ContinuLoc: Continuous Pose Inference over Neural Fields for UAV Geo-Localization

Weikun Zhou[1,2,3] **Jinliang Lin**[1] **Yu Zang**[1,2,3,*] **Lin Xie**[1,2,3] **Haiqing Fu**[1,2,3]
Sheng Ao[1,2,3] **Shangshu Yu**[4] **Zhiming Luo**[1] **Cheng Wang**[1,2,3]
[1]Fujian Key Laboratory of Urban Intelligent Sensing and Computing,
Xiamen University, Xiamen, China
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, Xiamen, China
[3]The Belt and Road Research Institute, Xiamen University, Xiamen, China
[4]School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China
zhouweikun@stu.xmu.edu.cn zangyu7@xmu.edu.cn

Abstract

Cross-view UAV geo-localization involves three design choices: how to represent the UAV observation, how to represent the geo-referenced map, and how to infer pose from their interaction. The dominant paradigm organizes the map as a discretized 2DoF candidate library, coupling map representation to pose inference and thus bounding spatial localization precision by candidate sampling resolution. We decouple map representation from pose inference by encoding the geo-referenced map as a continuously queryable semantic neural field, enabling localization to be formulated as explicit optimization over a continuous 4DoF pose space — parameterized by 2D ground position, in-plane heading, and viewing scale. Pose hypotheses are evaluated on demand using global descriptors in a shared semantic feature space, bypassing the sampling bottleneck of retrieval. We learn 4DoF pose-sensitive visual representations under full 4DoF supervision provided by ContinuScenes, a multi-city benchmark with dense near-nadir UAV imagery and complete pose annotations. Inference proceeds hierarchically, progressing from broad pose-space coverage to adaptive concentration on high-compatibility regions. Experiments demonstrate substantial gains over representative localization paradigms under progressively stricter 4DoF accuracy criteria. Code is available at `https://github.com/ZhouWeikun/ContinuLoc/`.

1 Introduction

Cross-view UAV geo-localization is often framed as a two-degree-of-freedom (2DoF) retrieval problem aimed at bridging severe appearance discrepancies between UAV observations and geo-referenced remote-sensing maps [1–8]. However, describing the task purely in terms of visual matching obscures its internal structure. Beyond learning discriminative visual representations, a localization system must also determine how reference-map evidence is represented and how the query pose is inferred from observation–map compatibility. This reveals three interdependent components: visual representation, map representation, and pose inference. How these components are organized jointly shapes the achievable localization accuracy, map storage requirements, and online inference cost.

Viewed through this three-component perspective, the prevailing 2DoF retrieval paradigm [1–8] has primarily emphasized learning geographically discriminative visual representations. The geo-referenced map is typically organized as a discrete reference library, while pose inference reduces to one-step similarity-based candidate selection, restricting pose estimates to predefined hypotheses.

*Corresponding Author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

For visual representation, the question is which pose cues should be preserved for localization. Under the typical near-nadir viewing setting [8], cross-view UAV localization can be viewed as placing a UAV observation into a geo-referenced remote-sensing map through geometric alignment. This setting can be parameterized by a task-specific 4D localization state, consisting of 2D ground position, in-plane heading, and viewing scale. The viewing scale captures the relative ground footprint of the UAV observation with respect to the reference map, from which flight altitude can be recovered given known camera intrinsics. This reveals a gap between the pose information available from observation–map alignment and what the prevailing 2DoF retrieval paradigm explicitly recovers. Although such alignment can constrain all four pose variables, 2DoF retrieval explicitly estimates only position. Moreover, position-only supervision encourages observations from the same location to have similar representations regardless of heading or viewing scale, potentially suppressing the cues needed to recover these two variables. In this work, we instead learn 4DoF pose-sensitive visual representations, enabled by ContinuScenes, a multi-city benchmark providing full 4DoF pose annotations. Here, we use the term 4DoF pose to describe planar similarity alignment, where viewing scale constitutes an alignment degree of freedom rather than a kinematic one [9].

While 4DoF-aware visual representations preserve richer pose cues, recovering these variables also exposes a structural limitation of the retrieval paradigm. Extending a discrete reference library to jointly cover position, heading, and viewing scale requires increasingly dense sampling, leading to rapidly growing map storage and indexing costs. The underlying issue is that the reference library simultaneously serves as a representation of map evidence and an enumeration of pose hypotheses, directly coupling inference resolution to the density of stored map evidence. From the three-component perspective, however, map representation need not be tied to the fixed set of pose hypotheses used for inference. Indeed, a dense retrieval library is generated from the same geo-referenced map under different pose hypotheses and thus contains substantial redundancy [10]. We exploit this redundancy by representing the geo-referenced map as a compact pose-conditioned map function [11, 12] that generates map descriptors on demand for arbitrary continuous pose hypotheses, rather than storing each sampled hypothesis independently. For a new scene, the field is fitted solely from its geo-referenced map, without retraining the visual representation model.

This architectural decoupling, in turn, invites a rethinking of pose inference. Matching the visual representation against this pose-conditioned map function defines a continuously evaluable compatibility landscape over the pose space. Localization can thus be formulated as a global optimization problem over this landscape. However, high-compatibility regions occupy only a small fraction of the pose space, making them costly to identify reliably through uniform sampling alone. This gives pose inference the characteristic structure of a rare-event search problem [13, 14]. Motivated by distribution-adaptation principles in stochastic optimization [15, 16], we develop a hierarchical pose inference scheme that exploits this sparsity by progressively concentrating sampling from broad pose-space exploration toward a small number of high-compatibility regions. Together, these designs enable pose inference to shift from one-step retrieval among fixed candidates to adaptive search over a continuous pose space, as shown in Figure 1.

In this work, we introduce ContinuLoc, a framework that rethinks UAV geo-localization from a decoupled perspective. Our contributions are threefold:

- We propose a neural-field representation of geo-referenced maps that exploits redundancy in descriptors extracted from pose-conditioned reference crops, enabling compact storage and on-demand evaluation.
- We formulate pose inference as a rare-event search problem and propose a hierarchical inference scheme that exploits the sparsity of high-compatibility regions to adaptively concentrate computation, avoiding dense enumeration of pose hypotheses.
- We construct ContinuScenes, a multi-scene UAV geo-localization benchmark with complete scene-level reference maps and refined 4DoF annotations. Experiments demonstrate the benefits of continuous pose inference and its complementarity with geometric refinement.

2 Related Work

Discrete Reference Retrieval. Early cross-view localization benchmarks primarily formulate the task as center-aligned retrieval [1, 17] or landmark classification [18]. Game4Loc [19] extends this



Figure 1: Illustration of ContinuLoc for UAV geo-localization. Left and center: visualization of the compatibility landscape, with pose candidates from the three inference stages. Right: the query image and reference views corresponding to the ground-truth pose and the highest-scoring pose at each stage.

Table 1: Comparison of representative localization paradigms.

Paradigm	**Map Representation**	**Pose Hypotheses**	**Inference Strategy**	**Pose Evaluation Cost**
Descriptor Retrieval [1, 19]	Global descriptors	Discrete	Enumerative	Low, vector similarity
Template Matching [23, 26]	Spatial local features	Discrete grid	Enumerative	Template matching
APR [33, 35]	Regressor weights	-	Regression	Network forward pass
Neural-field Alignment [39–41]	Neural field	Continuous	Local iteration	High, render-based alignment
ContinuLoc	**Neural field**	**Continuous**	**Global search**	**Low, vector similarity**

paradigm to unaligned UAV localization through retrieval over discretized satellite tiles, where finite reference sampling introduces a gap between the continuous true pose and the discrete candidate set, as also evidenced by DReSS [20]. Local geometric matching [5–7] can further refine retrieved poses beyond the reference resolution. However, such post-retrieval refinement depends on the correct region being retrieved and is typically restricted to a small top-K shortlist due to computational cost [21, 22]. Thus, localization remains sensitive to discrete candidate coverage and initial retrieval recall.

Grid-Based Template Matching. In the broader setting of 2D-map-based localization, some methods [23–26] rely on dense local feature maps to evaluate pose hypotheses over discretized 3DoF spaces, producing rotation-sensitive score distributions suitable for sequential probabilistic localization [27, 28]. However, compared with the global-vector similarity used in descriptor retrieval, dense template matching requires aggregating similarities across multiple local features, increasing per-hypothesis evaluation costs. Combined with enumerative pose evaluation, this exacerbates the sampling-density bottleneck discussed in the Introduction.

Absolute Pose Regression. Beyond planar-map localization, research on 6DoF localization addressing full pose estimation in 3D environments [29–32] has extensively explored absolute pose regression (APR) [33–35], which directly predicts continuous poses from visual observations by amortizing scene-specific inference into learned network parameters. While computationally efficient, embedding the appearance-to-coordinate mapping within the regressor can complicate transfer to unseen scenes. More fundamentally, standard point-regression models do not explicitly score alternative pose hypotheses, limiting their ability to assess prediction confidence.

Neural Scene-Based Pose Alignment. Another line of 6DoF camera relocalization research builds on neural implicit map representations [11, 36, 37], which provide continuous alternatives to discrete scene storage. Existing methods typically recover poses through iterative alignment against rendered RGB or feature observations [38–40], or geometric refinement using rendered scene evidence [41–43]. Although these methods provide continuous scene access, inference remains predominantly local and often requires expensive per-hypothesis rendering or alignment. Table 1 highlights the trade-offs among these localization paradigms in map representation and pose inference.

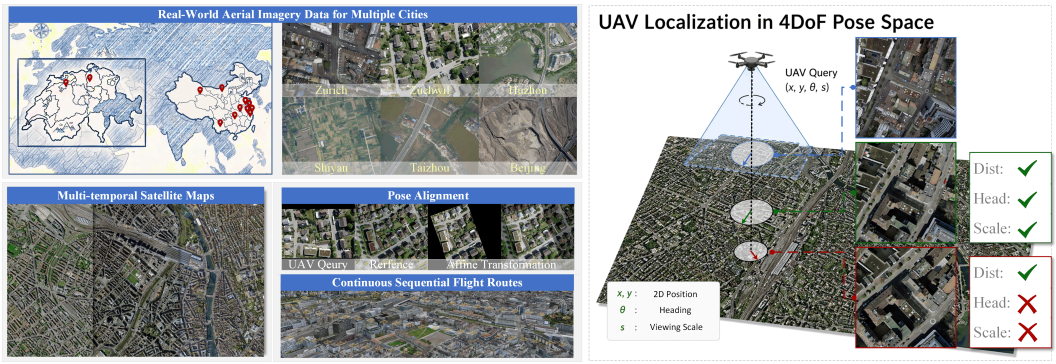


Figure 2: Overview of ContinuScenes and the 4DoF localization task. Left: ContinuScenes covers multiple cities with real-world UAV imagery, multi-temporal maps, and continuous flight routes. Right: The task estimates a 4DoF pose (x, y, θ, s) placing the UAV query into the reference map.

Table 2: Comparison with representative UAV geo-localization datasets.

Dataset	UAV Query				Reference Map				Task Type
	Platform	Head	Scale	Seq	Regions	GNSS	Multi-date	Map-Form	
University-1652 [18]	Virtual	✗	✗	✗	72	✗	✗	Instance-Set	Landmark classification
SUES-200 [44]	Drone	✗	✗	✗	200	✗	✗	Instance-Set	Landmark classification
ALTO [17]	Drone	✗	✗	✓	1	✓	✗	Center-aligned Views	Center-aligned retrieval
DenseUAV [1]	Drone	✗	✓	✓	14	✓	✓	Center-aligned Views	Center-aligned retrieval
GTA-UAV [19]	Virtual	✗	✓	✗	1	✓	✗	Tiled Map	Map Tile Localization
Ours	Drone	✓	✓	✓	12	✓	✓	Continuous Map	Continuous Pose Estimation

Note. Seq denotes whether UAV queries provide sequential observations. GNSS indicates whether absolute sensor metadata, e.g., RTK, is available.

3 ContinuScenes: Enabling 4DoF-Aware UAV Geo-Localization

To address the limitation of conventional 2D representation learning, we first construct ContinuScenes, as shown in Figure 2 and Table 2, to provide the 4DoF supervision required for learning pose-sensitive visual representations. We collect UAV images and reference maps from two sources: the publicly released UAV-VisLoc [8] dataset for the Chinese scenes, and orthophotos from swisstopo for the Swiss scenes. We derive refined 4DoF pose annotations from coarse sensor metadata by aligning each UAV image with its corresponding reference-map crop and extracting translation, heading, and viewing-scale corrections from the estimated 2D similarity transform. Standardized data splits and unified evaluation protocols are also provided for this continuous 4DoF localization task. Further details are provided in Appendix A.

4 ContinuLoc: Continuous Pose Inference over Neural Fields

4.1 Problem Formulation

Given a UAV query image I_q, a visual encoder parameterizes a mapping $\psi(\cdot)$ from an image to a global visual descriptor, yielding the query representation $\mathbf{z}_q = \psi(I_q) \in \mathbb{R}^d$. On the map side, the geo-referenced map M is represented by a semantic neural field parameterized by Θ_M, which defines a mapping from the continuous pose space $\mathcal{P}$ to pose-conditioned map descriptors, $\mathbf{z}_M(\mathbf{p}) = \phi_{\Theta_M}(\mathbf{p}) \in \mathbb{R}^d$, where $\mathbf{p} \in \mathcal{P}$. Given these two representations, a compatibility function $\mathcal{C}(\cdot, \cdot)$ evaluates their compatibility under each pose hypothesis, yielding a scalar score $S_q(\mathbf{p}) = \mathcal{C}(\mathbf{z}_q, \mathbf{z}_M(\mathbf{p}))$. Here, $\mathcal{C}(\cdot, \cdot)$ can be instantiated as cosine similarity, negative Euclidean distance, or a learnable matching function. Evaluating this score over $\mathcal{P}$ induces a continuous compatibility landscape. Accordingly, as illustrated in Figure 3, ContinuLoc formulates UAV geo-localization as compatibility-based inference over a continuous pose space, with the objective of identifying the pose that maximizes the compatibility score:

$$\hat{\mathbf{p}} = \arg\max_{\mathbf{p} \in \mathcal{P}} S_q(\mathbf{p}). \tag{1}$$

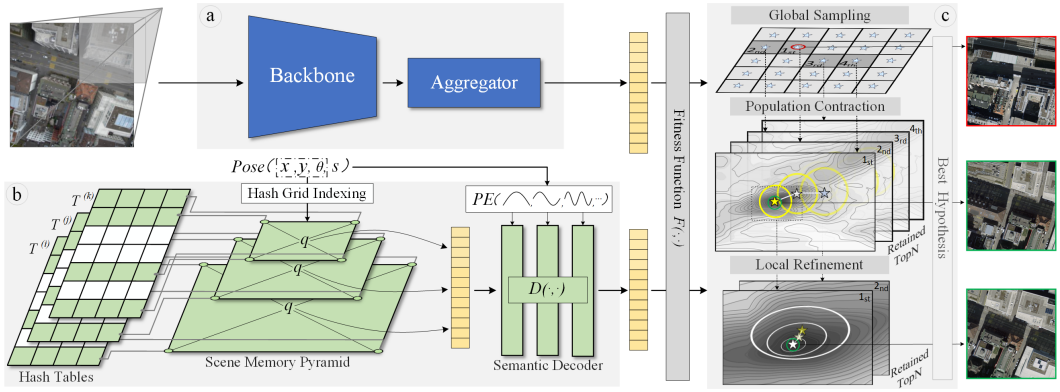


Figure 3: **Overview of ContinuLoc from the three-component perspective of UAV geo-localization.** (a) visual representation via a visual encoder maps the UAV query to a global descriptor; (b) pose-conditioned map representation via a semantic neural field provides pose-conditioned map descriptors; (c) pose inference searches the induced compatibility landscape from global coverage to local refinement. These three components interact via a shared compatibility function.

4.2 Learning Pose-Sensitive Visual Representations

Equation 1 formulates localization as an argmax problem, which implicitly induces a relative ranking over pose hypotheses: the true pose should achieve a higher query–map compatibility score than competing hypotheses. Under the near-nadir setting, each pose is parameterized as $\mathbf{p} = (x, y, \theta, s)$, where (x, y), θ, and s denote ground position, in-plane heading, and viewing scale, respectively. The ranking objective thus requires the visual encoder ψ to produce 4DoF pose-sensitive descriptors in a shared cross-view feature space.

To learn such representations, we adopt metric learning [45, 46], a common paradigm in cross-view geo-localization [1, 19, 47], using a triplet loss [48] to encourage higher similarity between the descriptors of UAV observations and their pose-matched reference crops than between those of observations and reference crops associated with competing pose hypotheses. Once trained, the visual encoder ψ serves as a scene-generalizable component that can be reused on unseen maps without retraining. Unlike conventional approaches, we omit rotation and scale augmentations to preserve the learned visual representations' sensitivity to heading θ and viewing scale s. Although we focus on 4DoF, our framework can also support traditional 2DoF localization with visual representations sensitive only to 2D position (x, y).

4.3 Neural Compression of Pose-Conditioned Semantic Fields

Continuous pose inference requires evaluating query–map compatibility at arbitrary poses $\mathbf{p} \in \mathcal{P}$. However, a finite reference library provides direct descriptor access only at predefined poses. Approximating arbitrary-pose access through denser sampling requires storing an increasingly large number of high-dimensional descriptors. Drawing on neural compression [49, 50], we view the pose-indexed descriptors $\{\mathbf{z}_M^*(\mathbf{p}_i)\}_{i=1}^N$ as discrete samples of an underlying high-dimensional, vector-valued semantic signal over the continuous pose space. Rather than storing these sampled descriptors independently, we formulate map representation as a compression problem and learn a compact, scene-specific neural field to directly represent the underlying semantic signal:

$$\phi_{\Theta_M} : \mathcal{P} \to \mathbb{R}^d, \qquad \mathbf{p} \mapsto \mathbf{z}_M(\mathbf{p}) = \phi_{\Theta_M}(\mathbf{p}), \tag{2}$$

where Θ_M denotes the scene-specific parameters that encode the pose-conditioned descriptor signal of map M. We implement this neural field using two complementary components: a multi-resolution map memory and a pose-conditioned semantic decoder. The multi-resolution memory compactly encodes map information through shared feature parameters, while the pose-conditioned decoder combines the retrieved memory features with the full pose encoding to generate pose-conditioned map descriptors.

Multi-Resolution Map Memory. Inspired by Instant-NGP [51], we organize the map memory $\mathcal{G}$ as a multi-resolution feature pyramid, with each level implemented as an independent hash table. At each pyramid level, the continuous coordinate space is covered by a regular grid with a level-specific resolution, and each grid vertex is mapped to an entry in a hash table by a hash encoding function. Given a query coordinate, the entries stored at its adjacent grid vertices will be interpolated and returned as the result. For a pose hypothesis $\mathbf{p} = (x, y, \theta, s)$, we use its position-and-heading component $\mathbf{q} = (x, y, \theta)$ to address the hash-grid map memory. For level l, the corresponding readout can be written as:

$$\mathbf{f}_{\text{grid}}^{(l)}(\mathbf{q}) = \text{Interp}\Big(T^{(l)}, \mathbf{q}\Big)\,, \tag{3}$$

where $T^{(l)}$ denotes the hash table at level l. The readouts from different pyramid levels provide features with varying spatial granularities. We concatenate them into a memory vector:

$$\mathbf{f}_{\text{grid}}(\mathbf{q}) = \text{Concat}_{l=1}^{L}\, \mathbf{f}_{\text{grid}}^{(l)}(\mathbf{q}). \tag{4}$$

Pose-Conditioned Semantic Readout. To perform pose-dependent interpretation of the memory vector, we introduce a lightweight MLP decoder that injects full-pose positional encoding (PE) [52] into its hidden transformations as a modulation signal [53]. Let $\tilde{\mathbf{p}} = (x, y, \cos\theta, \sin\theta, \log s)$ denote the transformed full-pose vector, where the heading is represented by periodic coordinates and the viewing scale is represented in log space. We apply a sinusoidal positional encoding with K frequency bands to each scalar component $\xi \in \tilde{\mathbf{p}}$:

$$\gamma(\xi) = \left[\xi,\, \sin(2^0\pi\xi),\, \cos(2^0\pi\xi),\, \ldots,\, \sin(2^{K-1}\pi\xi),\, \cos(2^{K-1}\pi\xi)\right]. \tag{5}$$

The predicted map descriptor is then written as:

$$\begin{aligned} \mathbf{e}(\mathbf{p}) &= \underset{\xi\in\tilde{\mathbf{p}}}{\text{Concat}}\, \gamma(\xi), \\ \mathbf{z}_M(\mathbf{p}) &= \phi_{\Theta_M}(\mathbf{p}) = D(\mathbf{f}_{\text{grid}}(\mathbf{q}), \mathbf{e}(\mathbf{p}))\,. \end{aligned} \tag{6}$$

Semantic Field Distillation. Under this parameterization, learning the underlying semantic signal amounts to fitting the neural field to the sampled descriptors at their corresponding poses. For each sampled pose hypothesis $\mathbf{p}_i$, we extract the corresponding reference crop from the map and obtain its target map descriptor $\mathbf{z}_M^*(\mathbf{p}_i)$ using the frozen visual encoder. The field parameters are optimized to minimize the following regression objective:

$$\mathcal{L}_{\text{map}} = \frac{1}{N}\sum_{i=1}^{N} \|\phi_{\Theta_M}(\mathbf{p}_i) - \mathbf{z}_M^*(\mathbf{p}_i)\|_2^2\,. \tag{7}$$

4.4 Hierarchical Search over the Compatibility Landscape

Assuming that the compatibility landscape induced by the visual encoder and map field is continuous and attains its unique global maximum at the ground-truth pose, the region containing the true pose can be defined as follows:

$$\mathcal{E}_\gamma(q) = \{\mathbf{p} \in \mathcal{P} : S_q(\mathbf{p}) \geq \gamma\}\,. \tag{8}$$

The score threshold γ is a sufficiently high value to exclude competing but distant hypotheses and restricts the region to the vicinity of the ground-truth pose. This region occupies only a small fraction of the entire pose space. Therefore, successful localization can be viewed as a rare-event search problem: the challenge is to efficiently reach a small set of near-optimal poses within a large pose space.

The cross-entropy method [13] provides a natural principle for this rare-event search problem: adapt the proposal distribution toward high-performing samples so that high-compatibility poses become more likely to be sampled. Standard cross-entropy optimizers characterize the search state with an explicit parameterized distribution such as a Gaussian distribution. However, the compatibility landscape may be extremely rugged and cannot be faithfully characterized by a single simple distribution. A direct extension is to first decompose the global pose space into many local regions and then run a separate parametric population within each plausible region [54]. But it comes with substantial online computation. In fact, the sparsity of high-compatibility regions suggests that most local regions do not warrant expensive distribution adaptation. Accordingly, we propose a three-stage, parameter-light evolutionary inference procedure [15, 16] that progressively filters candidate regions and concentrates computational effort on the most promising ones before local refinement.

Stage I: Global Exploration. We first conduct a global evaluation over the 4DoF pose space, guided by prior knowledge, to ensure broad coverage of the compatibility landscape. The top-scoring coarse hypotheses are retained as anchors for subsequent local refinement.

Stage II: Population Contraction. Lightweight particle populations are initialized around the retained anchors, one for each local candidate region. We evolve these populations through repeated selection and resampling. Within each population, top-scoring particles are selected as elites; their aggregate location re-centers the population toward the local high-compatibility region, while radius contraction progressively reduces the search range. Across populations, persistently weak regions are eliminated, whereas strong populations are retained in parallel to preserve competing hypotheses. This stage implements an evolutionary contraction process that preserves multi-modal uncertainty while reducing the initial coarse regions to a compact set of retained modes.

Stage III: Local Refinement. After population contraction, the search has been narrowed to a small number of plausible local modes, shifting the inference problem from multi-modal exploration to local exploitation. Since only a few modes remain, it becomes affordable to use covariance-adaptive evolutionary refinement [55], adaptively exploring the local compatibility landscape around each retained mode. The final prediction is thus selected as the highest-scoring pose across all covariance-adaptive refinements. Further algorithmic details are provided in Appendix B.

5 Experiments

Datasets. We evaluate on four representative ContinuScenes scenes (Section 3): Taizhou03, Taizhou04, Zurich, and Zuchwil. Taizhou03/04 cover larger areas with higher-altitude UAV observations, while Zurich and Zuchwil are smaller urban scenes with lower-altitude observations, enabling evaluation of both global hypothesis search and fine-grained local pose estimation. Viewing scale is supervised by the scene-normalized label $s^* = h^*/\bar{h}$, where h^* is the query footprint and $\bar{h}$ is the average UAV footprint of the scene. Unless otherwise stated, the main paper reports same-area results. Detailed dataset statistics and cross-area results are provided in Table 7 and Table 11 in the appendix.

Evaluation Metric. We evaluate localization accuracy by measuring distance, heading, and scale errors. For a predicted pose $\hat{\mathbf{p}} = (\hat{x}, \hat{y}, \hat{\theta}, \hat{s})$ and ground-truth $\mathbf{p}^* = (x^*, y^*, \theta^*, s^*)$:

$$e_d = \|(\hat{x}, \hat{y}) - (x^*, y^*)\|_2 , \tag{9}$$

$$e_h = \min\left(|\hat{\theta} - \theta^*|,\ 360^\circ - |\hat{\theta} - \theta^*|\right), \tag{10}$$

$$e_s = \max\left(\frac{\hat{s}}{s^*}, \frac{s^*}{\hat{s}}\right). \tag{11}$$

We report nested recall R@1 under three cumulative criteria: distance only (D), distance and heading ($D\&H$), and all three ($D\&H\&S$). Recall values are reported as percentages; distance and heading errors are measured in meters and degrees, respectively, and scale errors are dimensionless ratios. We use two median-error metrics in this paper: overall median errors computed over all queries, and Median@Success computed only over queries satisfying the corresponding recall criterion.

5.1 Implementation Details

We use DINOv2 [56] as the visual backbone and fine-tune it with parallel adapters following SeLaVPR [57]. SALAD [58] is adopted as the global aggregator to produce 1024-dimensional descriptors for both UAV observations and pose-conditioned reference crops. For each scene, the semantic neural field is trained to output 1024-dimensional descriptors in the same embedding space. Detailed architecture and training hyperparameters are provided in Appendix C.1.

5.2 Localization Performance across Scenes

Table 3 compares representative localization paradigms, including 2DoF retrieval, discrete 4DoF retrieval, APR, and our continuous inference framework, under increasingly strict pose constraints.

Table 3: Comparison of representative localization paradigms on four scenes.

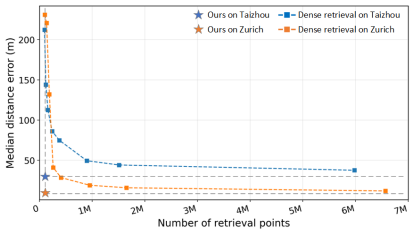
		Taizhou03						Taizhou04					
		R@1 (100 m, 10°, 1.2×) ↑			Median@Success ↓			R@1 (100 m, 10°, 1.2×) ↑			Median@Success ↓		
Paradigm	Method	D	D&H	D&H&S	Dist	Head	Scale	D	D&H	D&H&S	Dist	Head	Scale
2DoF Ret.	DenseUAV [1]	30.87	–	–	71.99	–	–	20.42	–	–	78.13	–	–
	QDFL [4]	32.21	–	–	72.53	–	–	18.31	–	–	76.87	–	–
4DoF Ret.	NetVLAD [59]	30.20	16.78	15.44	75.32	3.63	1.060	17.61	11.97	11.97	76.76	5.40	1.044
	SALAD [58]	28.86	19.46	18.79	75.32	4.82	1.086	14.08	9.86	9.86	74.59	**3.83**	1.044
	SeLaVPR [57]	28.86	21.48	20.13	72.67	**3.09**	1.083	17.61	11.97	11.97	78.13	5.40	1.059
APR	PoseNet [35]	1.34	0.00	0.00	–	–	–	1.41	0.00	0.00	–	–	–
Cont. inf.	Ours	**91.28**	**81.21**	**81.21**	**30.98**	4.27	**1.047**	**83.10**	**76.76**	**76.76**	**48.53**	4.03	**1.020**
		Zurich						Zuchwil					
		R@1 (50 m, 10°, 1.2×) ↑			Median@Success ↓			R@1 (50 m, 10°, 1.2×) ↑			Median@Success ↓		
Paradigm	Method	D	D&H	D&H&S	Dist	Head	Scale	D	D&H	D&H&S	Dist	Head	Scale
2DoF Ret.	DenseUAV [1]	19.48	–	–	24.04	–	–	23.01	–	–	24.75	–	–
	QDFL [4]	27.31	–	–	19.82	–	–	29.76	–	–	25.83	–	–
4DoF Ret.	NetVLAD [59]	27.31	13.45	8.82	24.28	3.40	1.112	32.18	16.26	12.11	27.95	3.95	1.053
	SALAD [58]	26.89	11.34	6.30	19.66	3.94	**1.055**	25.95	13.15	10.38	20.59	4.48	1.064
	SeLaVPR [57]	31.09	19.33	11.34	25.36	3.94	1.085	30.10	23.18	19.03	25.78	3.71	1.084
APR	PoseNet [35]	7.56	2.94	2.94	32.51	3.10	1.063	10.03	3.81	3.81	23.23	4.40	**1.013**
Cont. inf.	Ours	**89.92**	**81.93**	**70.59**	**9.31**	**2.86**	1.079	**76.12**	**61.94**	**61.25**	**10.79**	**2.72**	1.070

Table 4: Effect of GIM$_{dkm}$-based local refinement on progressive recall.

	Taizhou03			Taizhou04			Zurich			Zuchwil		
Method	D	D&H	D&H&S	D	D&H	D&H&S	D	D&H	D&H&S	D	D&H	D&H&S
QDFL [4]	86.58	86.58	86.58	73.24	73.24	72.54	48.74	48.74	39.50	50.87	49.48	48.79
SeLaVPR [57]	67.11	67.11	67.11	64.79	64.79	64.08	48.74	48.32	43.28	47.75	46.37	46.02
Ours	**99.33**	**99.33**	**99.33**	**88.03**	**88.03**	**87.32**	**91.60**	**91.18**	**81.93**	**81.31**	**78.55**	**77.16**

Although APR offers a compact map representation, its localization accuracy is substantially lower than that of explicit retrieval. In contrast, our framework achieves the highest R@1 under all three criteria across all four scenes. Table 4 further isolates the role of matching-based local refinement using GIM [21] applied to the top-5 retained candidates. Local refinement benefits both retrieval-based methods and our framework. Notably, before using GIM, our framework already outperforms GIM-refined retrieval baselines on three of the four scenes under the full $D\&H\&S$ criterion. This indicates that the benefits of local matching remain fundamentally bounded by the quality and coverage of the initial hypotheses.

5.3 Dense Retrieval versus Continuous Inference



Metric	Dense retrieval only					w/ S-II	Full
InitPts	**98k**	259k	408k	954k	6.6M	**98k**	**98k**
Storage (GB)↓	**0.4**	1.1	1.7	3.9	26.9	2.8	2.8
Time (ms)↓	**1.3**	3.6	5.7	33.3	350.6	98.0	336.4
MedErr (m)↓	231.1	41.1	28.5	19.0	11.9	14.5	**10.2**
R@1 ↑	8.0	19.3	24.0	30.3	36.1	63.0	**70.6**

Figure 4: Accuracy-efficiency trade-off between dense retrieval and our continuous inference framework. Left: Median localization error versus reference-library density. Right: Storage, runtime, and localization accuracy for dense retrieval and our inference stages.

A natural question is whether retrieval-based localization can be made sufficiently accurate simply by increasing the sampling density of the reference library. We therefore compare dense-retrieval libraries with different numbers of reference points. As shown in Figure 4, denser retrieval libraries

initially reduce localization error, but they exhibit clear diminishing marginal returns as the number of reference points continues to grow. This reflects a structural limitation of dense retrieval: improving precision requires a disproportionately larger reference library. In contrast, our approach achieves a comparable median localization error after Stage II using substantially less storage and runtime. The optional Stage III refinement further improves localization accuracy at additional computational cost. And this efficiency advantage is expected to become more pronounced in higher-dimensional pose spaces (e.g., 6DoF).

5.4 Ablation Study

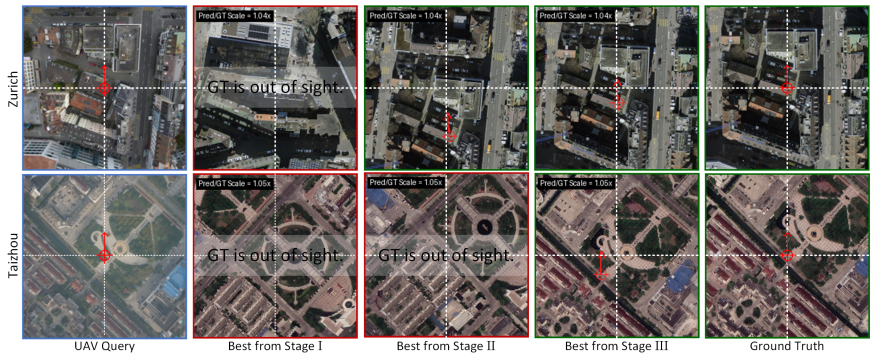
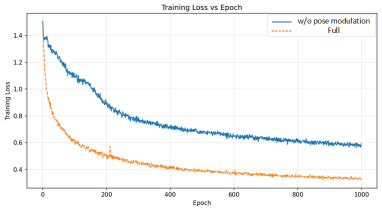


Figure 5: Qualitative visualization of stage-wise pose refinement. The red arrows indicate the ground-truth query positions and orientations in the coordinate frames of the reference views corresponding to the predicted poses.

Table 5: Stage-wise progression of the inference algorithm.

	Taizhou03						Zurich					
	R@1 (100 m, 10°, 1.2×) ↑			Median ↓			R@1 (50 m, 10°, 1.2×) ↑			Median ↓		
Stage	D	D&H	D&H&S	Dist	Head	Scale	D	D&H	D&H&S	Dist	Head	Scale
Global Exploration	30.20	24.16	24.16	160.83	12.94	1.044	31.93	22.27	20.17	115.11	29.97	1.102
Population Contraction	89.26	75.17	75.17	43.04	5.13	1.044	87.82	74.79	63.03	14.55	4.42	1.102
Local Refinement	91.28	81.21	81.21	34.45	5.00	1.044	89.92	81.93	70.59	10.15	3.55	1.102

Stage-wise Inference Analysis. To evaluate how each stage of our hierarchical inference algorithm contributes to the final localization result, we take the top-scoring pose hypothesis produced at each stage and compare it with the ground-truth pose. Stage I evaluates a sparse set of pose hypotheses, providing only a coarse sampling of the entire landscape. Consequently, the highest-scoring sampled hypothesis may not identify the most promising region across the underlying landscape. As shown in Table 5 and Figure 5, Stages II and III progressively resample and refine the retained candidate regions, providing denser local evaluations that can revise their relative ranking and yield more accurate pose estimates. The Taizhou example is deliberately selected to illustrate that visually consistent localization can still be achieved despite slight residual errors in the ground-truth annotation.



	Stage	D ↑	D&H ↑	D&H&S ↑	ΔD&H&S ↑
w/o pose-PE mod.	S-I	28.15	15.55	13.03	–
	S-II	33.19	20.17	17.65	+4.62
	S-III	33.19	22.69	20.17	+7.14
Full	S-I	31.93	22.27	20.17	–
	S-II	87.82	74.79	63.03	+42.86
	S-III	**89.92**	**81.93**	**70.59**	**+50.42**

Figure 6: Effect of pose modulation in the semantic field. Left: map-field training loss with and without pose modulation. Right: stage-wise R@1 under D, $D\&H$, and $D\&H\&S$ criteria.

Pose Modulation in the Semantic Field. Figure 6 shows the effect of pose-PE modulation. Removing pose modulation only moderately affects coarse-stage recall, but substantially limits the gains from later-stage refinement. The training curves suggest that this is due to weaker field learnability, as the unmodulated field converges more slowly and reaches a higher fitting loss. Thus, pose-PE modulation helps produce pose-conditioned map descriptors that better support continuous refinement.

5.5 Geometric Analysis of Compatibility Landscapes for Visual Localization

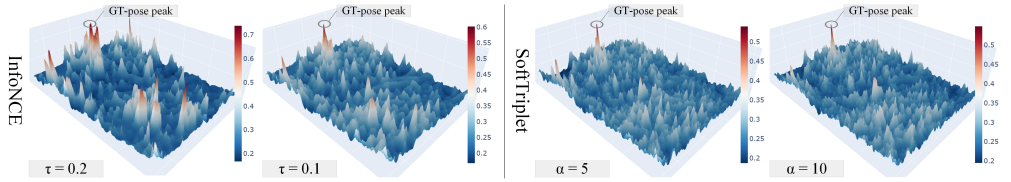


Figure 7: Compatibility landscapes induced by different loss configurations. Circles mark the ground-truth poses.

Viewing localization as continuous optimization over a continuous landscape provides a geometric perspective for analyzing the localization process. With heading and viewing scale fixed to their ground-truth values, we examine the resulting 2D compatibility landscapes and R@1 under varying spatial sampling densities. Table 6 shows that InfoNCE with $\tau = 0.2$ achieves only 66.53% R@1 even at the densest setting. This failure is geometrically reflected in Figure 7, where its corresponding landscape exhibits multiple prominent competing responses. In contrast, the other three configurations approach saturated R@1 as candidate density increases, consistent with their more clearly defined ground-truth-associated peaks in Figure 7. This suggests that finite candidate sampling provides only a sparse observation of the continuous landscape, and R@1 under a given sampling density may therefore not faithfully reflect its localization potential.

Table 6: R@1 at different sampling densities with different loss parameters.

Parameter	R@1 at Sampling Density ↑			
	1×	1.31×	2.56×	64×
InfoNCE				
$\tau = 0.2$	21.76	29.71	47.70	66.53
$\tau = 0.1$	**46.03**	**51.88**	71.97	98.33
SoftTriplet				
$\alpha = 5$	38.08	40.59	67.78	**100.00**
$\alpha = 10$	39.33	43.93	**74.06**	**100.00**

These observations suggest a broader perspective on visual localization. Conventional approaches have primarily focused on improving the discriminability of visual representations, which implicitly shapes the resulting compatibility landscape. From an optimization perspective, however, a desirable landscape should not only exhibit a well-defined structure that distinguishes the correct pose, but also be sufficiently observable under finite sampling.

6 Conclusion

This paper revisits UAV geo-localization through a three-component perspective encompassing visual representation, map representation, and pose inference. Building on this perspective, ContinuLoc learns 4DoF pose-sensitive observation representations, encodes geo-referenced maps as compact, continuously queryable pose-conditioned neural fields, and performs hierarchical search over the resulting observation–map compatibility landscape. By decoupling map representation from pose inference, ContinuLoc shifts UAV geo-localization from discrete retrieval to continuous-space optimization. Experiments demonstrate consistent improvements over discrete retrieval and absolute pose regression, while providing stronger initial hypotheses for subsequent local matching refinement. More broadly, our analysis suggests that localization depends not only on feature discriminability, but also on whether the induced compatibility landscape exhibits a clear geometric structure that makes the correct pose readily identifiable.

7 Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62471415 and 62401225), the Natural Science Foundation of Fujian Province (Grant Nos. 2023J01004, 2025J0141, and 2024J01115), and the Xiamen Municipal Natural Science Foundation (Grant No. 3502Z202472018).

References

[1] Ming Dai, Enhui Zheng, Zhenhua Feng, Lei Qi, Jiedong Zhuang, and Wankou Yang. Vision-based uav self-positioning in low-altitude urban environments. *IEEE Transactions on Image Processing*, 33:493–508, 2022.

[2] Jinliang Lin, Zhiming Luo, Dazhen Lin, Shaozi Li, and Zhun Zhong. A self-adaptive feature extraction method for aerial-view geo-localization. *IEEE Trans. Image Process.*, pages 126–139, 2025. doi: 10.1109/TIP.2024.3513157.

[3] Jinliang Lin, Zhedong Zheng, Zhun Zhong, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Joint representation learning and keypoint detection for cross-view geo-localization. *IEEE Trans. Image Process.*, pages 3780–3792, 2022. doi: 10.1109/TIP.2022.3175601.

[4] Shuyu Hu, Zelin Shi, Tong Jin, and Yunpeng Liu. Query-driven feature learning for cross-view geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025.

[5] Yuan Chen and Jie Jiang. An oblique-robust absolute visual localization method for gps-denied uav with satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–13, 2024. doi: 10.1109/tgrs.2023.3342142. URL `https://doi.org/10.1109/tgrs.2023.3342142`.

[6] Liang Cheng, Yi Yuan, Nan Xia, Song Chen, Yanming Chen, Kang Yang, Lei Ma, and Manchun Li. Crowd-sourced pictures geo-localization method based on street view images and 3d reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, pages 72–85, 2018. doi: 10.1016/j.isprsjprs.2018.04.006. URL `https://doi.org/10.1016/j.isprsjprs.2018.04.006`.

[7] Ganchao Liu, Chao Li, Sihang Zhang, and Yuan Yuan. Vl-mfl: Uav visual localization based on multisource image feature learning. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–12, 2024. doi: 10.1109/tgrs.2024.3383509. URL `https://doi.org/10.1109/tgrs.2024.3383509`.

[8] Wenjia Xu, Yaxuan Yao, Jiaqi Cao, Zhiwei Wei, Chunbo Liu, Jiuniu Wang, and Mugen Peng. Uav-visloc: A large-scale dataset for uav visual localization. *ArXiv*, abs/2405.11936, 2024.

[9] John J. Craig. *Introduction to Robotics: Mechanics and Control*. Addison-Wesley, 1986.

[10] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[11] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[12] Andrei N Kolmogorov. Three approaches to the quantitative definition of information. *Problems of information transmission*, 1(1):1–7, 1965.

[13] Reuven Y. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology And Computing In Applied Probability*, 1:127–190, 1999.

[14] Zdravko I. Botev, Dirk P. Kroese, Reuven Y. Rubinstein, and Pierre L'Ecuyer. The cross-entropy method for optimization. In Venu Govindaraju and C. R. Rao, editors, *Handbook of Statistics: Machine Learning: Theory and Applications*, volume 31, pages 35–59. Elsevier, 2013. doi: 10.1016/B978-0-444-53859-8.00003-5.

[15] Yann Ollivier, Ludovic Arnold, Anne Auger, and Nikolaus Hansen. Information-geometric optimization algorithms: A unifying picture via invariance principles. *Journal of Machine Learning Research*, 18(18):1–65, 2017. URL `https://www.jmlr.org/papers/v18/14-467.html`.

[16] Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *Journal of Machine Learning Research*, 15(27):949–980, 2014. URL https://www.jmlr.org/papers/v15/wierstra14a.html.

[17] Ivan Cisneros, Peng Yin, Ji Zhang, Howie Choset, and Sebastian A. Scherer. Alto: A large-scale dataset for uav visual place recognition and localization. *ArXiv*, abs/2207.12317, 2022.

[18] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[19] Yuxiang Ji, Boyong He, Zhuoyue Tan, and Liaoni Wu. Game4Loc: A UAV geo-localization benchmark from game data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3913–3921, 2025. doi: 10.1609/aaai.v39i4.32409.

[20] Panwang Xia, Lei Yu, Yi Wan, Qiong Wu, Peiqi Chen, Liheng Zhong, Yongxiang Yao, Dong Wei, Xinyi Liu, Lixiang Ru, et al. Cross-view geo-localization with panoramic street-view and vhr satellite imagery in decentrality settings. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227:1–11, 2025.

[21] Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. Gim: Learning generalizable image matcher from internet videos. *ArXiv*, abs/2402.11095, 2024.

[22] Yibin Ye, Xichao Teng, Shuo Chen, Zhang Li, Leqi Liu, Qifeng Yu, and Tao Tan. Exploring the best way for uav visual localization under low-altitude multi-view observation condition: a benchmark. *ArXiv*, abs/2503.10692, 2025.

[23] Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulò, Richard A. Newcombe, Peter Kontschieder, and Vasileios Balntas. Orienternet: Visual localization in 2d public maps with neural matching. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21632–21642, 2023. URL https://api.semanticscholar.org/CorpusID:257921605.

[24] Xiangze Meng, Wulong Guo, Kai Zhou, Ting Sun, Lei Deng, Shijie Yu, and Yuhao Feng. Airgeonet: A map-guided visual geo-localization approach for aerial vehicles. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–14, 2024. doi: 10.1109/tgrs.2024.3456912. URL https://doi.org/10.1109/tgrs.2024.3456912.

[25] Zhen Wang, Dianxi Shi, Chunping Qiu, Songchang Jin, Tongyue Li, Ziteng Qiao, and Yang Chen. Vecmaplocnet: Vision-based uav localization using vector maps in gnss-denied environments. *ISPRS Journal of Photogrammetry and Remote Sensing*, pages 362–381, 2025. doi: 10.1016/j.isprsjprs.2025.04.009. URL https://doi.org/10.1016/j.isprsjprs.2025.04.009.

[26] Changan Chen, Rui Wang, Christoph Vogel, and Marc Pollefeys. F3loc: Fusion and filtering for floorplan localization. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18029–18038, 2024. URL https://api.semanticscholar.org/CorpusID:268253225.

[27] M. Sanjeev Arulampalam, Simon Maskell, Neil J. Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Trans. Signal Process.*, 50:174–188, 2002. URL https://api.semanticscholar.org/CorpusID:55577025.

[28] Rico Jonschkowski and Oliver Brock. End-to-end learnable histogram filters. In *Workshop on Deep Learning for Action and Interaction at NIPS*, 2016. URL https://www.static.tu.berlin/fileadmin/www/10002220/Publications/Jonschkowski-16-NIPS-WS.pdf.

[29] 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*, volume 37, 2024. doi: 10.52202/079017-3782. URL https://papers.nips.cc/paper/2024/hash/d78ece6613953f46501b958b7bb4582f-Abstract-Conference.html.

[30] Oussema Dhaouadi, Riccardo Marin, Johannes Meier, Jacques Kaiser, and Daniel Cremers. Ortholoc: Uav 6-dof localization and calibration using orthographic geodata. *ArXiv*, abs/2509.18350, 2025.

[31] Xiaoya Cheng, Long Wang, Yan Liu, Xinyi Liu, Hanlin Tan, Yu Liu, Maojun Zhang, and Shen Yan. PiLoT: Neural pixel-to-3d registration for UAV-based ego and target geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://nudt-sawlab.github.io/PiLoT/.

[32] Oussema Dhaouadi, Zuria Bauer, Johannes Michael Meier, Olaf Wysocki, Marc Pollefeys, and Daniel Cremers. OrthoTrack: Continuous 6-DoF UAV trajectory estimation anchored in public orthophotos. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2606.25245.

[33] Shuai Chen, Yash Bhalgat, Xinghui Li, Jia-Wang Bian, Kejie Li, Zirui Wang, and Victor Adrian Prisacariu. Neural refinement for absolute pose regression with feature synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20987–20996, 2024. doi: 10.1109/cvpr52733.2024.01983. URL https://doi.org/10.1109/cvpr52733.2024.01983.

[34] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6555–6564, 2017.

[35] Alex Kendall, Matthew Koichi Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2938–2946, 2015.

[36] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4460–4470, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Mescheder_Occupancy_Networks_Learning_3D_Reconstruction_in_Function_Space_CVPR_2019_paper.html.

[37] Jeong Joon Park, Peter R. Florence, Julian Straub, Richard A. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019. URL https://api.semanticscholar.org/CorpusID:58007025.

[38] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5721–5731, 2021. URL https://api.semanticscholar.org/CorpusID:233219786.

[39] Jingyu Lin, Jiaqi Gu, Bojian Wu, Lubin Fan, Renjie Chen, Ligang Liu, and Jieping Ye. Learning neural volumetric pose features for camera localization. In *European Conference on Computer Vision*, 2024. URL https://api.semanticscholar.org/CorpusID:268532608.

[40] Yen-Chen Lin, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting neural radiance fields for pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330, 2021. URL https://yenchenlin.me/inerf/.

[41] Changkun Liu, Shuai Chen, Yash Bhalgat, Siyan Hu, Ming Cheng, Zirui Wang, Victor Adrian Prisacariu, and Tristan Braud. GS-CPR: Efficient camera pose refinement via 3D gaussian splatting. In *ICLR*, 2025. URL https://gsloc.active.vision/.

[42] Zhiwei Huang, Hailin Yu, Yichun Shentu, Jin Yuan, and Guofeng Zhang. From sparse to dense: Camera relocalization with scene-specific detector from feature gaussian splatting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27059–27069, 2025. doi: 10.1109/cvpr52734.2025.02520. URL https://doi.org/10.1109/cvpr52734.2025.02520.

[43] Gabriele Trivigno, Carlo Masone, Barbara Caputo, and Torsten Sattler. The unreasonable effectiveness of pre-trained features for camera pose refinement. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12786–12798, 2024. doi: 10.1109/cvpr52733.2024.01215. URL https://doi.org/10.1109/cvpr52733.2024.01215.

[44] Runzhe Zhu, Ling Yin, Mingze Yang, Fei Wu, Yunchen Yang, and Wenbo Hu. Sues-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. *IEEE Transactions on Circuits and Systems for Video Technology*, 33:4825–4839, 2022.

[45] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006. URL https://api.semanticscholar.org/CorpusID:8281592.

[46] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. URL https://api.semanticscholar.org/CorpusID:206592766.

[47] Hongyang Zhang, Maonan Wang, Ziyao Wang, Hongrui Yin, and Man-On Pun. InfoGeo: Information-theoretic object-centric learning for cross-view generalizable UAV geolocalization. *arXiv preprint arXiv:2605.07099*, 2026. URL https://arxiv.org/abs/2605.07099.

[48] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. URL https://arxiv.org/abs/1703.07737.

[49] Emilien Dupont, Adam Golinski, Milad Alizadeh, Yee Whye Teh, and A. Doucet. Coin: Compression with implicit neural representations. *ArXiv*, abs/2103.03123, 2021. URL https://api.semanticscholar.org/CorpusID:232110691.

[50] Sharath Girish, Abhinav Shrivastava, and Kamal Gupta. Shacira: Scalable hash-grid compression for implicit neural representations. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17467–17478, 2023. URL https://api.semanticscholar.org/CorpusID:263151321.

[51] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41:1 – 15, 2022. URL https://api.semanticscholar.org/CorpusID:246016186.

[52] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *ArXiv*, abs/2006.10739, 2020. URL https://api.semanticscholar.org/CorpusID:219791950.

[53] Michael Niemeyer, Lars M. Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3512, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Niemeyer_Differentiable_Volumetric_Rendering_Learning_Implicit_3D_Representations_Without_3D_Supervision_CVPR_2020_paper.html.

[54] Amir Omeradzic and Hans-Georg Beyer. Progress analysis of a multi-recombinative evolution strategy on the highly multimodal rastrigin function. *Theoretical computer science*, 978, 2023. URL https://api.semanticscholar.org/CorpusID:261852588.

[55] Nikolaus Hansen. The cma evolution strategy: A tutorial. *ArXiv*, abs/1604.00772, 2016.

[56] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. URL https://api.semanticscholar.org/CorpusID:258170077.

[57] Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. *ArXiv*, abs/2402.14505, 2024.

[58] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17658–17668, 2024.

[59] Relja Arandjelović, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5297–5307, 2016.

[60] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018. URL https://github.com/ethanjperez/film.

[61] Sung-Jin Hong, Jisu Nam, Seokju Cho, Susung Hong, Sangryul Jeon, Dongbo Min, and Seung Wook Kim. Neural matching fields: Implicit representation of matching fields for visual correspondence. *ArXiv*, abs/2210.02689, 2022. URL https://api.semanticscholar.org/CorpusID:252734872.

ContinuLoc: Continuous Pose Inference over Neural Fields for UAV Geo-Localization

Supplementary Material

A Detailed Dataset Descriptions

A.1 Dataset Collection

We introduce a novel UAV-satellite image pairing and geo-localization dataset with more accurate geospatial metadata than existing benchmarks. The dataset comprises 12 large-scale scenes, of which 10 are sourced from the UAV-VisLoc [8] and 2 are newly collected from publicly available UAV imagery provided by Wingtra, covering a diverse range of topographies including urban areas, croplands, deserts, hilly terrain, and forests.

The absolute altitudes in our dataset encompass an extensive range from 30 m to approximately 2,570 m, while UAV headings exhibit an omnidirectional distribution. Together, these factors provide significant viewpoint and scale diversity. Each UAV-view image is paired with a corresponding reference image crop and annotated with refined geospatial metadata, comprising GPS coordinates, ground coverage, and the heading angle. As illustrated in Figure 8, sampled UAV-view images exhibit substantial visual variation across the 12 scenes, primarily attributable to differences in flight altitude and heading angle.

Table 7: Summary of the ContinuScenes dataset.

	Scene-1	Scene-2	Scene-3	Scene-4	Scene-5	Scene-6	Scene-7	Scene-8	Scene-9	Scene-10	Scene-11	Scene-12
Scene / City	Zurich	Zuchwil	Jiujiang01	Jiujiang02	Taizhou03	Taizhou04	Kunming	Ankang	Huzhou01	Huzhou02	Beijing	Zhangye
UAV images	2413	3951	817	1071	768	738	473	344	30	1033	144	590
Map area (km^2)	12	12	20	30	64	52	4	6	54	113	2	34

A.2 Dataset Construction

To mitigate cross-view registration errors, inaccurate geospatial metadata, and missing or unreliable heading annotations present in existing benchmarks, we develop an automated pipeline for UAV-to-reference image pairing and geospatial attribute refinement. This pipeline consists of two primary stages: deep-feature-based correspondence estimation, and an affine geometry-based metadata recalibration that builds upon localized patch acquisition and coarse alignment. These steps jointly enable the recovery of accurate GPS coordinates, heading angles, and scale information.

Refined Georeferencing via Affine Matching Since neither the UAV-VisLoc dataset nor the Wingtra dataset provides pre-paired reference counterparts, we extract GPS-centered patches from a high-resolution base map and roughly align them to the UAV images by incorporating heading data and phase-domain information, thereby narrowing the search space for refined affine matching. For each UAV image, we estimate a scaling factor s such that the reference patch coverage slightly exceeds the UAV field of view, providing a sufficient search margin for subsequent matching. The patch is cropped from the reference base map, centered on the GPS coordinates recorded in the original flight logs. If heading information is available, we optionally pre-rotate the UAV image to approximately align its heading with the North-aligned reference patch, thereby reducing the rotational search space and facilitating more reliable feature matching.

Affine Matching and Geospatial Metadata Recalibration To establish precise pixel-level correspondences between UAV and reference views, we employ GIM [21] with DKM weights. An overview of this alignment process and the subsequent metadata recalibration is visually summa-

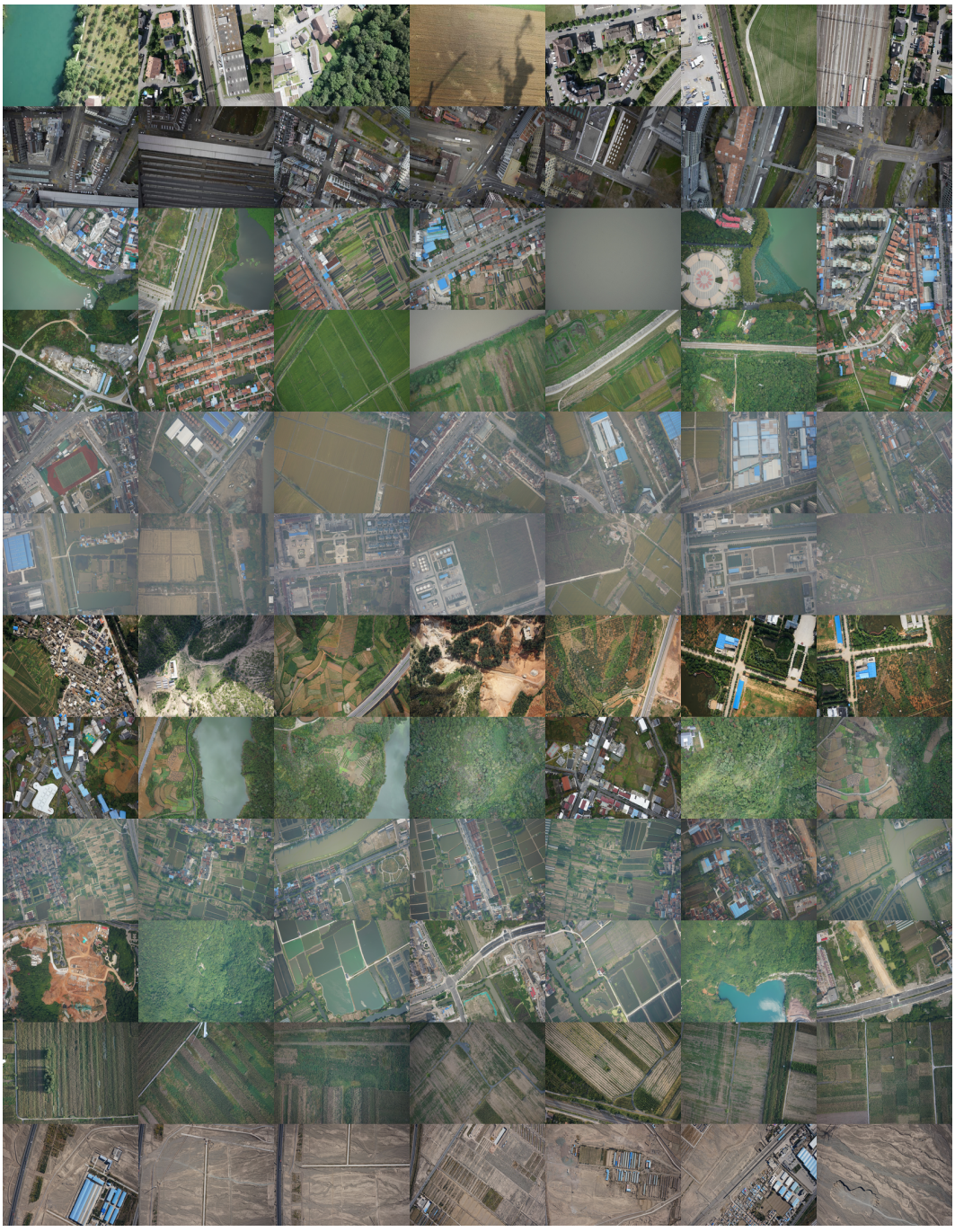


Figure 8: Visual overview of our ContinuScenes cross-view localization dataset. Sample UAV-view images across 12 evaluation scenes. The dataset covers diverse terrains (e.g., urban, farmland, deserts), demonstrating substantial variations in flight altitude and azimuth critical for robust cross-view registration.

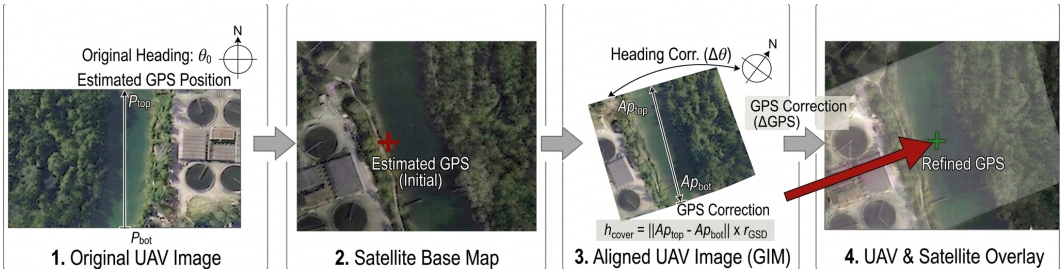


Figure 9: Overview of the UAV and satellite image alignment process. The framework estimates an affine transformation to map the original UAV image (1) to the satellite base map (2). The resulting alignment parameters (3) explicitly drive the recalibration of the UAV's heading ($\Delta\theta$), GPS position (ΔGPS), and physical footprint (h_{cover}), yielding the final geo-registered overlay (4).

rized in Figure 9. For each cross-view pair, we fit a 2×3 affine transformation matrix $\mathbf{A}$:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \mathbf{A} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & t_x \\ c & d & t_y \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{12}$$

where (x, y) and (x', y') denote pixel coordinates in the UAV and reference views, respectively. The linear sub-matrix $\mathbf{M} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ encodes rotation and scale, while the translation vector (t_x, t_y) reflects the GPS error in pixel space. Leveraging the estimated $\mathbf{A}$, we recalibrate three geospatial attributes.

(i) Position: the UAV image center $\mathbf{p}_c = [W_u/2,\ H_u/2,\ 1]^\top$ is projected onto the reference frame via $\mathbf{p}'_c = \mathbf{A}\mathbf{p}_c$, and the resulting pixel displacement relative to the reference patch center is converted into geographic increments to obtain the refined location $(\hat{\text{lat}},\ \hat{\text{lon}})$.

(ii) Heading: the rotational component is extracted from $\mathbf{M}$ via SVD, $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, yielding the nearest rotation matrix $\mathbf{R} = \mathbf{U}\mathbf{V}^\top$. The angular offset $\theta = \text{atan2}(R_{10},\ R_{00})$ is then combined with the raw flight heading and normalized to $[-180°, 180°]$.

(iii) Ground Coverage and Scale: The absolute spatial footprint of each UAV image is quantified by its ground coverage h_{cover}. This metric represents the actual physical extent of the UAV image projected onto the reference map space, defined as:

$$h_{\text{cover}} = \|\mathbf{A}\mathbf{p}_{\text{top}} - \mathbf{A}\mathbf{p}_{\text{bot}}\| \times r \tag{13}$$

where $\mathbf{A}$ is the ground-truth affine matrix, $\mathbf{p}_{\text{top}}$ and $\mathbf{p}_{\text{bot}}$ denote the centers of the top and bottom edges of the UAV image, and r represents the Ground Sampling Distance (GSD) of the base map. Based on this absolute measurement, we explicitly formulate the relative **scale factor** as the ratio $h_{\text{cover}}/h_{\text{mean}}$. This conversion from physical coverage to a normalized scale factor enables fine-grained, scale-aware performance analysis across different flight altitudes. Finally, to maintain strict benchmark integrity, any corresponding pairs exhibiting extreme geometric anomalies are systematically curated out.

A.3 Analysis of Metadata Corrections

To quantify the impact of our refinement pipeline, we analyze the distribution of corrections applied to GPS coordinates and heading angles on Taizhou03, as illustrated in Figure 10.

The latitude and longitude offsets, Δlat and Δlon, reflect the displacement between the original GPS-recorded positions and the affine-corrected locations. While both distributions are approximately centered near zero, the corrections span a considerable range, demonstrating that instance-level positional errors can be substantial and are not adequately captured by the mean alone. The heading offset $\Delta\theta$ exhibits a notably wider distribution, indicating that raw flight headings recorded in the flight logs can deviate significantly from the true North-relative orientation, and underscoring the necessity of our affine-based heading correction. In this scene, a small proportion of samples are identified as low-confidence and excluded during the outlier rejection step, confirming the effectiveness of our quality filtering.

Collectively, these results demonstrate that raw geospatial metadata in existing UAV datasets can contain non-negligible positional and angular errors, and validate the necessity of our automated refinement pipeline for producing reliable annotations.

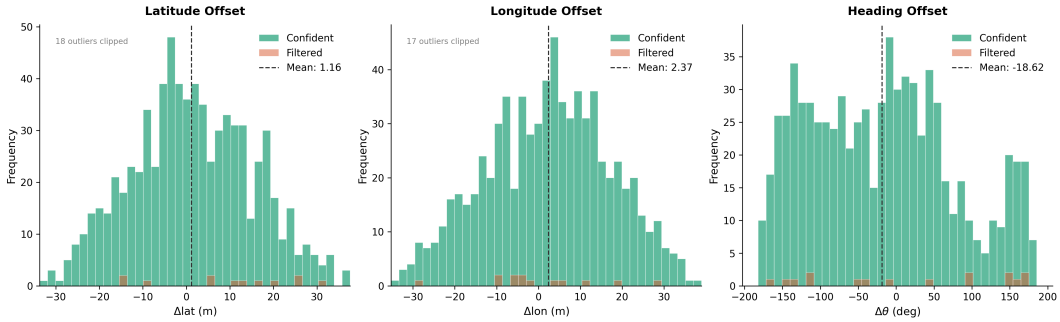


Figure 10: Distribution of metadata corrections (Δlat, Δlon, $\Delta\theta$) for Taizhou03. Green bars denote confident samples retained after outlier rejection; orange bars denote filtered low-confidence samples.

B Hierarchical Evolutionary Search: Algorithmic Details

In this section, we provide the detailed formulation and algorithmic implementation of our hierarchical evolutionary search procedure, which was outlined in Section 4.4. As described in the main text, the search decomposes continuous 4DoF UAV localization into three stages: (1) *Coarse Exploration* to obtain a broad coverage of the compatibility landscape, (2) *Population Contraction* to iteratively refine local candidate regions, and (3) *Local Refinement* using a covariance-adaptive evolutionary optimizer for final pose estimation. Here, we provide a detailed description of the symbols, parameters, and update rules used in each stage of the hierarchical search procedure; the corresponding pseudocode is presented in Algorithm 1.

Stage I: Coarse Prior Sampling We first draw a finite support $\mathcal{H} = \{\mathbf{p}_i\}_{i=1}^{N}$ from a uniform prior over the 4DoF pose space and evaluate each hypothesis using the compatibility score $S(\mathbf{p})$. To retain only high-compatibility hypotheses, we compute a quantile threshold

$$\gamma_0 = \mathrm{Quantile}_{1-\rho_0}(S(\mathcal{H})), \tag{14}$$

which corresponds to the score such that the top fraction ρ_0 of hypotheses in $\mathcal{H}$ have compatibility scores above γ_0. From the hypotheses exceeding this threshold, we select K_0 coarse candidate anchors:

$$\mathcal{A}_0 = \{\mathbf{a}_k\}_{k=1}^{K_0}. \tag{15}$$

Each anchor $\mathbf{a}_k$ is treated as the center of a coarse candidate region and is used to initialize a local population.

Stage II: Population Contraction Around each anchor $\mathbf{a}_k \in \mathcal{A}_0$, we initialize a local particle population centered at $\mathbf{m}_k^{(0)} = \mathbf{a}_k$ with an initial sampling radius $\delta_k^{(0)} = \eta \cdot r_{\mathbf{a}_k}$, where $r_{\mathbf{a}_k}$ is the radius of the local region inherited from Stage I, defined as half the distance to the nearest neighboring sample in $\mathcal{H}$ and scaled by the factor η. This defines a local neighborhood $\mathcal{B}(\mathbf{m}_k^{(0)}, \delta_k^{(0)})$ as an axis-aligned box of half-width $\delta_k^{(0)}$ centered at $\mathbf{m}_k^{(0)}$, from which the initial particles $\mathcal{X}_k^{(0)} = \{\mathbf{p}_{k,i}^{(0)}\}_{i=1}^{M}$ are drawn. Then, all populations are iteratively refined over T_{II} rounds. Each refinement round operates at two levels.

Within each population, particles are sampled from the current neighborhood $\mathcal{B}(\mathbf{m}_k^{(t-1)}, \delta_k^{(t-1)})$ and evaluated using the compatibility score $S(\mathbf{p})$. A threshold $\gamma_k^{(t)} = \mathrm{Quantile}_{1-\rho}\big(S(\mathcal{X}_k^{(t)})\big)$ is used to select the top-scoring fraction of particles, forming the local elites $\mathcal{E}_k^{(t)}$. The population center and

Algorithm 1: Hierarchical Evolutionary Search

Input: compatibility score $S(\mathbf{p})$, pose space $\mathcal{P}$
Output: optimal pose estimate $\hat{\mathbf{p}}$

Stage I: Coarse Exploration
Draw a finite support $\mathcal{H} = \{\mathbf{p}_i\}_{i=1}^{N} \sim \mathrm{Unif}(\mathcal{P})$
$\gamma_0 \leftarrow \mathrm{Quantile}_{1-\rho_0}(S(\mathcal{H}))$
$\mathcal{A}_0 = \{\mathbf{a}_k\}_{k=1}^{K_0} \leftarrow \mathrm{Top}_{K_0}\{\mathbf{p} \in \mathcal{H} : S(\mathbf{p}) \geq \gamma_0\}$

Stage II: Particle-based Population Contraction
Initialize local populations around $\mathbf{a}_k \in \mathcal{A}_0$: $\mathbf{m}_k^{(0)} \leftarrow \mathbf{a}_k$, $\delta_k^{(0)} \leftarrow \eta r_{\mathbf{a}_k}$
for $t = 1, \ldots, T_{\mathrm{II}}$ **do**
  foreach *active population* k **do**
    Draw a particle set $\mathcal{X}_k^{(t)} = \{\mathbf{p}_{k,i}^{(t)}\}_{i=1}^{M} \subset \mathcal{B}(\mathbf{m}_k^{(t-1)}, \delta_k^{(t-1)})$ and evaluate $S(\mathbf{p})$
    $\gamma_k^{(t)} \leftarrow \mathrm{Quantile}_{1-\rho}\big(S(\mathcal{X}_k^{(t)})\big)$
    $\mathcal{E}_k^{(t)} \leftarrow \{\mathbf{p} \in \mathcal{X}_k^{(t)} : S(\mathbf{p}) \geq \gamma_k^{(t)}\}$
    $\mathbf{m}_k^{(t)} \leftarrow \frac{1}{|\mathcal{E}_k^{(t)}|} \sum_{\mathbf{p} \in \mathcal{E}_k^{(t)}} \mathbf{p}$, $\delta_k^{(t)} \leftarrow \delta_k^{(0)} \alpha^t$
  Prune populations k whose population-center score $S(\mathbf{m}_k^{(t)})$ falls below the $(1 - \rho_{\text{pop}})$ quantile across all active populations
$\mathcal{A}_1 = \{\mathbf{m}_k^{(T_{\mathrm{II}})}\}_{k=1}^{K_1}$, $K_1 \ll K_0$

Stage III: Parametric Local Refinement
Select K_2 modes from $\mathcal{A}_1$
foreach *selected mode* $\mathbf{m}_k^{(T_{\mathrm{II}})}$ **do**
  Initialize local optimizer: $\tilde{\mathbf{m}}_k^{(0)} \leftarrow \mathbf{m}_k^{(T_{\mathrm{II}})}$, $\tilde{\delta}_k^{(0)} \leftarrow \delta_k^{(T_{\mathrm{II}})}$, $\tilde{\mathbf{C}}_k^{(0)} \leftarrow \mathbf{I}$
  Refine a candidate $\tilde{\mathbf{p}}_k$ by locally maximizing $S(\mathbf{p})$ with covariance-adaptive updates
$\mathcal{A}_2 = \{\tilde{\mathbf{p}}_k\}_{k=1}^{K_2}$, $K_2 \leq K_1$
$\hat{\mathbf{p}} \leftarrow \arg\max_{\mathbf{p} \in \mathcal{A}_2} S(\mathbf{p})$
return $\hat{\mathbf{p}}$

sampling radius are updated at each round t as:

$$\mathbf{m}_k^{(t)} = \frac{1}{|\mathcal{E}_k^{(t)}|} \sum_{\mathbf{p} \in \mathcal{E}_k^{(t)}} \mathbf{p}, \qquad \delta_k^{(t)} = \delta_k^{(0)} \cdot \alpha^t, \tag{16}$$

where $\mathcal{E}_k^{(t)}$ denotes the set of local elite particles whose scores $S(\mathbf{p})$ exceed the quantile threshold $\gamma_k^{(t)}$, i.e., the top fraction ρ of particles in $\mathcal{X}_k^{(t)}$. Here, $\delta_k^{(0)} = \eta\, r_{\mathbf{a}_k}$ is the initial neighborhood radius inherited from Stage I, and $\alpha \in (0, 1)$ controls the contraction rate of the sampling neighborhood.

Across populations, each population's survival is determined by the score at the mean of its elite particles, and populations whose center score falls below the $(1 - \rho_{\text{pop}})$ quantile are pruned. By combining per-population contraction with cross-population pruning, Stage II reduces the initial candidate regions to a compact set of retained modes:

$$\mathcal{A}_1 = \{\mathbf{m}_k^{(T_{\mathrm{II}})}\}_{k=1}^{K_1}, \qquad K_1 \ll K_0. \tag{17}$$

Stage III: Local Parametric Refinement After Stage II, the search has been narrowed from the coarse candidate anchors $\mathcal{A}_0$ to a smaller set of retained local modes $\mathcal{A}_1$, each representing a plausible local basin where a single-distribution assumption is reasonable. Since $K_1 \ll K_0$, applying local parametric distribution updates to the retained modes is computationally affordable.

We then select at most $K_2 \leq K_1$ modes from $\mathcal{A}_1$ and apply a covariance-adaptive evolutionary optimizer to each mode. For the k-th selected mode, the optimizer is initialized with the population center and radius from Stage II:

$$\tilde{\mathbf{m}}_k^{(0)} \leftarrow \mathbf{m}_k^{(T_{\mathrm{II}})}, \quad \tilde{\delta}_k^{(0)} \leftarrow \delta_k^{(T_{\mathrm{II}})}, \quad \tilde{\mathbf{C}}_k^{(0)} \leftarrow \mathbf{I}, \tag{18}$$

where $\tilde{\mathbf{C}}_k^{(0)}$ is the initial covariance, set to identity, and subsequently adapted during optimization.

Table 8: Data splits for training and evaluation across the four main scenes. Cross-area splits evaluate generalization to unseen trajectory regions, while same-area splits evaluate interpolation along the same trajectory.

Scene	Cross-Area (train/test)	Same-Area (train/test)
Taizhou03	0.8 / 0.2	0.8 / 0.2
Taizhou04	0.8 / 0.2	0.8 / 0.2
Zurich	0.9 / 0.1	0.9 / 0.1
Zuchwil	0.9 / 0.1	0.9 / 0.1

Each optimizer locally maximizes the compatibility score $S(\mathbf{p})$, producing a refined candidate $\tilde{\mathbf{p}}_k \in \mathcal{A}_2 = \{\tilde{\mathbf{p}}_k\}_{k=1}^{K_2}$, after which the final estimated pose is

$$\hat{\mathbf{p}} = \arg\max_{\mathbf{p} \in \mathcal{A}_2} S(\mathbf{p}). \tag{19}$$

C Experimental Setup and Implementation Details

C.1 Implementation Details

Visual representation. We use DINOv2 [56] as the visual backbone and follow the parallel-adapter strategy of SeLaVPR [57] to fine-tune the last four transformer blocks. For global aggregation, we use SALAD [58] with $K = 16$ cluster centers of dimension $D = 64$, yielding a $K \times D$-dimensional descriptor of size 1024 for each UAV observation and each pose-conditioned reference crop.

Semantic neural field. For each scene, we instantiate a multi-resolution hash-grid memory and a lightweight MLP-based decoder that maps retrieved memory to a pose-conditioned semantic descriptor. The 4DoF pose $\mathbf{p}$ is encoded and injected into a residual-style MLP via feature-wise affine modulation, following the FiLM design [60]. The residual-style MLP structure is commonly used in neural rendering frameworks [53, 61].

Hierarchical Evolutionary Search. Although Stage I scores a sparse 4DoF grid, we do not carry the discrete scale dimension into the later evolutionary search. For each spatial-heading anchor $\mathbf{a} = (x, y, \theta)$, we collect its compatibility scores over the sampled scale bins $\{s_j\}$ and compute a soft scale estimate

$$\pi_j(\mathbf{a}) = \frac{\exp(\beta S_j(\mathbf{a}))}{\sum_l \exp(\beta S_l(\mathbf{a}))}, \qquad \hat{s}(\mathbf{a}) = \exp\left(\sum_j \pi_j(\mathbf{a}) \log s_j\right), \tag{20}$$

where $S_j(\mathbf{a}) = \mathcal{C}(\mathbf{z}_q, \phi_{\Theta_M}((x, y, \theta, s_j)))$. The resulting non-grid scale estimate can be directly carried into subsequent inference because the map field is continuously queryable and admits hypotheses at arbitrary poses beyond the initial discrete grid, a property that standard retrieval libraries do not possess. Empirically, this weighted estimate provides more stable scale initialization than a naïve $\arg\max$ over the discrete bins. We therefore keep $\hat{s}$ fixed in Stage II and Stage III, and allocate the evolutionary search budget entirely to the remaining (x, y, θ) dimensions.

C.2 Experimental Setup

Dataset Partitioning For training and evaluation, we provide two types of data splits for each scene: **cross-area** splits, which use the first portion of the UAV trajectory for training and the latter portion for testing, and **same-area** splits, which sample queries along the trajectory while maintaining the same-area distribution. The corresponding configurations and train/test ratios are summarized in Table 8.

Discrete pose sampling configuration. For the initial finite support $\mathcal{H}$ used in Stage I, we construct a deterministic grid as a finite-support approximation to a uniform prior over the 4DoF pose space. For the spatial dimensions, the reference map is tiled into candidate crops according to the

image crop size, with zero overlap between adjacent crops; the crop centers define the sampled (x, y) locations. For the heading dimension, we sample the full range $[0^\circ, 360^\circ)$ at a 10° interval, resulting in 36 heading bins. For the scale dimension, we first estimate the footprint-size range of UAV images in each scene, and then uniformly sample four scale values within this range. The Cartesian product of the spatial grid, heading bins, and scale bins forms the initial hypothesis set $\mathcal{H}$ evaluated in Stage I:

$$\mathcal{H} = \mathcal{G}_{xy} \times \Theta \times \mathcal{S}, \qquad \Theta = \{0^\circ, 10^\circ, \ldots, 350^\circ\}, \quad |\mathcal{S}| = 4. \tag{21}$$

For the discrete retrieval-based 4DoF baselines in Table 3, we use the same sampling configuration to construct their reference libraries. For the 2DoF retrieval-based baselines, only the spatial grid is used: the heading is fixed to 0°, and the scale is set to the scene-specific mean UAV footprint size.

For the sampling-density experiment in Section 5.3, we progressively increase the spatial sampling density by raising the overlap ratio between adjacent map crops, starting from zero overlap. The heading and scale dimensions are kept unchanged, with 36 heading bins and 4 scale bins, respectively.

Ablation configuration. For the stage-wise inference ablation experiment in Table 5, we retain the top 5% of hypotheses from Stage I as the initialization set for Stage II, and further select the top-5 retained modes from Stage II for Stage III refinement. This configuration is kept consistent with the inference setting used in the main results in Table 3. As shown in Figure 4, Stage III accounts for most of the inference time, since covariance-adaptive refinement must maintain and update parametric statistics, including the mean, step size, and covariance matrix, for each selected mode. This motivates our hierarchical design: a parameter-light population contraction stage first reduces the number of plausible modes, so that the expensive covariance-adaptive optimizer is applied only to a few high-compatibility candidates.

Training Configuration Table 9 shows the main training configurations for Visual Representation Learning and Map Representation Learning, including optimizer, learning rate, weight decay, and number of epochs. Visual Representation Learning trains the visual aggregation head and adapters with metric learning, while Map Representation Learning freezes the visual encoder and trains the map field to regress L2-normalized visual descriptors. All experiments were performed on Ubuntu 20.04.6 LTS using a single NVIDIA RTX 4090 GPU with 24 GB of memory, with CPU computations limited to a single thread by default.

Table 9: Training configurations for Visual Representation Learning and Map Representation Learning.

Module	Trainable	Optimizer	LR / WD / Epochs
Visual Representation Learning	Aggregator / Adapters	Adam	1e-4 / 1e-6 / 100
Map Representation Learning	Semantic Neural Field	Adam	1e-2 / 1e-6 / 1000

D Additional Experiments

Main Results under the Cross-area Setting As shown in Table 10, we present the performance of our method under the cross-area setting. Compared to its performance in the same-area setting (Table 3), our continuous 4DoF inference largely maintained recall and precision on Taizhou03 and Zurich, indicating that the method possesses strong cross-region generalization. Although some performance drop occurs in Taizhou04 and Zuchwil, this trend is not unique to our approach: most baseline methods also exhibit decreased recall and precision in these scenes, suggesting that the decline is primarily driven by scene-specific factors. Nevertheless, our method consistently achieves the highest R@1 under all three criteria and the lowest median position error among successful localizations across all four scenes. Furthermore, after applying GIM-based [21] local refinement, our method achieves the highest R@1 under all three criteria across all four scenes, as shown in Table 11. Qualitative results in the cross-area setting are shown in Figure 11.

Detailed Analysis of Dense Retrieval Libraries. Table 12 further examines the impact of reference library density, complementing Figure 4. Under the full $D\&H\&S$ criterion, increasing the

Table 10: Comparison of representative localization paradigms on four scenes in the cross-area setting.

	Taizhou03						Taizhou04					
	R@1 (100m, 10°, 1.2x) ↑			Median@Success ↓			R@1 (100m, 10°, 1.2x) ↑			Median@Success ↓		
Method	D	D&H	D&H&S	Dist	Head	Scale	D	D&H	D&H&S	Dist	Head	Scale
DenseUAV [1]	30.20	–	–	69.56	–	–	16.08	–	–	73.61	–	–
QDFL [4]	22.67	–	–	60.76	–	–	16.78	–	–	74.49	–	–
NetVLAD [59]	26.00	15.33	14.67	66.05	5.81	1.062	14.69	11.89	11.89	74.46	3.83	1.039
SALAD [58]	22.67	17.33	17.33	65.21	4.24	1.106	12.50	8.33	8.33	74.59	**2.19**	1.055
SeLaVPR [57]	26.67	18.67	18.67	66.05	**3.75**	1.070	15.38	13.29	13.29	74.46	2.27	1.063
PoseNet [35]	0.00	0.00	0.00	–	–	–	0.00	0.00	0.00	–	–	–
Ours	**87.33**	**74.67**	**74.67**	**36.30**	4.48	**1.051**	**58.04**	**53.15**	**53.15**	**41.65**	3.94	**1.038**

	Zurich						Zuchwil					
	R@1 (50m, 10°, 1.2x) ↑			Median@Success ↓			R@1 (50m, 10°, 1.2x) ↑			Median@Success ↓		
Method	D	D&H	D&H&S	Dist	Head	Scale	D	D&H	D&H&S	Dist	Head	Scale
DenseUAV [1]	19.83	–	–	24.21	–	–	18.24	–	–	33.97	–	–
QDFL [4]	27.62	–	–	23.56	–	–	24.48	–	–	27.07	–	–
NetVLAD [59]	32.22	17.99	8.79	29.49	4.46	1.099	31.72	17.93	13.45	27.86	4.07	1.042
SALAD [58]	30.54	14.23	9.62	29.31	3.71	1.114	30.00	16.21	13.45	26.20	5.13	1.040
SeLaVPR [57]	28.87	16.74	10.04	29.49	**3.07**	1.105	23.45	13.10	13.10	28.78	4.37	**1.025**
PoseNet [35]	0.00	0.00	0.00	–	–	–	2.07	0.34	0.34	33.80	8.76	1.187
Ours	**84.94**	**65.27**	**62.76**	**9.80**	3.10	**1.063**	**45.86**	**33.45**	**33.45**	**13.57**	**3.75**	1.159

Table 11: Effect of GIM$_{dkm}$-based local refinement on progressive recall under the cross-area setting.

	Taizhou03			Taizhou04			Zurich			Zuchwil		
Method	D	D&H	D&H&S	D	D&H	D&H&S	D	D&H	D&H&S	D	D&H	D&H&S
QDFL [4]	72.67	72.67	72.67	63.64	63.64	62.94	54.39	53.97	48.12	48.62	47.93	45.52
SeLaVPR [57]	75.33	75.33	75.33	51.05	51.05	50.35	56.90	55.23	48.54	45.86	45.86	44.48
Ours	**98.00**	**98.00**	**98.00**	**81.12**	**81.12**	**81.12**	**89.12**	**89.12**	**83.26**	**60.69**	**59.31**	**57.59**

sampling density improves dense-retrieval R@1 from 7.98% to 36.13% on Zurich and from 21.48% to 73.15% on Taizhou03. In comparison, our continuous inference achieves R@1 of 70.59% and 81.21%, respectively, with 98k initial hypotheses per scene. These results demonstrate higher recall than the densest retrieval configurations while using substantially fewer initial hypotheses.

Table 12: Recall@1 of dense retrieval under increasing gallery density on Zurich and Taizhou03.

Paradigm	Overlap	Zurich R@1 (50m, 10°, 1.2x)				Taizhou03 R@1 (100m, 10°, 1.2x)			
		Gallery ↓	D ↑	$D\&H$ ↑	$D\&H\&S$ ↑	Gallery ↓	D ↑	$D\&H$ ↑	$D\&H\&S$ ↑
Dense Ret.	0.000	98k	29.41	20.17	7.98	98k	28.19	22.82	21.48
	0.125	135k	31.93	23.11	10.50	120k	32.89	24.83	22.15
	0.250	187k	39.50	31.09	16.39	160k	40.94	32.89	30.87
	0.375	259k	55.04	42.86	19.33	240k	60.40	46.98	43.62
	0.500	408k	67.65	54.62	23.95	378k	75.17	61.74	54.36
	0.675	954k	79.83	68.07	30.25	900k	88.59	75.84	70.47
	0.750	1.65M	85.71	73.95	31.09	1.51M	90.60	77.85	71.14
	0.875	6.57M	89.92	79.41	36.13	5.98M	90.60	79.87	73.15
Ours	0.000	**98k**	**89.92**	**81.93**	**70.59**	**98k**	**91.28**	**81.21**	**81.21**

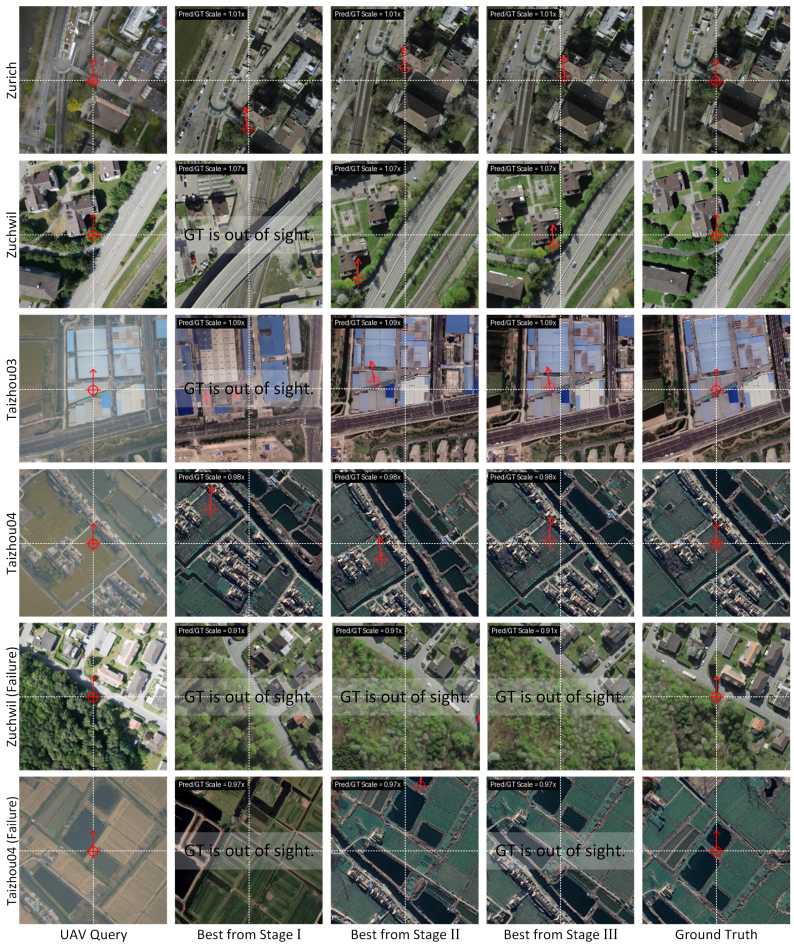


Figure 11: Qualitative visualization of stage-wise pose refinement in the cross-area setting. The red arrows mark the relative positions and orientations of the ground-truth query poses within the coordinate frames of the hypothesized reference views.

E Limitations

Despite achieving strong performance compared to existing paradigms, our method has several limitations: (1) Constructing a continuous map representation for a new map incurs an additional time overhead, ranging from 0.5 to 3 hours, depending primarily on the size of the map. (2) Fitting a map is, in fact, an underlying scientific challenge; not all feature spaces learned by visual encoders are amenable to such fitting. (3) Although our dataset encompasses a variety of real-world drone scenarios from different cities, only the "Zurich" and "Zuchwil" scenarios include flights conducted at varying altitudes; in the other scenarios, the range of flight altitudes remains relatively consistent. This is why our inference algorithm does not optimize for scale as the primary variable.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We explicitly summarize our core contributions in the abstract and Section 1 (Introduction), including the formulation of UAV geo-localization as a continuous 4DoF optimization via semantic neural fields, and the introduction of the ContinuScenes benchmark. The scope is clearly bounded to cross-view UAV localization, and our empirical claims match the experimental results presented in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our proposed approach in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: Our paper is an empirical methodology and dataset paper. We introduce the ContinuLoc framework and the ContinuScenes benchmark, but we do not propose new theoretical theorems or mathematical proofs that require validation.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We comprehensively detail the experimental setup required for reproducibility in Appendix C.1 and Appendix C.2. This includes explicit descriptions of the network architectures (e.g., DINOv2 and SALAD), the multi-resolution hash grid configurations for the semantic neural field, and the evaluation protocols.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Please visit our project homepage, which is publicly available at `https://anonymous.4open.science/r/Loc-ContinuLoc`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, all the necessary training and testing details—including data partitioning, optimizer type, and hyperparameters—are covered in Appendix C.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to the significant computational cost of training semantic neural fields and evaluating them across the large-scale ContinuScenes benchmark, running multiple training seeds to compute error bars is computationally prohibitive. However, our evaluation across multiple diverse scenes and increasingly strict 4DoF criteria sufficiently demonstrates the robustness and clear performance margins of our method.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: As stated in the supplementary materials, all experiments were conducted on an Ubuntu 20.04.6 LTS system using a single NVIDIA RTX 4090 GPU with 24 GB of memory, with CPU computations limited to a single thread by default. The time cost of the algorithm was also reported in detail in the experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper strictly conforms to the NeurIPS Code of Ethics. The development of the ContinuLoc framework and the ContinuScenes benchmark focuses on advancing foundational geo-localization techniques for academic purposes. We ensure that the dataset collection process respects privacy, and we do not promote any harmful, malicious, or unethical applications of UAV technologies.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Our work is technology oriented, with the main goal of improving the localization accuracy of UAV. It does not have a direct social impact, but we certainly hope that it can indirectly have a positive impact on the industry, economy, and society.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The data sources in this work are all legally collected.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Yes, all external assets (code, data, models) used in this paper are properly cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces a novel benchmark, the ContinuScenes dataset, and the ContinuLoc codebase. The dataset collection, automated pairing pipeline, and geospatial metadata statistics are extensively documented in Appendix A. A comprehensive README and usage guidelines are also included in the supplementary material alongside the assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: Our research focuses exclusively on UAV geo-localization algorithms and aerial imagery datasets (ContinuScenes). It does not involve any crowdsourcing experiments, manual data annotation by paid workers, or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our research focuses exclusively on UAV geo-localization algorithms and the processing of aerial imagery for the ContinuScenes benchmark. It does not involve any research with human subjects, making IRB approval or equivalent ethical reviews inapplicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology of our research relies exclusively on vision foundation models (e.g., DINOv2) and continuous neural fields for UAV geo-localization. Large Language Models (LLMs) are not utilized as an important, original, or non-standard component of our proposed ContinuLoc framework.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.