Geo-Localization in OpenStreetMap-Derived Semantic Map Space Using Structural Embeddings

Tomas Kanapickas [1], Ugnius Lukošius [1] and Andrius Kriščiūnas[1]

[1] *Kaunas University of Technology, Studentu St. 50, Kaunas, LT-51368, Lithuania*

Abstract

Unmanned aerial vehicle (UAV) localization is important in many applications, especially when Global Positioning System (GPS) signals are unreliable or unavailable. While many approaches rely on matching RGB imagery, this work investigates geo-localization using only structural semantic information derived from OpenStreetMap (OSM) data. The proposed method operates entirely within a rasterized semantic map containing roads and buildings, allowing isolation of structural information without visual effects such as lighting, vegetation, or seasonal changes. Localization is formulated as an embedding-based retrieval task. Semantic map tiles are encoded using a convolutional–transformer network. Two additional variants are evaluated: an object-wise representation and a graph-based approach using a GINE model. The models are tested on a database of 13,330 semantic tiles using cosine similarity. Results show that structural and topological information alone can support large-scale localization. Although strict Top-1 performance is limited at small spatial thresholds, Top-1% retrieval consistently identifies the correct geographic region within practical distances. The method is computationally efficient, requiring only milliseconds per query, making it suitable for real-time applications.

Keywords

UAV localization, geo-localization, semantic maps, OpenStreetMap, embedding retrieval, contrastive learning, transformer encoder, GIS-based localization

1. Introduction

Accurate localization is a fundamental problem across many Unmanned Aerial Vehicle applications [11], including mapping, infrastructure inspection, agriculture monitoring, and search-and-rescue operations. In many real-world scenarios, Global Positioning System (GPS) signals may be unreliable or unavailable due to environmental conditions, signal obstruction or interference [8, 10]. Such limitations have motivated the development of vision-based approaches for robust localization.

Currently, most existing research addresses this issue through cross-view geo-localization, where UAV images are matched against satellite imagery using convolutional neural networks (CNNs), NetVLAD-based methods, or transformer-based approaches [1, 4, 12]. While these methods have shown promising results, they often rely on large geo-tagged datasets and operate directly on raw RGB imagery [15]. Consequently, their performance is sensitive to environmental variations such as lighting conditions and seasonal changes [8].

An alternative solution is to move beyond traditional image-based localization and instead rely on semantic abstractions that provide a more stable and invariant representation of the scene. OpenStreetMap (OSM) [5] offers structured and annotated information about roads, buildings, and other geographic features, removing many visual factors that affect raw imagery, while preserving structural properties of the environment. In this work, localization is performed directly within semantic representation derived from OSM, allowing to evaluate structural information importance for embedding-based localization without introducing variability caused by segmentation.

*IVUS 2026: Information Society and University Studies 2026, May 12, Kaunas, Lithuania

tomas.kanapickas@ktu.edu (T. Kanapickas); ugnius.lukosius@ktu.lt (U. Lukosius); andrius.krisciunas@ktu.lt (A. Krisciunas)

0009-0006-3870-9287 (T. Kanapickas); 0000-0002-3691-0423 (A. Krisciunas)

Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-06

CEUR-WS.org/Vol-4317/paper14.pdf

Localization performance was formulated as an embedding-based image retrieval task over a large database of map tiles generated from OSM data. The evaluation methodology measured Rank-k recall under increasing spatial error tolerance, allowing analysis of how embedding similarity correlates with geographic proximity. This formulation was motivated by the fact that generated query tiles do not exactly correspond to the database tiles; therefore, spatial proximity of retrieved tiles provides a more meaningful evaluation metric than strict tile-level matching.

2. Related Work

Vision-based localization has been widely studied as an alternative for GPS-denied environments where signals are unreliable. Traditional approaches include matching onboard imagery against aerial or satellite orthophotos, or solving cross-view geo-location between ground, UAV, and satellite views, using handcrafted features or global image descriptors learned by convolutional or transformer-based models [3, 9]. Because these methods rely on RGB imagery, their performance is sensitive to appearance changes such as illumination, shadows, seasonal effects and vegetation.

Moreover, several works have explored the use of semantic representations to reduce the influence of visual factors. Semantic maps or segmentation masks preserve structural and geometric information of the environment, which can be sufficient for localization tasks [7, 14]. In contrast, this work performs localization directly on semantic tiles derived from OpenStreetMap data without relying on raw imagery or segmentation pipelines, thereby isolating structural information and removing variability introduced by segmentation errors.

The focus of this study is on evaluating how well structural and topological information alone can support largescale localization through embedding similarity.

3. Proposed Approach

3.1. Problem Formulation

In this work, the localization task is formulated as an embedding-based retrieval problem over a database of map tiles. Instead of relying on raw imagery, a semantic map derived from OpenStreetMap data is used to evaluate whether structural information alone is sufficient to capture geographic relationships in the embedding space.

The semantic map tiles are denoted as:

$$M = \{m_1, m_2, \dots, m_N\}, \tag{1}$$

where each tile m_i associates with a known position $p_i = (x_1, y_1)$ in a global map coordinate system. The coordinate system is defined in image space, where distance between points can be converted to metric units, scale of 0.25 meters per pixel. The goal is to retrieve the tile whose geographic position is closest to the query

The embedding function maps both the query q and the map tiles m_i into a d-dimensional feature space. Localization is achieved by computing the cosine similarity between embeddings and retrieving the most similar tiles

$$\widehat{m} = \arg\max_{m_i \in M} \frac{f(q) \cdot f(m_i)}{\|f(q)\| \|f(m_i)\|}, \tag{2}$$

This formulation allows analysis of how embedding similarity correlates with spatial proximity. The objective is to learn an embedding space in which semantically similar regions are mapped close, without relying on visual appearance.

3.2. OSM Data Processing and Tile Generation

The tiles are generated from an image constructed from OpenStreetMap vector data of a selected geographic region. OpenStreetMap provides structured semantic information in vector form; in this study, only roads and buildings were used.

The vector data was rasterized into an image, with spatial resolution of 0.25 x 0.25 meters per pixel. This resolution preserves sufficient detail while maintaining manageable image sizes for embedding extraction.

The rasterized map is divided into tiles of size 600 x 450 pixels, which corresponds to 150 x 112.5 meters. The tile dimensions were selected to approximately match the altitude of a UAV operating at typical altitude of 80-100 meters. Horizontal and vertical overlaps of 120 and 90 pixels are introduced between adjacent tiles to preserve spatial continuity, while limiting the total number of tiles.

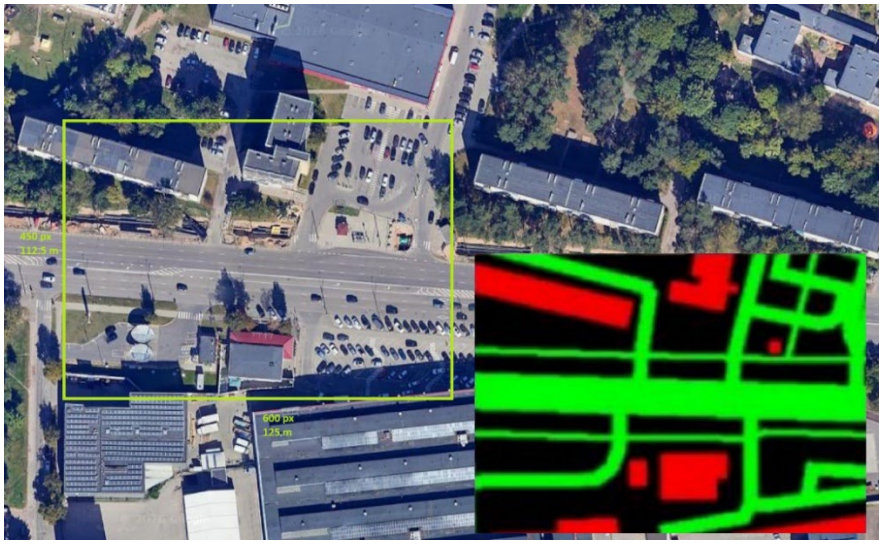


Figure 1: Segmentation from OpenStreetMap data in relation with orthophotos

Tiles containing no semantic structures (i.e., without roads or buildings) were excluded from the dataset.

Rasterized OSM provides a simplified semantic view that differs from orthophotos, as it excludes vegetation, shadows and seasonal effects, while preserving only structural information of the environment.

3.3. Embedding model

Each semantic map tile is mapped to a fixed-length embedding vector, using an encoder composed of a convolutional stem followed by a transformer encoder. To analyze how different forms of structural information contribute to semantic localization, three different strategies were evaluated. The primary approach processed the full semantic tile using CNN-transformer encoder to preserve global spatial layout. In addition, an object-wise representation of primary approach decomposes the scene into individual semantic objects to test whether object level features alone are sufficient. Finally, when the scene is formed into a relational graph using GINE architecture [6] to evaluate the topological connectivity between objects. These variants allow us to analyze how geometric layout, components level features, and relational topology influences localization performance. The architecture of the primary embedding model is described below.

Convolutional feature extraction. Each semantic tile is resized to 224 x 224 and passed through three convolutional blocks to generate patch-level descriptors. Each tile is represented as a three-channel RGB semantic mask, where buildings, roads, and background are encoded using distinct colors. The convolutional stem consists of three layers with channel dimensions $3 \rightarrow 64$, $64 \rightarrow 128$ and $128 \rightarrow 256$, using kernel sizes of 7x7, 3x3 and 3x3, respectively. Each layer uses stride 2 and is followed by batch normalization and ReLU activation.

Transformer aggregation. The patch descriptors are combined with positional embeddings and processed by a transformer encoder [13] consisting of four layers and multi-head self-attention. The global tile descriptor is obtained from the output token.

Embedding head. The resulting descriptor is projected into a d-dimensional embedding space and L2-normalized for cosine similarity-base retrieval during localization.

Training objective. The encoder is trained using an NT-Xent [2] contrastive loss on two augmented views of the same tile. The augmentation include random horizontal and vertical flips, encouraging the model to learn representations that are invariant to tile orientation while focusing on structural information.

This architecture was chosen because CNNs capture local geometric shapes, while transformer layers model spatial relationships between features across tile. As a result the formed embedding represents both geometry and overall layout, allowing to distinguish between map regions based on structural configuration rather than visual appearance.

4. Experimental Setup

Evaluation was performed on a database of 13330 semantic tiles that were generated from the OpenStreetMap representation.

To simulate localization queries, random coordinates were sampled within the rasterized map, and then a tile of size H x W pixel were extracted around each sampled location, while the sampled coordinate was the center of the tile. Consequently, as the query tiles are generated independently from the database grid, they did not match any pre-generated database tiles exactly.

Each query image is processed by the embedding model, and the resulting descriptor is compared against all database tile embeddings using cosine similarity. Database tiles then are ranked according to similarity.

Localization performance is evaluated using Top-1, Top-5, Top-10 and Top 1 % (133 top tiles) candidate retrieval. A query was considered successful, if at least one retrieved tile lies within the predefined spatial distance thresholds from the query location. Performance is reported under defined distances (5, 10, 25, 50, 100 meters and larger distances).

5. Results

5.1. Comparative Localization Performance

Table 1
Localization performance of the CNN-Transformer model. Success under spatial distance thresholds

Threshold (m)	Top1	%	Top5	%	Top10	%	Top1%	%
25	41	8.2	53	10.6	66	13.2	84	16.8
50	68	13.6	114	22.8	146	29.2	226	45.2
100	86	17.2	154	30.8	207	41.4	350	70.0

Results (Table 1) represent localization accuracy under spatial distances. At very close distances (5-25 meters), accuracies remained low, indicating that retrieval of the exact tile is challenging. However, while the distance tolerance increases, Top-5 and Top-10 retrieval accuracy improves steadily, indicating that embeddings consistently retrieve tiles that are geographically close to query location. The most significant observation is the behavior of Top-1% recall. Where at the 100 meter distance, the correct region is selected within the top 133 retrieved tiles in 70 % of cases. These results demonstrate that embedding space strongly preserves geographic structure, while exact tile matching fails. This behavior is important as, for practical localization systems, the correct region selection is important, when later in the localization stage the algorithm could be improved to try select the more optimal solution by providing a another image to try increase the accuracy of the query.

Table 2
Localization performance of the Object-based CNN-Transformer model. Success under spatial distance thresholds.

Threshold (m)	Top1	%	Top5	%	Top10	%	Top1%	%
25	7	2.6	12	2.4	18	3.6	40	8.0
50	14	2.8	26	5.2	37	7.4	96	19.0
100	25	5.0	52	10.4	74	14.8	183	36.6

Table 3
Localization performance of the GINE Graph model. Success under spatial distance thresholds.

Threshold (m)	Top1	%	Top5	%	Top10	%	Top1%	%
25	13	2.6	19	3.8	24	4.8	44	8.8
50	17	3.4	38	7.6	53	10.6	109	21.8
100	17	3.4	42	8.2	65	13.0	169	33.8

The full-tile approach consistently outperforms the other methods across all spatial thresholds. Although the object-wise method also uses a transformer, processing roads and buildings independently removes the global spatial layout of the scene. As a result, the transformer receives fragmented object features without their original geometric arrangement, reducing localization accuracy. Similarly, the graph-based representation captures relational structure but lacks dense spatial detail, leading to weaker performance compared to the full semantic tile representation.

5.2. Analysis of the CNN-Transformer Model

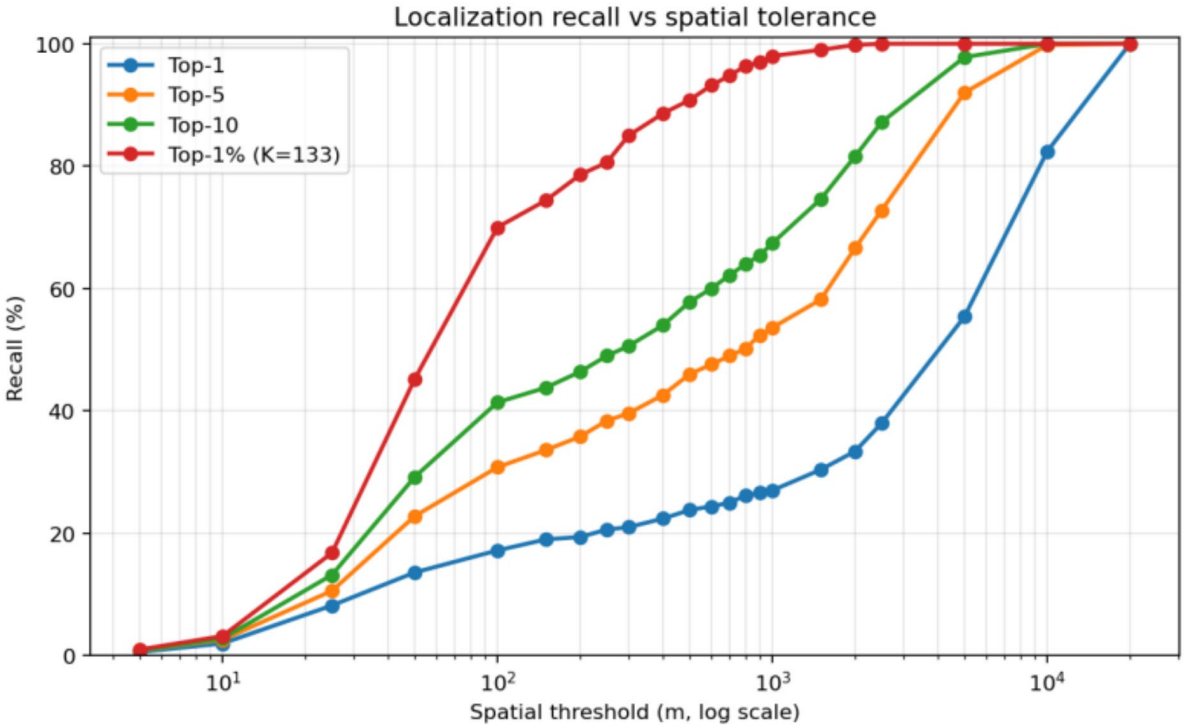


Figure 2: Top-1, Top-5, Top-10 and Top 1%, recall change while spatial threshold increases, CNN-Transformer model.

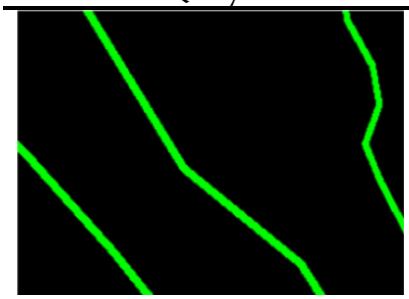
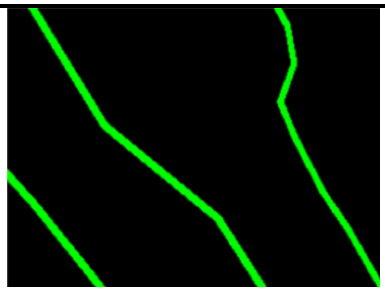
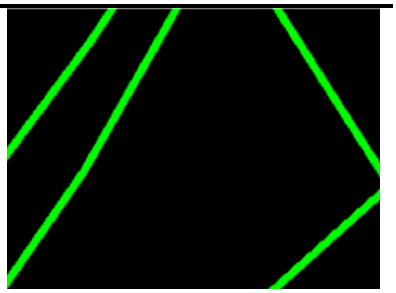
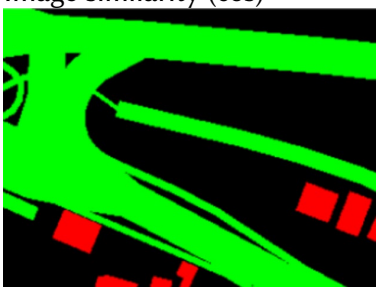
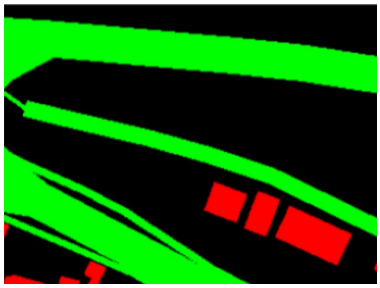
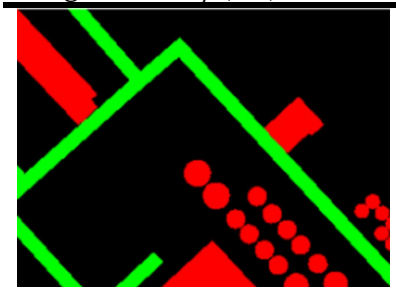
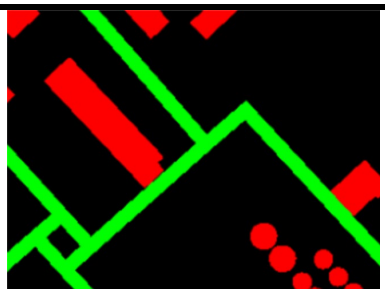
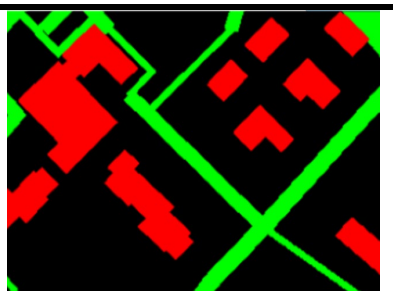
The primary model results (Fig. 2) show that recall increases consistently as spatial tolerance grows for all Top-N settings. Performance remains low below 25 m, indicating that exact tile retrieval is difficult due to discretization and arbitrary query placement. A rapid improvement is observed between 50 m and 100 m, where the number of correctly retrieved candidates increases substantially. Beyond this range, recall continues to grow more gradually, with Top-1% approaching near-complete retrieval at kilometer-scale tolerances, while Top-5 and Top-10 follow a similar but slower trend. At very large thresholds (≈10 km), all Top-N configurations eventually achieve nearly 100% recall, as expected when spatial constraints become loose. Overall, the 100 m threshold represents a practically meaningful operating point, indicating that the model reliably retrieves geographically overlapping regions rather than exact tiles.

Table 4
Localization performance of the GINE Graph model. Success under spatial distance thresholds.

Top-N retrieved	Query embedding time (s)	Similarity search time (s)	Total processing time per query (s)
10	0.008348	0.000497	0.039873
50	0.012301	0.000506	0.164380
133 (1%)	0.010405	0.000515	0.445829

Furthermore, the proposed method demonstrates high computational efficiency. In all experiments, the average embedding extraction time was approximately 10 ms per query, while similarity comparison against the full database of 13,330 tiles required less than 1 ms. The total processing time per query was approximately 40 ms, 164 ms, and 446 ms for retrieval of the Top-10, Top-50, and Top-1% (133) candidates, respectively. The embedding and similarity computation together account for only about 10–20% of the total processing time, while the remaining time is mainly consumed by additional SSIM, NCC, and cosine-based image comparison metrics applied to the retrieved candidates.

Table 5
Example retrieval results illustrating structural similarity between query tiles and retrieved candidates.

Query	Rank 1	Rank 2
		
Similarity (cos)	0.9975	0.9830
Distance	116 (px) 29 (m)	4744 (px) 1186 (m)
SSIM	0.6790	0.6990
Image similarity (cos)	0.0000	0.0799
		
Similarity (cos)	0.8657	0.8405
Distance	149 (px) 37 (m)	13522 (px) 3380 (m)
SSIM	0.2684	0.2739
Image similarity (cos)	0.4527	0.4758
		
Similarity (cos)	0.9434	0.8898
Distance	156 (px) 39 (m)	38118 (px) 9529 (m)

Query	Rank 1	Rank 2
SSIM	0.3693	0.3088
Image similarity (cos)	0.1660	0.2525
		
Similarity (cos)	0.9521	0.9476
Distance	1609 (px) 402 (m)	2060 (px) 515 (m)
SSIM	0.2159	0.2147
Image similarity (cos)	0.3795	0.3602

Qualitative retrieval examples demonstrate that the model captures structural layout, returning tiles with similar road and building patterns to the query. In some cases, structurally similar regions from distant locations are also retrieved, indicating that the embedding represents topology rather than absolute geographic position.

6. Conclusions

The proposed localization approach operates entirely within a semantic space generated from OpenStreetMap (OSM) vector data. The rasterized tiles contain only roads and buildings, and a convolutional–transformer encoder is trained to produce embeddings for similarity-based tile retrieval.

The results show that structural and topological information alone is sufficient to support large-scale localization. Although strict Top-1 accuracy is limited at small spatial thresholds, Top-1% retrieval consistently identifies the correct geographic region within practical distances. These findings indicate that the learned embedding space preserves spatial structure while allowing further improvements in the localization strategy. In addition, the method is computationally efficient, requiring only milliseconds per query, which makes it suitable for real-time applications.

Overall, the results suggest that semantic GIS data can serve as an alternative to raw imagery-based localization approaches. Future work will explore multi-query candidate refinement to incorporate more scene information and reduce the search region. Further integration with raw imagery and three-dimensional semantic representations may improve robustness and localization precision.

7. Funding

This research was funded by the Research Council of Lithuania, Agreement Number S- 662 MIP-23-88.

8. Declaration on Generative AI

The authors have not employed any Generative AI tools.

9. References

[1] Chen, N. et al.: OBTPN: A Vision-Based Network for UAV Geo-Localization in Multi-Altitude Environments. Drones. 9, 1, (2025). https://doi.org/10.3390/drones9010033.
[2] Chen, T. et al.: A Simple Framework for Contrastive Learning of Visual Representations. (2020).
[3] Cui, Z. et al.: A Novel Geo-Localization Method for UAV and Satellite Images Using Cross-View Consistent Attention. Remote Sens. (Basel). 15, 19, (2023). https://doi.org/10.3390/rs15194667.
[4] Duan, R. et al.: A Scene Graph Encoding and Matching Network for UAV Visual Localization. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 17, 9890–9902 (2024). https://doi.org/10.1109/JSTARS.2024.3396168.
[5] Haklay, M., Weber, P.: OpenStreetMap: User-Generated Street Maps. IEEE Pervasive Comput. 7, 4, 12–18 (2008). https://doi.org/10.1109/MPRV.2008.80.
[6] Hu, W. et al.: Strategies for Pre-training Graph Neural Networks. (2020).
[7] Hu, Y. et al.: Combining OpenStreetMap with Satellite Imagery to Enhance Cross-View Geo-Localization. Sensors. 25, 1, (2025). https://doi.org/10.3390/s25010044.
[8] Kinnari, J. et al.: Season-Invariant GNSS-Denied Visual Localization for UAVs. IEEE Robot. Autom. Lett. 7, 4, 10232–10239 (2022). https://doi.org/10.1109/LRA.2022.3191038.
[9] Lin, H.Y., Chen, L.Y.: GNSS-denied UAV localization with satellite and aerial image matching. Results in Engineering. 28, (2025). https://doi.org/10.1016/j.rineng.2025.108132.
[10] Liu, H. et al.: A survey of sensors based autonomous unmanned aerial vehicle (UAV) localization techniques. Complex and Intelligent Systems. 11, 8, (2025). https://doi.org/10.1007/s40747-025-01961-2.
[11] Meer, I.A. et al.: Cellular localizability of unmanned aerial vehicles. Vehicular Communications. 44, (2023). https://doi.org/10.1016/j.vehcom.2023.100677.
[12] Nassar, A. et al.: A Deep CNN-Based Framework For Enhanced Aerial Imagery Registration with Applications to UAV Geolocalization.
[13] Vaswani, A. et al.: Attention Is All You Need. (2017).
[14] Wang, Z. et al.: VecMapLocNet: Vision-based UAV localization using vector maps in GNSS-denied environments. ISPRS Journal of Photogrammetry and Remote Sensing. 225, 362–381 (2025). https://doi.org/10.1016/j.isprsjprs.2025.04.009.
[15] Ye, Q. et al.: A coarse-to-fine visual geo-localization method for GNSS-denied UAV with oblique-view imagery. ISPRS Journal of Photogrammetry and Remote Sensing. 212, 306–322 (2024). https://doi.org/10.1016/j.isprsjprs.2024.05.006.