Towards Zero-Shot Global Localization via Dense Embedding Maps Based on Satellite Priors

Michal Bouda, Michal Werner, David Čapek, and Tomáš Báča
Faculty of Electrical Engineering, Czech Technical University in Prague
wernemic@fel.cvut.cz

***Abstract*— Effectively leveraging prior maps, such as top-down satellite imagery, remains a significant challenge for enabling robust robot perception and navigation at low altitudes (e.g., below 30 m) in GNSS-denied environments. Traditional visual matching struggles with the extreme viewpoint gap, while learned descriptors typically require extensive training on specific environments. In this paper, we explore the extraction of open-vocabulary semantic priors from remote sensing maps using zero-shot Vision Foundation Models (VFMs). Rather than matching raw pixels, we abstract dense semantic embeddings from onboard sensors and aerial priors into discrete spatial semantic regions. This representation forms a lightweight topology that allows for robust, neighborhood-based structural matching. Validated on multiple custom real-world flights and on a real-world dataset, our preliminary findings demonstrate how structuring semantic priors can overcome perspective constraints to achieve training-free semantic UAV geolocalization.**

I. Introduction

Autonomous localization and navigation for low-flying Unmanned Aerial Vehicles (UAVs) in GNSS-denied environments is an open challenge with practical relevance in scenarios such as search-and-rescue and inspection in areas with active interference. In these applications, onboard localization must often rely on globally accessible prior maps, such as aerial or satellite imagery [1]. However, utilizing top-down satellite priors is exceptionally challenging for low-flying robots (e.g., below 30 m AGL) [2]. At these altitudes, the extreme viewpoint mismatch between onboard sensors capturing vertical facades and orthorectified satellite priors capturing building roofs renders traditional direct visual or geometric matching [3]–[5] ineffective.

Recent cross-view geo-localization methods attempt to bridge this perspective gap using learned descriptors, but they predominantly require extensive training on paired image datasets from specific environments [6]–[8]. To overcome the reliance on environment-specific training, zero-shot Vision Foundation Models (VFMs) have emerged as powerful tools, providing robust, class-agnostic semantic embeddings across diverse domains [9], [10]. However, VFMs natively output dense, high-dimensional feature tensors. Processing and matching these continuous embeddings directly at the pixel level is computationally prohibitive for onboard UAV hardware and lacks the discrete topological structure necessary for efficient global search algorithms [11]. To address this, we reframe cross-view localization around the efficient processing of semantic features using a structured data representation [12], [13].

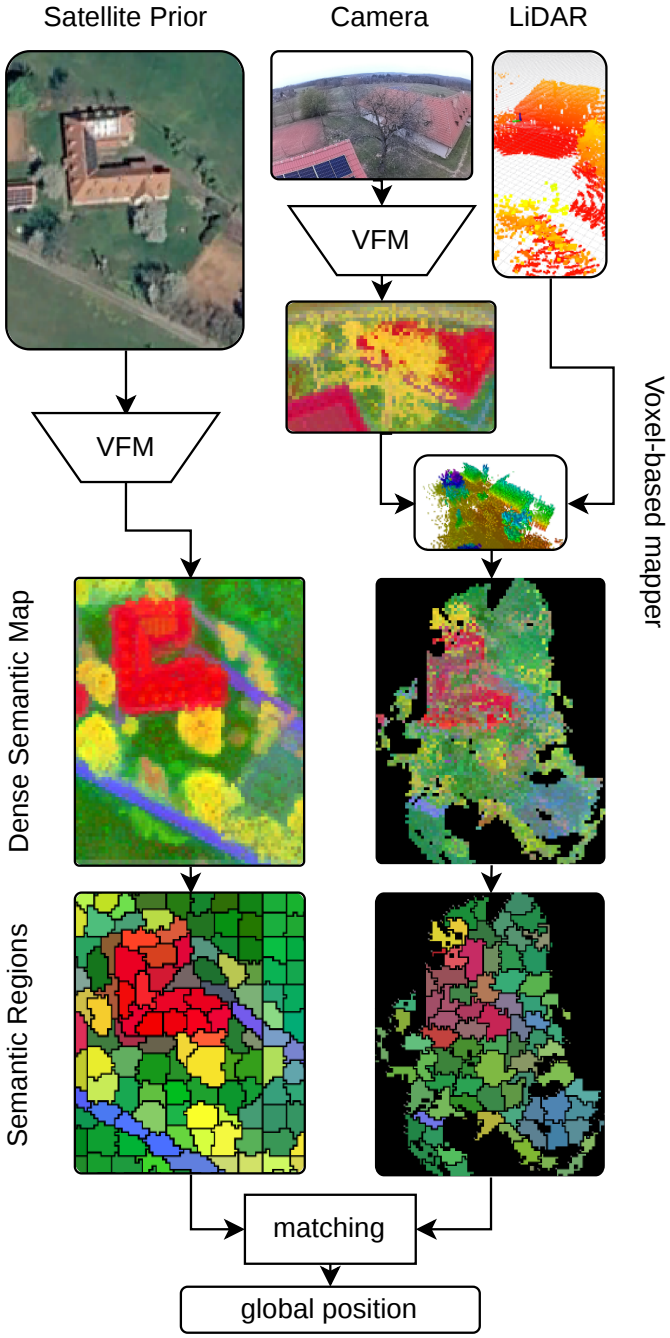


Fig. 1: Overview of the proposed localization pipeline. The dense embedding vectors are projected into RGB space using PCA projection fitted on the satellite prior.

Rather than performing an exhaustive search over raw dense maps, our system abstracts semantic data from both the onboard voxel-based map and the satellite reference into discrete spatial *semantic regions*. This structured abstraction drastically reduces the memory footprint and converts flat embedding maps into efficient, region-based topology. By grouping dense features into this structured representation, we enable structurally-aware correspondence search across different viewpoints without requiring any fine-tuning.

A. Contributions

We present preliminary results of our zero-shot cross-view localization pipeline, featuring the following contributions:

- A novel global localization pipeline utilizing zero-shot Vision Foundation Models to bridge the perspective gap without environment-specific retraining.
- An efficient, region-based map representation that mitigates outdoor semantic aliasing by abstracting dense embeddings into meaningful spatial structures.
- Preliminary real-world experiments demonstrating the feasibility of matching onboard semantic region maps to globally available aerial priors.

II. Method

A. System Overview

The overall architecture of the proposed localization pipeline is illustrated in Figure 1. The system continuously fuses multimodal onboard sensor data to construct a local semantic map, which is subsequently matched against a globally available aerial prior.

B. Onboard Semantic Mapping

To construct the local semantic representation, RGB images from the onboard camera are processed through the RADIO [14] VFM equipped with a SigLIP-2 [15] adaptor, yielding dense, per-patch semantic embeddings. These 2D semantic features are spatially registered by projecting them onto the 3D point cloud acquired by the onboard LiDAR. To ensure computational efficiency during subsequent mapping and matching stages, the native high-dimensional embedding vectors are compressed into a 512-dimensional space via a precomputed linear projection. The resulting semantic point cloud (where each 3D point is associated with its nearest patch-semantic vector in the image) is then incrementally fused into a volumetric map. This mapping module leverages a VDB-tree data structure to efficiently maintain the spatial environment. Within this volumetric representation, each occupied voxel stores a single 512-dimensional semantic descriptor, which is continuously updated via temporal averaging as new sensor measurements are integrated. To align the local 3D map with the 2D aerial prior, the voxel map is projected onto a top-down grid. At each spatial coordinate, only the semantic vector from the highest occupied voxel is retained. This isolates the topmost visible surfaces, such as tree canopies and building roofs, effectively mimicking the satellite's nadir perspective.

C. Prior map processing

For satellite imagery, an inference pipeline processes the image at a resolution of one feature vector per square meter of ground area, using overlapping tiles with a symmetric crop-margin strategy to ensure all resulting pixels were processed with enough context. The output is a dense feature tensor of shape $H \times W \times 512$, where each spatial location encodes a 512-dimensional semantic vector and H with W are the height and width of the whole satellite area in meters.

D. Semantic regions

Both the drone and satellite feature vector maps are clustered into semantic regions, in order to lower the number of primitives and reduce the significant memory footprint of the unsegmented maps. A target semantic region area of $50\,\mathrm{m}^2$ is used, providing enough spatial extent to capture meaningful semantic regions such as individual trees or building footprints. Each region S_i is represented by its barycenter X_{S_i} and its intra-region descriptor R_i, defined as the cosine-distance medoid of all feature vectors F within the region:

$$R_i = \arg\min_{f_i \in F} \frac{1}{|F|} \sum_{f_j \in F} \left(1 - \frac{f_i^\top f_j}{\|f_i\|_2 \|f_j\|_2}\right) \tag{1}$$

Unlike a mean vector, the medoid is an observed output of the model that has the lowest dissimilarity to all vectors in the region, preserving the properties of vectors the VFM actually produces.

E. Matching of the semantic regions

To robustly localize the UAV without the computational burden of an exhaustive search, we leverage an iterative, neighborhood-based optimization approach adapted for semantic regions. The system evaluates spatial correspondences by requiring alignment of both the internal semantic content of a region and its external topological context, effectively filtering out visually similar but structurally inconsistent aliases. By exploiting the physical rigidity of the locally mapped environment, adjacent regions iteratively propagate their optimal spatial alignments through the map's network. This cooperative, local-to-global optimization enables structurally accurate matches to rapidly cascade across the graph, naturally resolving ambiguities and bridging extreme perspective shifts to converge on a single, unified global position estimate suitable for resource-constrained robotic applications.

III. Preliminary Experiments

The proposed method has been validated using real-world dataset UAVScenes [16] and our own real-world experiments.

A. UAVScenes Dataset

Most works focusing on GNSS-denied global localization rely on camera images as their primary data source. Datasets providing both LiDAR and camera data for low-flying aerial vehicles are severely underrepresented. One

Context Level	Success Rate (Error ≤ 15 m)	Matched
Size 50 m		
High Context	62.1%	18/29
Low Context	46.4%	13/28
No Context	25.0%	7/28
Size 100 m		
High Context	83.7%	36/43
Low Context	68.2%	15/22
No Context	33.3%	4/12
Size 150 m		
High Context	86.4%	19/22
Low Context	55.6%	5/9
No Context	14.3%	1/7
Total (All Sizes)		
High Context	77.7%	73/94
Low Context	55.9%	33/59
No Context	25.5%	12/47

TABLE I: Localization success rates across context levels and submap sizes for the proposed localization pipeline.

notable exception is the UAVScenes dataset [16], which provides both LiDAR and top-down onboard camera data. For our statistical evaluation, we utilize a subset of this dataset comprising flights AMtown01, AMvalley01, HKisland01 in both urban and non-urban environments. The dataset contains multiple flights at altitudes of approximately 30 to 60 m Above Ground Level (AGL), recording data from a down-facing RGB camera and a LiDAR. For the evaluation, a dense semantic map of the entire flight is first constructed using our mapping pipeline. Map tiles of 50 m, 100 m, and 150 m are then cropped and localized using the proposed method. To better assess localization performance, the 200 map tiles are manually classified into three categories based on the amount of semantic context present in the scene. A *High Context* scene represents an area with multiple buildings, intersections, or a distinct combination of semantic objects. Conversely, *Low Context* scenes represent repetitive areas lacking clear localization markers. Semantic maps were matched against satellite images, which were all over 600 m tall and over 1100 m wide.

Success rates for our proposed pipeline are shown in Table I, with visual examples in Figure 2. The results reveal a strong correlation between localization success, the available semantic context, and the size of the queried submap. For *High Context* environments, increasing the map size from 50 m to 150 m substantially improves the success rate from 62.1% to 86.4%. Overall, the experiments validate our hypothesis that the proposed lightweight abstraction of dense semantic maps can be used for global localization, although further validation on more datasets is needed.

Conversely, the performance degrades significantly in *No Context* environments. Interestingly, while larger maps generally improve localization, expanding the map size in *No Context* areas actually decreases the success rate (dropping from 25.0% at 50 m to 14.3% at 150 m). This indicates that accumulating large areas of repetitive, non-distinct semantic regions (such as flat fields or uniform forest canopies) increases the likelihood of structural aliasing, leading the system to converge on false positive matches.

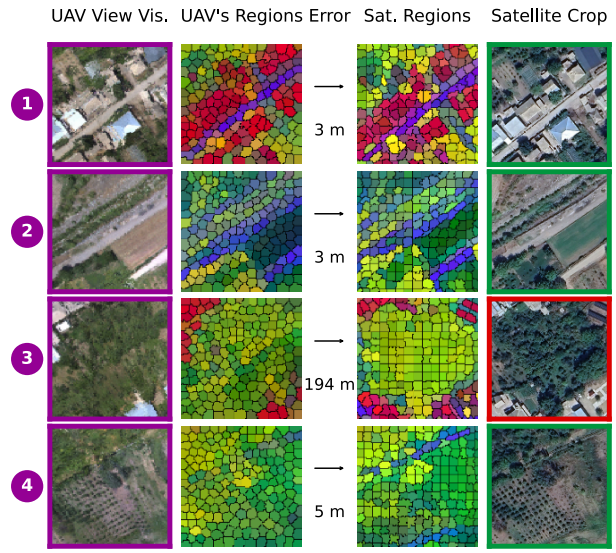
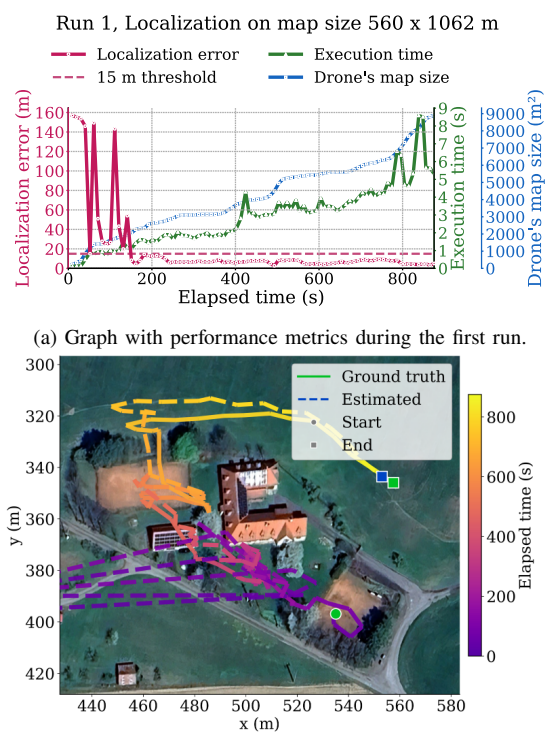


Fig. 2: Examples of correct and incorrect localization matches on the UAVScenes dataset, including the PCA visualisation of the semantic regions.

B. Real-world Flights

We also evaluated the pipeline on continuous real-world flight trajectories. Metrics, including localization error, matching duration, and accumulated drone map size, were recorded throughout the duration of each flight.



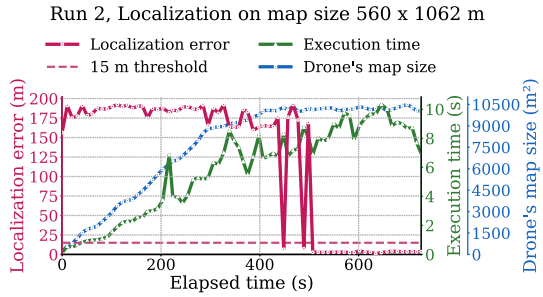
(a) Graph with performance metrics during the first run.

(b) Map with trajectories of the first run.

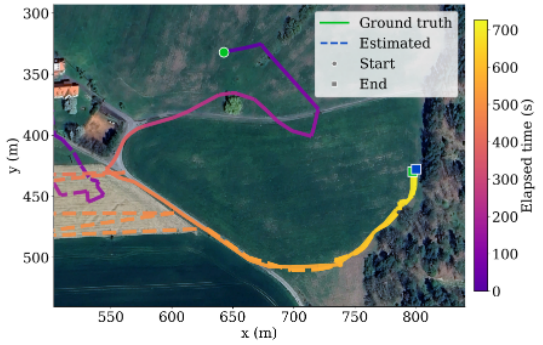
Fig. 3: Overview of Run 1.

As observed in Figure 3 (Run 1), the localization error rapidly converges below the 15 m threshold once a sufficient map size is accumulated, remaining below 10 m for the majority of the flight. Notably, the current implementation does not employ any probabilistic filtering. Therefore, these results represent pure, single-epoch global localization without reliance on prior state estimates or odometry integration. The onboard and prior semantic maps from this flight are visualized in Figure 1.

Run 2 (Figure 4) demonstrates how the method consistently achieves correct global alignment as soon as the drone's aggregated map captures sufficient structural scene context.



(a) Graph with performance metrics during the second run.



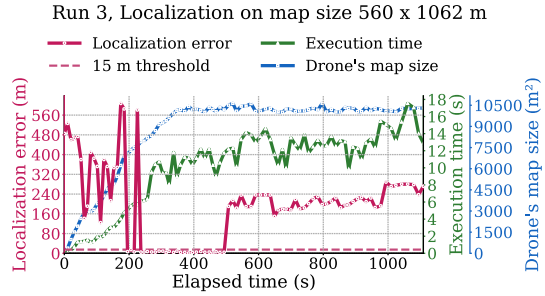
(b) Map with trajectories of the second run.

Fig. 4: Overview of Run 2.

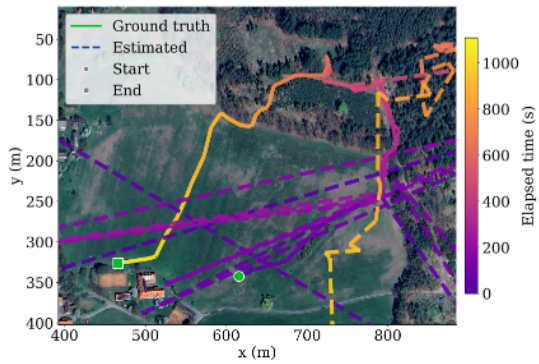
Conversely, Run 3 (Figure 5) highlights the operational limits of the current system in highly unstructured environments. The method successfully localized the UAV between 240 s and 490 s while flying above a forest road, which we hypothesize provided a uniquely discriminative structural feature within an otherwise repetitive landscape. However, outside this window, the absence of distinct geometric or semantic boundaries led to severe spatial aliasing, resulting in significant degradation of localization accuracy.

IV. Discussion

The preliminary results provide insights into the feasibility and limitations of the proposed semantic localization. Bridging the perspective gap using zero-shot VFM embeddings is achievable, but it depends heavily on the structural uniqueness of the environment. As evidenced by Table I and Run 3, the method achieves global alignment in *High Context* areas with distinct topologies, such as buildings or isolated roads.



(a) Graph with performance metrics during the third run.



(b) Map with trajectories of the third run.

Fig. 5: Overview of Run 3.

Conversely, performance drops in *No Context* environments define the operational limits of purely semantic region abstraction. In repetitive landscapes like uniform forests or flat fields, semantic clusters lack the distinctiveness needed to prevent structural aliasing.

We utilize the RADIO VFM paired with a SigLIP-2 adaptor. While prioritizing high-level semantics over visual texture successfully overcomes perspective differences in structured areas, our findings suggest it may be insufficient for featureless terrain. Despite current localization limitations, these abstracted maps offer significant downstream potential. We hypothesize their underlying semantic structure could support broader navigation and scene understanding tasks. Ultimately, this approach provides a versatile framework for exploring more advanced autonomous capabilities.

V. Conclusion and Future Work

This workshop paper presented preliminary results for a novel, zero-shot cross-view localization pipeline for low-flying UAVs. By leveraging Vision Foundation Models to abstract dense features into semantic regions, we demonstrated the feasibility of matching onboard sensor data to aerial priors without environment-specific training.

Ongoing work is focused on formalizing the region-matching algorithm, benchmarking the system across diverse environmental and seasonal conditions, and establishing comparisons with state-of-the-art methods and an ablation study with other Vision Foundation Models. We look forward to community feedback to further refine the pipeline as we work toward a full publication detailing the complete mathematical framework and comprehensive quantitative results.

References

[1] A. Durgam, S. Paheding, V. Dhiman, and V. Devabhaktuni, "Cross-view geo-localization: A survey," *IEEE Access*, vol. 12, pp. 192 028–192 050, 2024.

[2] D. Avola, L. Cinque, E. Emam, F. Fontana, G. Luca Foresti, M. Raoul Marini, A. Mecca, and D. Pannone, "UAV Geo-Localization for Navigation: A Survey," *IEEE Access*, vol. 12, pp. 125 332–125 357, 2024.

[3] H. Goforth and S. Lucey, "GPS-Denied UAV Localization using Pre-existing Satellite Imagery," in *2019 International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 2974–2980.

[4] A. Viswanathan, B. R. Pires, and D. Huber, "Vision based robot localization by ground to satellite matching in GPS-denied situations," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sep. 2014, pp. 192–198.

[5] R. Kümmerle, B. Steder, C. Dornhege, A. Kleiner, G. Grisetti, and W. Burgard, "Large scale graph-based SLAM using aerial images as prior information," *Autonomous Robots*, vol. 30, no. 1, pp. 25–39, Jan. 2011.

[6] S. Hu, M. Feng, R. M. H. Nguyen, and G. H. Lee, "CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7258–7267.

[7] S. Zhu, M. Shah, and C. Chen, "TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization," Mar. 2022.

[8] Y. He, I. Cisneros, N. Keetha, J. Patrikar, Z. Ye, I. Higgins, Y. Hu, P. Kapoor, and S. Scherer, "FoundLoc: Vision-based Onboard Aerial Localization in the Wild," Oct. 2023.

[9] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "AnyLoc: Towards Universal Visual Place Recognition," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1286–1293, Feb. 2024.

[10] Y. Tao, D. Ong, F. Cladera, J. Hughes, C. J. Taylor, P. Chaudhari, and V. Kumar, "HALO: High-Altitude Language-Conditioned Monocular Aerial Exploration and Navigation," Nov. 2025.

[11] N. Hughes, Y. Chang, S. Hu, R. Talak, R. Abdulhai, J. Strader, and L. Carlone, "Foundations of spatial perception for robotics: Hierarchical representations and real-time systems," *The International Journal of Robotics Research*, vol. 43, no. 10, pp. 1457–1505, Sep. 2024.

[12] A. Gawel, C. D. Don, R. Siegwart, J. Nieto, and C. Cadena, "X-View: Graph-Based Semantic Multi-View Localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1687–1694, Jul. 2018.

[13] S. D. Sarkar, O. Miksik, M. Pollefeys, D. Barath, and I. Armeni, "SGAligner: 3D Scene Alignment with Scene Graphs," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, pp. 21 870–21 880.

[14] M. Ranzinger, G. Heinrich, J. Kautz, and P. Molchanov, "AM-RADIO: Agglomerative Vision Foundation Model – Reduce All Domains Into One," Apr. 2024.

[15] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai, "SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features," Feb. 2025.

[16] S. Wang, S. Li, Y. Zhang, S. Yu, S. Yuan, R. She, Q. Guo, J. Zheng, O. K. Howe, L. Chandra, S. Srijeyan, A. Sivadas, T. Aggarwal, H. Liu, H. Zhang, C. Chen, J. Jiang, L. Xie, and W. P. Tay, "UAVScenes: A Multi-Modal Dataset for UAVs," Jul. 2025.