to trajectory shifts. Results on the Swiss-EPFL dataset (Tab. 4) further confirm the benefits of integrating semantics and depth cues. In the in-Place scenario, recall values are comparable to LoD-Loc, while translation and heading errors are consistently reduced. In the challenging out-of-Place setting, SemCityLoc significantly improves recall (up to two-fold) and substantially reduces mean positional error from 12.54 m to 3.09 m (Tab. 4).

Overall, these results indicate that the proposed semantic-geometric alignment strategy generalizes effectively across heterogeneous environments comprising mixed urban and sub-urban regions, varying structural densities, and diverse acquisition altitudes, while consistently outperforming wireframe-based localization methods.

We note that UAVD4L-LoD and Swiss-EPFL provide lower-precision ground-truth poses compared to SemCityLockeD, which may introduce additional evaluation noise. Nevertheless, consistent improvements across datasets demonstrate the robustness of our approach. Additional details are provided in Suppl.

Table 3: Camera pose evaluation on the UAVD4L-LoD dataset. Pose recall is reported in %, rotation error in degrees, and positional errors in meters. Our SemCityLoc achieves competitive performance, outperforming the baseline in most metrics. **Bold** = best result; underlined = second best.

Method		in-Traj					out-of-Traj				
		Recall (%)			Mean Error		Recall (%)			Mean Error	
		2m-2°	3m-3°	5m-5°	Rotation (°)	XYZ (m)	2m-2°	3m-3°	5m-5°	Rotation (°)	XYZ (m)
LoD-Loc	LoD-Loc	83.42	<u>90.65</u>	<u>96.57</u>	<u>0.19</u>	<u>1.73</u>	<u>88.18</u>	98.49	<u>99.54</u>	<u>0.18</u>	<u>1.38</u>
SemCityLoc	<i>No Coarse Selection</i>	62.03	89.96	<u>99.05</u>	0.20	1.84	47.63	62.32	78.51	0.34	3.36
	<i>No Refinement</i>	47.07	74.88	91.21	0.34	2.48	48.49	83.94	99.13	0.68	2.18
	<i>Full model</i>	<u>70.45</u>	92.96	99.19	0.19	1.68	89.64	<u>98.40</u>	99.95	0.14	1.26

Table 4: Camera pose evaluation on the Swiss-EPFL dataset. Pose recall is reported in %, rotation error in degrees, and positional errors in meters. Our SemCityLoc outperforms the baseline in most metrics, with particularly strong performance on out-of-place scenarios. **Bold** = best result; underlined = second best.

Method		in-Place					out-of-Place				
		Recall (%)			Mean Error		Recall (%)			Mean Error	
		2m-2°	3m-3°	5m-5°	Rotation (°)	XYZ (m)	2m-2°	3m-3°	5m-5°	Rotation (°)	XYZ (m)
LoD-Loc	LoD-Loc	36.23	50.00	63.62	3.73	<u>9.17</u>	<u>17.41</u>	<u>30.61</u>	<u>48.55</u>	3.40	12.54
SemCityLoc	<i>No Coarse Selection</i>	5.34	15.73	31.32	2.15	11.60	6.86	17.15	35.88	1.62	9.05
	<i>No Refinement</i>	10.11	20.51	35.96	<u>2.04</u>	10.39	8.71	19.79	42.22	<u>1.23</u>	<u>7.86</u>
	<i>Full model</i>	<u>33.99</u>	50.70	<u>62.36</u>	1.31	7.13	35.36	62.01	89.18	0.93	3.09

Sensitivity Analysis to Segmentation Quality. We evaluate the impact of semantic quality by training a lightweight head on top of a shared DINOv3 backbone per dataset, achieving 88% mIoU on SemCityLockeD, 85% on UAVD4L-LoD, and 78% on Swiss-EPFL—sufficient for stable semantic-geometric alignment.

Fully zero-shot alternatives perform substantially worse: CLIP [24] + Semantic-SAM [17] yields 32% mIoU on SemCityLockeD and reduces recall to 15.93% / 28.57% / 52.74% at 2m-2° / 3m-3° / 5m-5°, while Grounded-SAM2 [26] achieves 37% mIoU with unstable masks on the same dataset. We attribute this degradation to domain mismatch between terrestrial pretraining data and UAV viewpoints. These results indicate that moderate dataset-adaptive supervision suffices for reliable localization, whereas fully zero-shot semantics remain insufficient (more details in Suppl.).

Sensitivity Analysis to Pose Prior Noise. The translation priors are initially perturbed by 5–15 m, on top of which we inject varying independent perturbation of up to 200 m per axis. Rotations are also perturbed by 0–8° (Tab. 5). The pose

Table 5: Pose prior sensitivity analysis. Localization recall (%) under increasing translational perturbations (up to 200 m per axis).

Metric		Perturbation Range per Axis (m)						
Threshold(Δm)		10–20	20–30	30–40	40–50	50–100	100–150	150–200
Recall (%)	2m–2°	68.09	67.02	65.96	62.77	45.74	29.79	18.09
	3m–3°	80.85	75.53	73.40	70.21	50.00	30.85	18.09
	5m–5°	87.23	81.91	78.72	76.60	53.19	36.17	19.15

prior sensitivity analysis shows that recall degrades smoothly as translational perturbations increase. Even with perturbations up to 50 m per axis, the method maintains relatively stable performance, with recall remaining at 62.77%, 70.21%, and 76.60% under the 2 m–2°, 3 m–3°, and 5 m–5° thresholds, respectively. This demonstrates robustness to substantial initialization errors. However, when the pose prior error exceeds 100 m, localization recall drops dramatically across all thresholds. This sharp decrease indicates that very large prior errors significantly affect localization reliability, revealing an operational boundary of the method.

Sensitivity Analysis Across LoDs. We analyze how structural abstraction in the 3D city model affects pose alignment (Tab. 6). While LoD2 offers a favorable trade-off between availability and fidelity, SemCityLockeD uniquely provides standardized LoD1–LoD3 models [40], enabling systematic evaluation. Performance

Table 6: Localization performance across different LoDs. Increasing geometric detail strengthens pose constraints, while semantic LoD2 achieves performance comparable to LoD3 despite reduced geometric complexity.

LoD Model	Recall (%)			Mean Error	
	2m-2°	3m-3°	5m-5°	Rotation (°)	XYZ (m)
LoD1	48.94	54.26	69.15	1.14	6.01
LoD3	70.21	75.53	89.36	0.50	2.48
LoD2(semantic)	<u>69.15</u>	84.04	89.36	0.42	<u>2.62</u>

improves from LoD1 to LoD3, reflecting increased structural discriminability. Notably, semantic LoD2 achieves accuracy close to or even surpasses that of LoD3 despite reduced geometric resolution. This indicates that semantically partitioned surface structure enhances pose observability and can partially compensate for missing fine-grained geometry. Pose accuracy is therefore governed not only by geometric density, but by the joint availability of discriminative structural cues and semantically coherent surface representations (more details in Suppl.).

Running time analysis. Using nvdiffrast-based [15] GPU rasterization and frustum culling, SemCityLoc requires only 0.878 s per image, including 0.116 s for model inference, 0.409 s for coarse pose search, and 0.353 s for particle-filter refinement. The resulting sub-second runtime highlights the feasibility of deploying semantic city-model-based localization in practical UAV systems, where accurate relocalization must be performed online under large pose uncertainty.

Limitations and Future Work. SemCityLoc relies on the availability and geometric accuracy of semantic 3D city models, and variations in model completeness or semantic richness may affect localization robustness across regions. While the method is generally robust to dynamic objects absent from the LoD models (e.g., vehicles and vegetation), performance may degrade in rare cases with severely limited geometric observability, such as scenes containing only a single visible building or facade. Finally, localization accuracy depends on the quality of semantic segmentation and depth estimation, although continued advances in foundation models are expected to further improve robustness.

7 Conclusion

We present SemCityLoc, a lightweight semantic–geometric localization system that leverages standardized 3D city models for accurate and scalable aerial 6DoF pose estimation. By jointly aligning semantic segmentation and monocular depth with structured LoD representations, our method enables robust localization without dense radiometric reconstructions. We further introduce SemCityLocked, a unique real-world benchmark comprising centimeter-accurate UAV poses, close-range urban canyon imagery, and standardized LoD1–LoD3 models. Across challenging scenarios, SemCityLoc improves recall at 2m–2° from 35.11% to 69.15%, demonstrating substantially stronger pose discriminability than existing map-based approaches while remaining robust to segmentation noise and prior inaccuracies. Overall, our results indicate that semantically structured 3D city models provide sufficient and scalable constraints for high-precision aerial localization, establishing structured semantic–geometric alignment as a promising direction for next-generation map-based aerial localization systems.

Acknowledgements

This work was supported in part by the Cambridge Centre for Smart Infrastructure and Construction (CSIC), the Laing O’Rourke Centre for Construction Engineering and Technology at the University of Cambridge, the Anhui Provincial Natural Science Foundation (Grant No. 2508085MF142), and the USTC Global Initiative Fund (MS-A-2026-1-19) at the University of Science and Technology of China. The authors thank the members of the Computer Vision for Digital Twins (CV4DT) group at the University of Cambridge for their support and valuable discussions. The authors also thank Yuhan Jiang for assistance with figure preparation. The authors further acknowledge the contributors to the TUM2TWIN project <https://tum2t.win/>, in particular Katharina Anders, Jiapan Wang, Xiaoyu Huang, and Sidi Liu, for their efforts in processing and publishing the data <https://zenodo.org/records/14548134>.

References

1. Anders, K., Wang, J., Chang, M., Letard, M., Schulte, F., Winiwarter, L.: Terrestrial and UAV laser scanning point clouds of TUM Campus Ottobrunn (Dec

- 2024). <https://doi.org/10.5281/zenodo.14443336>, <https://doi.org/10.5281/zenodo.14443336>
2. Barath, D., Matas, J., Noskova, J.: Magsac: marginalizing sample consensus. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10197–10205 (2019)
 3. Barbosa, J., Schwab, B., Wysocki, O., Heeramaglore, M., Huang, X.: tum2twin: repository of CityGML LoD3 models of the technical university of munich. <https://github.com/tum-gis/tum2twin/tree/main> (2023), Accessed: 2023-09-22
 4. Bieringer, A., Wysocki, O., Tuttas, S., Hoegner, L., Holst, C.: Analyzing the impact of semantic lod3 building models on image-based vehicle localization. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **10**, 55–62 (2024)
 5. Biljecki, F., Ledoux, H., Stoter, J., Zhao, J.: Formalisation of the level of detail in 3D city modelling. *Computers, Environment and Urban Systems* **48**, 1–15 (2014)
 6. Chen, D., Zhu, C., Zhang, Z., Na, J., Shen, Y., Chen, Y., Peethambaran, J., Zhang, L.: Public building geometric models from point clouds: A multidimensional quality evaluation framework. *IEEE Geoscience and Remote Sensing Magazine* (2025)
 7. Dhaouadi, O., Marin, R., Meier, J.M., Kaiser, J., Cremers, D.: Ortholoc: UAV 6-dof localization and calibration using orthographic geodata. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
 8. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19790–19800 (2024)
 9. Froech, T., Schwab, B., Wysocki, O.: CityGML2OBJ 2.0: Command line converter of citygml (.gml) to obj (.obj) files, while maintaining the semantics. <https://github.com/tum-gis/CityGML2OBJv2> (2023), Accessed: 2023-08-23
 10. Gaisbauer, S., Gyawali, P., Zhang, Q., Wysocki, O., Jutzi, B.: To glue or not to glue? classical vs learned image matching for mobile mapping cameras to textured semantic 3d building models. *arXiv preprint arXiv:2505.17973* (2025)
 11. Huang, S., Hou, Y., Liu, L., Yu, X., Shen, X.: Real-time object detection meets dinov3. *arXiv preprint arXiv:2509.20787* (2025)
 12. Ji, Y., He, B., Tan, Z., Wu, L.: Game4loc: A uav geo-localization benchmark from game data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3913–3921 (2025)
 13. Kolbe, T.H., Gröger, G., Plümer, L.: CityGML–3D city models and their potential for emergency response. *Geospatial Information Technology for Emergency Response* **257**, 273–290 (2008)
 14. Kolbe, T.H., Kutzner, T., Smyth, C.S., Nagel, C., Roensdorf, C., Heazel, C.: OGC City Geography Markup Language (CityGML) Part 1: Conceptual Model Standard v3.0 (Jun 2021)
 15. Laine, S., Hellsten, J., Karras, T., Seol, Y., Lehtinen, J., Aila, T.: Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (ToG)* **39**(6), 1–14 (2020)
 16. LDBV: Das Landesamt für Digitalisierung, Breitband und Vermessung (LDBV). <https://geodaten.bayern.de/opengeodata/OpenDataDetail.html?pn=lod2> (2025), accessed: 2025-06-22
 17. Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Li, C., Yang, J., Zhang, L., Gao, J.: Segment and recognize anything at any granularity. In: European Conference on Computer Vision. pp. 467–484. Springer (2024)

18. Li, X., Wang, S., Zhao, Y., Verbeek, J., Kannala, J.: Hierarchical scene coordinate classification and regression for visual localization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11983–11992 (2020)
19. Li, Y., Jiang, L., Xu, L., Xiangli, Y., Wang, Z., Lin, D., Dai, B.: Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3205–3215 (2023)
20. Maggio, D., Abate, M., Shi, J., Mario, C., Carlone, L.: Loc-nerf: Monte carlo localization using neural radiance fields. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 4018–4025. IEEE (2023)
21. Nex, F., Remondino, F.: Uav for 3d mapping applications: a review. *Applied geomatics* **6**(1), 1–15 (2014)
22. Nex, F., Stathopoulou, E., Remondino, F., Yang, M., Madhuanand, L., Yogender, Y., Alsadik, B., Weinmann, M., Jutzi, B., Qin, R.: Usegeo-a uav-based multi-sensor dataset for geospatial research. *ISPRS Open Journal of Photogrammetry and Remote Sensing* p. 100070 (2024)
23. Panek, V., Kukulova, Z., Sattler, T.: Visual localization using imperfect 3d models from the internet. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13175–13186 (2023)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
25. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 12179–12188 (2021)
26. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159* (2024)
27. Roschlaub, R., Batscheider, J.: An INSPIRE-conform 3D building model of Bavaria using cadastre information, LiDAR and image matching. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XLI-B4**, 747–754 (2016)
28. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 4938–4947 (2020)
29. Sarlin, P.E., DeTone, D., Yang, T.Y., Avetisyan, A., Straub, J., Malisiewicz, T., Buló, S.R., Newcombe, R., Kotschieder, P., Balntas, V.: Orienternet: Visual localization in 2d public maps with neural matching. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 21632–21642 (2023)
30. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4104–4113 (2016)
31. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khaidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. *arXiv preprint arXiv:2508.10104* (2025)
32. Stilla, U., Xu, Y.: Change detection of urban objects using 3d point clouds: A review. *ISPRS Journal of Photogrammetry and Remote Sensing* **197**, 228–255 (2023). <https://doi.org/https://doi.org/10.1016/j.isprsjprs.2023.01.010>

33. Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A.: Inloc: Indoor visual localization with dense matching and view synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7199–7209 (2018)
34. Toft, C., Maddern, W., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Pajdla, T., et al.: Long-term visual localization revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(4), 2074–2088 (2020)
35. Trivigno, G., Masone, C., Caputo, B., Sattler, T.: The unreasonable effectiveness of pre-trained features for camera pose refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12786–12798 (2024)
36. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5261–5271 (2025)
37. Wang, Y., He, X., Peng, S., Tan, D., Zhou, X.: Efficient lofr: Semi-dense local feature matching with sparse-like speed. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21666–21675 (2024)
38. Wu, H., Zhang, Z., Lin, S., Mu, X., Zhao, Q., Yang, M., Qin, T.: Maplocnet: Coarse-to-fine feature registration for visual re-localization in navigation maps. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 13198–13205. IEEE (2024)
39. Wu, R., Cheng, X., Zhu, J., Liu, Y., Zhang, M., Yan, S.: Uavd4l: A large-scale dataset for uav 6-dof localization. In: 2024 International Conference on 3D Vision (3DV). pp. 1574–1583. IEEE (2024)
40. Wysocki, O., Schwab, B., Beil, C., Holst, C., Kolbe, T.H.: 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)
41. Wysocki, O., Schwab, B., Willenborg, B., Knezevic, M.: Awesome CityGML. <https://github.com/01o0cki/awesome-citygml> (2024), Accessed: 2024-01-30
42. Wysocki, O., Xia, Y., Wysocki, M., Grilli, E., Hoegner, L., Cremers, D., Stilla, U.: Scan2LoD3: Reconstructing semantic 3D building models at LoD3 using ray casting and Bayesian networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* pp. 6547–6557 (2023)
43. Xia, Y., Lu, Y., Song, R., Dhaouadi, O., Henriques, J.F., Cremers, D.: Trafficloc: localizing traffic surveillance cameras in 3d scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28685–28695 (2025)
44. Yan, Q., Zheng, J., Reding, S., Li, S., Doytchinov, I.: Crossloc: Scalable aerial localization assisted by multimodal synthetic data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17358–17368 (2022)
45. Yan, S., Cheng, X., Liu, Y., Zhu, J., Wu, R., Liu, Y., Zhang, M.: Render-and-compare: Cross-view 6-dof localization from noisy prior. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 2171–2176. IEEE (2023)
46. Yang, S., Wang, H., Xing, Z., Chen, S., Zhu, L.: Segdino: An efficient design for medical and natural image segmentation with dino-v3. *arXiv preprint arXiv:2509.00833* (2025)
47. Ye, Y., Teng, X., Chen, S., Li, Z., Liu, L., Yu, Q., Tan, T.: Exploring the best way for uav visual localization under low-altitude multi-view observation condition: a benchmark. *arXiv preprint arXiv:2503.10692* (2025)

48. Zhang, Q., Wysocki, O., Jutzi, B.: Gs4buildings: Prior-guided gaussian splatting for 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **10**, 249–256 (2025)
49. Zhu, J., Peng, S., Wang, L., Tan, H., Liu, Y., Zhang, M., Yan, S.: 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*. pp. 26610–26621 (2025)
50. Zhu, J., Yan, S., Wang, L., Zhang, S., Liu, Y., Zhang, M.: Lod-loc: Aerial visual localization using lod 3d map with neural wireframe alignment. *Advances in Neural Information Processing Systems* **37**, 119063–119098 (2024)