Uncertainty-Aware Hierarchical Re-Localization in OpenStreetMap via Semantic Alignment

Yuchen Zou, Xiao Hu, Lihuang Fang, Yuqing Tang*

***Abstract*— Monocular re-localization enables robots to estimate camera poses from visual observations. However, many existing methods rely on dense maps or large reference image databases, which face scalability limitations and privacy risks. OpenStreetMap (OSM), as a lightweight privacy-preserving map, offers semantic and geometric information with global scalability. Nonetheless, OSM localization remains challenging due to cross-modal discrepancies between natural images and OSM, as well as the high cost of global map-based localization. In this paper, we propose an uncertainty-aware hierarchical search framework with semantic alignment for localization in OSM. First, object-centric DINO-ViT tokens are exploited to reduce the semantic gap between ground-view observations and OSM vectors. Second, global dense matching is decomposed into coarse FFT correlation and uncertainty-controlled local refinement. Extensive experiments demonstrate that our method significantly improves localization accuracy and speed. When trained on a single dataset, the 3° orientation recall of our method even outperforms the 5° recall of state-of-the-art methods.**

I. Introduction

Monocular visual re-localization estimates camera pose from a single image, serving as a core technology for autonomous driving [1], navigation [2], and virtual reality [3]–[5]. Compared to LiDAR and multi-camera or stereo systems, monocular systems offer advantages in hardware cost and deployment flexibility, making them scalable for commercial applications.

Traditional monocular localization often relies on dense maps [6] or extensive reference image databases [7], [8]. These representations face scalability issues due to construction and maintenance costs [9], alongside privacy leakage risks during data acquisition [10]. Consequently, OpenStreetMap (OSM) has emerged as a compelling alternative. As an open-source vector map, OSM continuously updates building footprints and road topologies [11], providing rich, privacy-friendly semantic and geometric priors for localization [12].

Despite its advantages, accurate localization on OSM remains extremely challenging. Monocular images capture perspective surface textures, illumination, and viewpoint variations, whereas OSM provides structured, top-down geometric abstractions. This severe modality gap makes direct feature matching highly ineffective. Furthermore, performing exhaustive fine-grained search across large-scale OSM regions incurs prohibitive computational costs, failing to meet the real-time requirements of robotic systems [13].

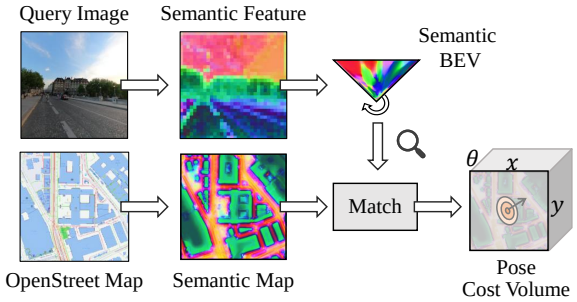

Fig. 1. Localization in the OSM-derived semantic map by transforming the query image into semantic features.

To address these hurdles, we re-formulate OSM-based localization as a cross-modal semantic alignment problem. Leveraging the object-centric semantic tokens of DINO-ViT [14], we decouple complex ground-view features to align with the vector semantics and geometric elements of OSM (as shown in Fig. 1). Moreover, we depart from traditional global dense matching paradigms [13], [15] by decomposing search into coarse FFT correlation and uncertainty-aware local refinement. This hierarchical approach progressively refines the location and heading direction, reducing computational overhead while ensuring high registration accuracy.

The main contributions are summarized as follows:

- We exploit object-centric DINO-ViT tokens to bridge the cross-modal gap between ground-view observations and OSM vector semantics.
- We propose an uncertainty-aware coarse-to-fine localization mechanism that converts global dense matching into coarse FFT correlation and adaptive local refinement.
- Experiments on street-view and autonomous driving datasets demonstrate that this mechanism triples localization speed. Remarkably, when trained on a single dataset, our 3° orientation recall outperforms the 5° recall of state-of-the-art methods.

II. Related Work

A. Monocular Re-Localization in OSM

Monocular re-localization has evolved from image retrieval [16]–[18] and map-based visual localization [19], [20] toward leveraging OpenStreetMap (OSM) [21] to circumvent

This work was completed during Y. Zou's internship at the International Digital Economy Academy.

Y. Zou is with the School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China.

X. Hu and Y. Tang are with the International Digital Economy Academy, Guangdong, Shenzhen 510085, China.

L. Fang is with the Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen 518055, China.

* Corresponding authors (e-mail: tangyuqing@idea.edu.cn).

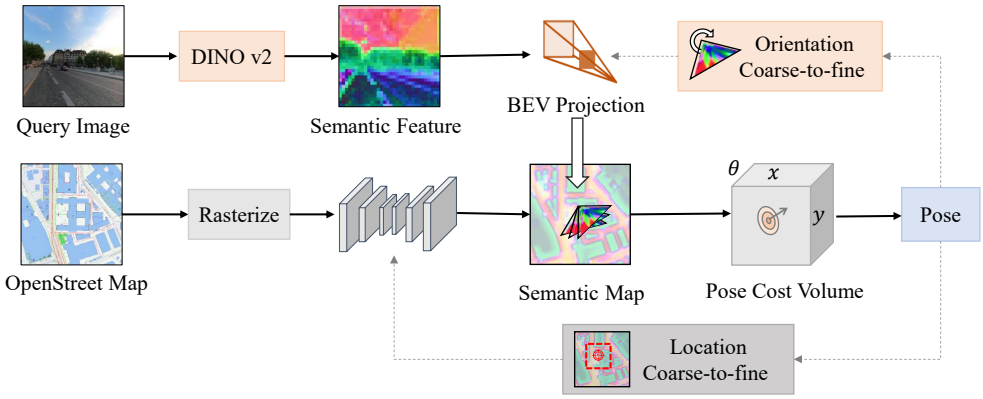

Fig. 2. Overview of the proposed coarse-to-fine OSM localization framework.

high maintenance costs and privacy issues. As a lightweight vector database, OSM provides structured geometric priors like road networks and building footprints [11], [22]–[24]. Although frameworks like OrienterNet [13] use OSM to build semantic neural maps, they struggle with the modality gap between ground-level perspective views and overhead vector abstractions. Furthermore, the computational burden of dense matching across extensive OSM regions limits real-time deployment.

B. Hierarchical and Neural Map-based Localization

To balance efficiency and precision, hierarchical localization typically adopts a progressive coarse-to-fine refinement paradigm. This approach utilizes global retrieval to provide an initial location hypothesis [25], which is subsequently verified and refined through local feature matching [26], [27]. In the domain of neural maps, MapLocNet [28] attempts to regress camera poses directly from heterogeneous map data. Nevertheless, direct regression frequently lacks the necessary robustness when addressing the high-frequency geometric variations found in single-view imagery. In contrast, our framework employs a matching-based hierarchical strategy. This method preserves the interpretability of geometric alignment while benefiting from the optimization capabilities of differentiable neural cost fields.

C. Semantic Representation via Vision Transformers

Vision Transformers (ViT) [29], particularly self-supervised models like DINO [14], [30], extract robust, object-centric representations invariant to lighting or texture fluctuations [31], [32]. Unlike convolutional features that prioritize local textures, the tokens generated by DINO are aware of global context and naturally decouple essential urban elements such as building facades and road boundaries. These semantic primitives correspond directly to the area and line features defined within OSM. By adopting DINO as a semantic backbone, we establish a robust intermodal bridge that aligns visual observations with geometric constraints from the map, ensuring strong generalization in complex, unconstrained environments.

III. Method

Given a query image I and a candidate local OSM map crop, we aim to estimate the 3-DoF camera pose $\xi = (x, y, \theta)$ within the map crop. Our framework aligns both the monocular observation and the local OSM map crop within a unified semantic space to perform robust registration (Fig. 2).

A. Semantic BEV Projection

We employ DINO to extract high-level semantic features $F_I \in \mathbb{R}^{U \times V \times N}$ from the query image I. To preserve pretrained semantics while adapting high-level features to localization, we freeze the first 8 layers of the DINO encoder and fine-tune the final 4 layers.

Subsequently, F_I is projected into the Bird's-Eye View (BEV) space using a monocular scale-distribution method [13]. Specifically, F_I is first mapped to polar coordinates $F_{polar} \in \mathbb{R}^{U \times D \times N}$ via depth-based ray interpolation:

$$F_{polar}(u,d,n) = \sum_{v} \alpha_{(d,v)} \cdot F_I(u,v,n), \tag{1}$$

where $\alpha_{(d,v)}$ denotes the interpolation probability between image column v and depth plane d. The polar feature is then horizontally resampled onto a Cartesian grid to yield the semantic BEV representation $F_{bev} \in \mathbb{R}^{L \times D \times N}$:

$$F_{bev}(l,d,n) = \sum_{u} \alpha_{(u,l)} \cdot F_{polar}(u,d,n), \tag{2}$$

where $\alpha_{(u,l)}$ is the transformation ratio determined by the camera intrinsics.

B. Semantic Map Encoding

OSM provides structured vector elements classified into areas (buildings), lines (roads), and points (POIs). To enable dense feature matching, we crop a local OSM map and project these discrete categorical vectors into a continuous feature space via a learnable embedding layer. Following [28], a U-Net architecture is employed to generate a rasterized neural semantic map $F_{map} \in \mathbb{R}^{W \times H \times N}$. This encoding ensures the channel dimension aligns with F_{bev} while preserving topological priors.

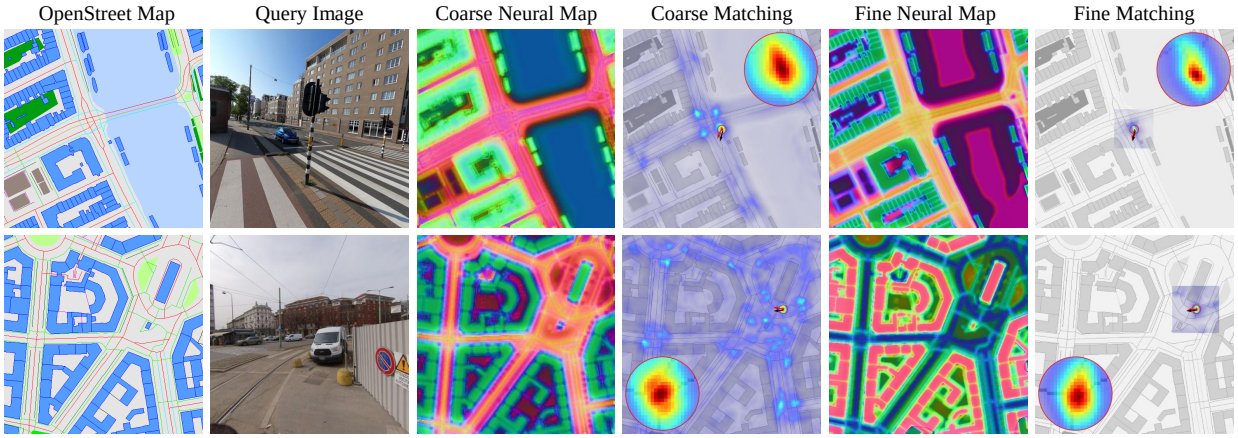

Fig. 3. Visualization of coarse-to-fine localization, including OSM neural-map refinement and pose-likelihood refinement. The red and black arrows denote the predicted and ground-truth poses.

C. Coarse-to-Fine Pose Estimation

Exhaustive global dense matching incurs prohibitive computational costs. To balance efficiency and accuracy, we implement a hierarchical correlation strategy.

Coarse Matching: We generate a discretized set of rotationally equivariant BEV features using a coarse angular step $\Delta\theta^c$:

$$\tilde{F}^c_{bev}(\theta) = \mathbf{R}(\theta) \cdot F^c_{bev}, \quad \theta \in \{0, \Delta\theta^c, \ldots, 2\pi - \Delta\theta^c\}, \quad (3)$$

where $\mathbf{R}(\theta) \in SO(2)$ is the rotation operator. Concurrently, low-resolution map features F^c_{map} are extracted from the U-Net bottleneck. We compute a 3-DoF correlation cost field $M(p,\theta)$ via the Fourier domain:

$$M(\cdot,\theta) = \mathscr{F}^{-1}\left[\mathscr{F}(F^c_{map}) \odot \mathscr{F}(\tilde{F}^c_{bev}(\theta))^*\right], \quad (4)$$

where $\mathscr{F}$ and $\mathscr{F}^{-1}$ denote 2D Fast Fourier Transform and its inverse, respectively, and $*$ is the complex conjugate. The optimal coarse pose hypothesis (p^c, θ^c) is obtained via Maximum Likelihood Estimation evaluated at point p:

$$(p^c, \theta^c) = \arg\max_{(p \in P, \theta \in \Theta)} M(p,\theta). \quad (5)$$

Fine Refinement: To adaptively constrain the high-resolution search space without compromising accuracy, we introduce a pose uncertainty measure U evaluated over the local neighborhood $\mathscr{N}(p^c, \theta^c)$:

$$U = 1 - \frac{\exp(M(p^c,\theta^c))}{\sum_{(p,\theta)\in\mathscr{N}} \exp(M(p,\theta))}. \quad (6)$$

This inverse-confidence metric dynamically contracts the fine-grained search space for orientation Θ^f and position P^f:

$$\Theta^f = [\theta^c - k_\theta U \Delta\theta^f, \theta^c + k_\theta U \Delta\theta^f], \quad (7)$$

$$P^f = [p^c - k_p U, p^c + k_p U], \quad (8)$$

where k_θ and k_p are scaling factors, and $\Delta\theta^f$ is the base angular radius of the fine search window. As shown in Fig. 3, uncertainty-aware refinement removes redundant spatial queries while maintaining registration accuracy.

D. Training Objective

To enable end-to-end optimization of the semantic feature extractor and the neural map encoder, we formulate pose estimation as a probability density estimation problem. Given the ground truth pose $\xi_{gt} = (p_{gt}, \theta_{gt})$, we supervise both the coarse and fine matching stages using a Negative Log-Likelihood (NLL) loss. The total training objective is defined as:

$$\mathscr{L} = \mathscr{L}_{coarse}(\xi_{gt}, M^c) + \lambda \mathscr{L}_{fine}(\xi_{gt}, M^f), \quad (9)$$

where λ is a balancing hyperparameter. For each stage, the loss minimizes the negative log-probability of the ground truth pose, derived by applying a softmax function over the position-orientation correlation volume. Jointly optimizing this objective maximizes the correlation response at true geometric intersections and suppresses false-positive alignments in repetitive urban environments.

IV. Experiments

A. Datasets and Implementation Details

We evaluate our framework on two benchmarks: the Mapillary Geo-Localization (MGL) dataset [13] and the KITTI odometry dataset [36]. MGL provides diverse street-level images across various urban environments. We extract corresponding local OSM map crops using the noisy GPS priors provided by the MGL metadata. For KITTI, which focuses on autonomous driving scenarios, we utilize the Test2 split [13], [35] to ensure zero geographic overlap with the MGL training set. Local OSM map crops are sampled within a 20 meter radius of the attached GPS coordinates.

Semantic extraction utilizes DINOv2 with registers [30]. OSM elements are quantized into 7 area, 10 line, and 33 point categories. Map encoding is performed via a VGG16 based U-Net architecture. During the hierarchical estimation, the coarse stage employs a 1 meter spatial resolution and a 6° angular interval, while the fine stage refines the search space to 0.5 meters and 2°. During training, spatial samples are drawn uniformly within a 32 meter radius of the ground truth. For MGL testing, GPS metadata is used only to define

TABLE I

LOCALIZATION PERFORMANCE ON KITTI USING SATELLITE AND OSM MAPS, WITH MODELS TRAINED ON KITTI OR MGL+KITTI.

Map	Method	Train dataset	Lateral R@Xm			Longitudinal R@Xm			Orientation R@X°		
			1m	3m	5m	1m	3m	5m	1°	3°	5°
Satellite	DSM [33]	KITTI	10.77	31.37	48.24	3.87	11.73	19.50	3.53	14.09	23.95
	VIGOR [34]	KITTI	17.38	48.20	70.79	4.07	12.52	20.14	5.46	15.79	27.15
	Refinement [35]	KITTI	27.82	59.79	72.89	5.75	16.32	26.48	18.42	49.72	71.00
OSM	OrienterNet [13]	KITTI	51.26	84.77	91.81	22.39	46.79	57.81	20.41	52.24	73.53
	Ours	KITTI	**68.05**	**92.84**	**95.81**	**27.45**	**57.89**	**68.34**	**37.28**	**80.16**	**92.12**
	OrienterNet [13]	MGL+KITTI	65.91	92.76	96.54	33.07	65.18	75.15	35.72	77.49	91.51
	Ours	MGL+KITTI	**72.28**	**94.42**	**97.49**	**37.87**	**71.40**	**79.61**	**42.83**	**87.26**	**96.37**

TABLE II

LOCALIZATION PERFORMANCE ON MGL. 'C' AND 'F' DENOTE COARSE AND FINE LOCALIZATION.

Method	Position R@Xm			Orientation R@X°			FPS
	1m	3m	5m	1°	3°	5°	
Retrieval	2.02	15.21	24.21	4.50	18.61	32.48	-
Refinement	8.09	26.02	35.31	14.92	36.87	45.19	-
OrienterNet	15.78	47.75	58.98	22.14	52.56	66.32	7.51
Ours-C	14.70	49.58	64.43	23.26	59.75	78.99	**29.54**
Ours-F	**17.36**	**53.67**	**65.81**	**30.66**	**68.71**	**80.53**	25.80

TABLE III

IMPACT OF DINO AND COARSE-TO-FINE (C2F) MATCHING ON LOCALIZATION PERFORMANCE.

Method	Position R@Xm			Orientation R@X°			FPS
	1m	3m	5m	1°	3°	5°	
Baseline	14.37	48.69	61.7	20.95	54.96	70.17	7.34
+ DINO	17.06	53.17	65.31	24.54	61.28	76.35	6.81
+ C2F	**17.36**	**53.67**	**65.81**	**30.66**	**68.71**	**80.53**	**25.80**

the queried 128×128 meter local OSM map crop, and no pose prior is used within the map crop. For KITTI testing, we incorporate an initial pose prior within a $\pm 10°$ window. Models are trained on three NVIDIA RTX 4090 GPUs, and all FPS results are measured on the same hardware.

B. Comparison Methods

We compare our framework with three categories of localization methods. For KITTI, we include satellite-based cross-view baselines, including DSM [33], VIGOR [34], and the refinement-based method in [35]. For MGL, we report retrieval-based and refinement-based monocular baselines [25], [37], [38]. Although newer OSM-based models such as [28] have been proposed, they are primarily optimized for multi-view autonomous-driving scenarios. Consequently, we prioritize comparison with OrienterNet, the open source state-of-the-art for monocular OSM localization.

C. Performance on KITTI

Evaluation on KITTI dataset (Table I) demonstrates the robust generalization of our semantic representations to unseen geographic regions. We assess models trained exclusively on KITTI alongside those fine tuned from MGL pre training. Our framework consistently improves across all spatial and angular metrics when compared to both satellite- and OSM-based paradigms. Notably, when trained solely on KITTI, our method achieves a 3° orientation recall that exceeds the less stringent 5° recall of the prior leading approach.

D. Performance on MGL

Table II details our performance on the MGL dataset. The proposed hierarchical strategy achieves substantial gains in both registration accuracy and inference efficiency. By dynamically pruning the search space, our full pipeline operates at 25.80 FPS on the same hardware, tripling the throughput of OrienterNet. Furthermore, the combination of DINO semantic alignment and dense local refinement enhances heading estimation, improving the rigorous 1° orientation recall by 38.48% relative to existing benchmarks.

V. ABLATION STUDIES

To isolate the contributions of our core components, we establish a baseline utilizing a ResNet101 feature extractor coupled with global dense matching. As detailed in Table III, integrating the DINO backbone significantly boosts pose recall, confirming its superiority in extracting reliable map-aligned semantics, albeit introducing a marginal computational overhead. The subsequent addition of the hierarchical matching strategy not only mitigates this overhead by increasing inference speed threefold, but also enables high-resolution angular sampling. This structural optimization yields clear improvements, particularly in strict orientation recall metrics.

VI. CONCLUSION

We presented a hierarchical semantic alignment framework designed for monocular pose estimation within OpenStreetMap. By extracting globally aware object semantics via Vision Transformers, we explicitly bridged the severe modality gap existing between ground-level perspectives and overhead vector maps. Coupled with an uncertainty-aware search strategy, our approach achieves strong accuracy while satisfying the real-time computational constraints of autonomous systems. This research demonstrates the viability of utilizing lightweight vector maps for highly scalable and privacy preserving urban localization.

REFERENCES

[1] T. Wen, K. Jiang, B. Wijaya, H. Li, M. Yang, and D. Yang, "Tm^{3}loc: Tightly-coupled monocular map matching for high precision vehicle localization," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 20 268–20 281, 2022.

[2] D. Maggio, M. Abate, J. Shi, C. Mario, and L. Carlone, "Loc-nerf: Monte carlo localization using neural radiance fields," in *ICRA*. IEEE, 2023, pp. 4018–4025.

[3] M. Smith, J. L. Gabbard, G. Burnett, C. Hare, H. Singh, and L. Skrypchuk, "Determining the impact of augmented reality graphic spatial location and motion on driver behaviors," *Applied ergonomics*, vol. 96, p. 103510, 2021.

[4] Y. S. W. W. Arachchige, H. K. Lukosch, J. Everett, and S. Lukosch, "Representing remote locations with location-based augmented reality game design," *Entertainment Computing*, p. 100932, 2025.

[5] J. Miao, K. Jiang, T. Wen, Y. Wang, P. Jia, B. Wijaya, X. Zhao, Q. Cheng, Z. Xiao, J. Huang, *et al.*, "A survey on monocular re-localization: From the perspective of scene map representation," *IEEE Transactions on Intelligent Vehicles*, 2024.

[6] Y. Xia, L. Shi, Z. Ding, J. F. Henriques, and D. Cremers, "Text2loc: 3d point cloud localization from natural language," in *CVPR*. IEEE, 2024, pp. 14 958–14 967.

[7] S. Garg, T. Fischer, and M. Milford, "Where is your place, visual place recognition?" in *IJCAI*, 2021, pp. 4416–4425.

[8] K. A. Tsintotas, L. Bampis, and A. Gasteratos, "The revisiting problem in simultaneous localization and mapping: A survey on visual loop closure detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 19 929–19 953, 2022.

[9] G. Elghazaly, R. Frank, S. Harvey, and S. Safko, "High-definition maps: Comprehensive survey, challenges, and future perspectives," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 4, pp. 527–550, 2023.

[10] J. Zhu, S. Yan, L. Wang, S. Zhang, Y. Liu, and M. Zhang, "Lod-loc: Aerial visual localization using lod 3d map with neural wireframe alignment," in *NeurIPS*, 2024.

[11] OSMstats, "Osmstats - statistics of the free wiki world map." [Online]. Available: https://osmstats.neis-one.org/

[12] C. Ho, J. Zou, O. Alama, S. M. J. Kumar, B. Chiang, T. Gupta, C. Wang, N. Keetha, K. Sycara, and S. Scherer, "Map it anywhere (mia): Empowering bird's eye view mapping using large-scale public data," in *Advances in Neural Information Processing Systems*, 2024.

[13] P. Sarlin, D. DeTone, T. Yang, A. Avetisyan, J. Straub, T. Malisiewicz, S. R. Bulò, R. A. Newcombe, P. Kontschieder, and V. Balntas, "Orienternet: Visual localization in 2d public maps with neural matching," in *CVPR*, 2023, pp. 21 632–21 642.

[14] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P. Huang, S. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, vol. 2024, 2024.

[15] A. B. Camiletto, A. Bochicchio, A. Liniger, D. Dai, and A. Gawel, "U-BEV: height-aware bird's-eye-view segmentation and neural map-based relocalization," in *IROS*. IEEE, 2024, pp. 5597–5604.

[16] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: CNN architecture for weakly supervised place recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1437–1451, 2018.

[17] K. Garg, S. S. Puligilla, S. Kolathaya, K. M. Krishna, and S. Garg, "Revisit anything: Visual place recognition via image segment retrieval," in *ECCV (68)*, ser. Lecture Notes in Computer Science, vol. 15126, 2024, pp. 326–343.

[18] N. V. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. A. Scherer, K. M. Krishna, and S. Garg, "Anyloc: Towards universal visual place recognition," *IEEE Robotics Autom. Lett.*, vol. 9, no. 2, pp. 1286–1293, 2024.

[19] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet, "A survey on visual-based localization: On the benefit of heterogeneous data," *Pattern Recognit.*, vol. 74, pp. 90–109, 2018.

[20] T. Sattler, B. Leibe, and L. Kobbelt, "Improving image-based localization by active correspondence search," in *ECCV (1)*, vol. 7572, 2012, pp. 752–765.

[21] J. E. Vargas-Munoz, S. Srivastava, D. Tuia, and A. X. Falcao, "Openstreetmap: Challenges and opportunities in machine learning and remote sensing," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 1, pp. 184–199, 2020.

[22] G. Floros, B. van der Zander, and B. Leibe, "Openstreetslam: Global vehicle localization using openstreetmaps," in *ICRA*, 2013, pp. 1054–1059.

[23] C. Mandel and O. Birbach, "Localization in urban environments by matching sensor data to map information," in *European Conference on Mobile Robots*, 2013, pp. 50–55.

[24] B. Suger and W. Burgard, "Global outer-urban navigation with openstreetmap," in *ICRA*, 2017, pp. 1417–1422.

[25] Z. Xia, O. Booij, M. Manfredi, and J. F. Kooij, "Visual cross-view metric localization with dense uncertainty estimates," in *ECCV*, 2022, pp. 90–106.

[26] P. Sarlin, F. Debraine, M. Dymczyk, and R. Siegwart, "Leveraging deep visual descriptors for hierarchical efficient localization," in *CoRL*, vol. 87, 2018, pp. 456–465.

[27] P. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *CVPR*, 2019, pp. 12 716–12 725.

[28] H. Wu, Z. Zhang, S. Lin, X. Mu, Q. Zhao, M. Yang, and T. Qin, "Maplocnet: Coarse-to-fine feature registration for visual re-localization in navigation maps," in *IROS*, 2024, pp. 13 198–13 205.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[30] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision transformers need registers," in *ICLR*, 2024.

[31] M. Wysoczanska, O. Siméoni, M. Ramamonjisoa, A. Bursuc, T. Trzcinski, and P. Pérez, "Clip-dinoiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation," in *ECCV*, vol. 15119, 2024, pp. 320–337.

[32] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel, "Deep vit features as dense visual descriptors," *arXiv preprint arXiv:2112.05814*, 2021.

[33] Y. Shi, X. Yu, D. Campbell, and H. Li, "Where am I looking at? joint location and orientation estimation by cross-view matching," in *CVPR*, 2020, pp. 4063–4071.

[34] S. Zhu, T. Yang, and C. Chen, "Vigor: Cross-view image geo-localization beyond one-to-one retrieval," in *CVPR*, 2021, pp. 3640–3649.

[35] Y. Shi and H. Li, "Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image," in *CVPR*, 2022, pp. 16 989–16 999.

[36] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robotics Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[37] N. Samano, M. Zhou, and A. Calway, "You are here: Geolocation by embedding maps and images," in *ECCV*, vol. 12368, 2020, pp. 502–518.

[38] P. Sarlin, A. Unagar, M. Larsson, H. Germain, C. Toft, V. Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl, and T. Sattler, "Back to the feature: Learning robust camera localization from pixels to pose," in *CVPR*, 2021, pp. 3247–3257.