Autonomous Identification of Distinctive Landmarks from Earth Surface Images

Zakhar Ostrovskyi[1], Oleksander Barmak[1] and Iurii Krak[2,3]

[1] Khmelnytskyi National University, 11, Institutes str., Khmelnytskyi, 29016, Ukraine
[2] Taras Shevchenko National University of Kyiv, 64/13, Volodymyrska str., Kyiv, 01601, Ukraine
[3] Glushkov Cybernetics Institute, 40, Glushkov Ave., Kyiv, 03187, Ukraine

Abstract

This article addresses the problem of identifying unique objects in aerial images of urban areas on the Earth's surface, which can serve as stable landmarks for UAV navigation without GPS signals. The main contribution lies in proposing an approach to transforming the image into an object-oriented vector representation (embedding) that retains structural information about those objects. The proposed approach automatically identifies the most distinctive objects, which can serve as navigation landmarks. The study focuses on urban and suburban landscapes, where buildings are chosen as landmarks and YOLOv11 is used as the deep learning model. By employing dimensionality reduction methods, in particular PCA and t-SNE, it is demonstrated that in the proposed embedding space, buildings with atypical structural or visual characteristics differ significantly from other buildings and are easily classified as outliers, making them natural landmarks for navigation. Experimental results confirm the effectiveness and potential of the proposed approach for ensuring stable UAV navigation in scenarios where GPS may be inaccessible—the accuracy of identifying buildings designated as landmarks is twice that of ordinary buildings (Recall@1 = 0.51 vs. 0.28).

Keywords

Instance embeddings, Landmark selection, Convolutional Neural Networks, UAV navigation, Satellite images, GPS-denied environments.

1. Introduction

Unmanned Aerial Vehicles (UAVs) increasingly operate in environments where GPS signals are unreliable or absent [1]. Under such conditions, visual landmarks identified from onboard camera images become the sole method for determining UAV location [2]. For accurate localisation, landmarks must be distinctive and visually recognisable under various conditions of illumination, altitude, and imaging type (e.g., UAV camera versus satellite imagery). Thus, using a set of landmarks for a given area, a route can be planned to a specified point, allowing UAV navigation without GPS signals, relying solely on environmental image analysis.

Depending on terrain characteristics, various types of objects can serve as landmarks. Given their critical practical relevance for tasks like search-and-rescue operations, deliveries, and path planning for long-range strike UAVs through densely populated urban areas, this study focuses on urban and suburban environments. Buildings, frequently prominent in these areas, have thus been chosen as potential landmarks in this research.

In recent years, deep neural networks have become widely used to obtain vector representations of features, such as embeddings. In computer vision tasks, embeddings of images or their fragments can be extracted from the hidden layers of pre-trained neural networks (such as convolutional networks or transformers), transforming visual data into compact, information-rich vectors while preserving meaningful similarity between inputs [3], [4]. Essentially, a neural network "encodes" an image into a point in latent space, placing images with similar content close together and enabling comparison using distance functions. Moreover, networks can be fine-tuned to produce embeddings with enhanced properties [5], such as invariance to common disturbances (changes in lighting, angles, etc.), thus improving their robustness in dynamic environments. Instead of raw pixel processing, UAV navigation systems can reliably recognise relevant objects using embeddings and distance metrics.

[1]CMIS-2025: Eighth International Workshop on Computer Modeling and Intelligent Systems, May 5, 2025, Zaporizhzhia, Ukraine

ostrovskyiz@khmnu.edu.ua (Z. Ostrovkyi); barmako@khmnu.edu.ua (O. Barmak); iurii.krak@knu.ua (I. Krak)
0009-0003-4644-3587 (Z. Ostrovkyi); 0000-0003-0739-9678 (O. Barmak); 0000-0002-8043-0785 (I. Krak)

Therefore, this research contributes to developing an approach to generate semantically rich vector representations of objects (buildings) based on their representations in the hidden layers of a convolutional neural network when processing satellite and UAV images. It doesn't require any additional training and thus can be directly applied to any segmentation CNN and other types of objects.

The proposed approach is practically significant, as it addresses the automatic identification of stable landmarks among numerous similar objects. For example, many urban buildings share architectural features, materials, or colour schemes, reducing their distinctiveness for reliable visual identification.

The structure of this paper is as follows: the *Literature Review* section surveys previous research on UAV visual localisation, including marker-based approaches, semantic dictionaries, and feature aggregation methods. The *Materials and Methods* section introduces the proposed model for extracting embeddings and automatically identifying landmark buildings. The *Results and Discussion* section presents experimental outcomes using the VPAIR dataset, their analysis, and potential improvements to the approach.

2. Literature review

Below, related research directions are reviewed.

One closely related approach is marker-based localisation, where artificial markers ensure uniqueness. For instance, YoloTag [6] employs a YOLO-based detector for fiducial markers, enabling position estimation through geometric algorithms (e.g., EPnP [7]). Although effective in indoor or restricted outdoor environments, it depends on physical marker placement, limiting scalability. Marker optimisation methods [8] face similar scalability challenges.

Semantic mapping and object dictionaries maintain recognised object annotations with geometric or class characteristics. For example, [9] combines detection with depth data (RGB-D cameras) to create semantic maps (doors, fire extinguishers, etc.). While conceptually similar, these solutions typically do not produce vector embeddings that differentiate objects within the same class, limiting the identification of distinctive landmarks. Additionally, dictionary-based approaches usually cover small-to-medium indoor areas, where objects appear repeatedly from various viewpoints within one route.

Local CNN-descriptor aggregation into global vectors has been investigated in object retrieval contexts (e.g., SPoC, CroW, R-MAC, NetVLAD [10], [11], [12], [13]), usually tested on datasets for ground-level place recognition. Despite conceptual similarities, these methods rarely perform precise object segmentation, particularly from aerial imagery. Furthermore, these methods typically produce a global vector representation for entire images, not considering individual object vector representations.

Domain adaptation methods such as CLDA-YOLO [14] address environmental variations (weather, lighting, etc.) in object detection tasks. These methods could enhance embedding robustness within the developed algorithm.

Comprehensive surveys of UAV navigation under GPS-denied conditions [15] emphasise the importance of tracking visual reference points. Yet, systems like SLAM primarily track key surface points without treating objects holistically as unique landmarks.

Literature analysis indicates limited attention to landmark-based UAV localisation methods. Most existing approaches generate descriptor vectors for entire UAV images, matching them against annotated databases. However, these descriptors may be sensitive to changes in imaging conditions and background noise.

Thus, this research aims to improve UAV localisation accuracy using stable landmarks derived from the hidden layers of a convolutional neural network when processing satellite images. These landmarks, based exclusively on unique objects, offer robustness to image noise. To achieve this goal, the following research tasks were formulated.

1. Develop a method for obtaining embeddings (vector representations) of buildings from convolutional neural network hidden layers, capable of preserving their visual and structural characteristics.
2. Create an automatic method for selecting landmark buildings based on embedding space analysis (e.g., via outlier detection).

3. Experimentally validate the proposed approach for accurately identifying buildings designated as landmarks (based on prior analysis of satellite images) on UAV images.

3. Materials and methods

3.1. Process model

For the problem under consideration, the input data consists of a set of satellite images, denoted by $I=\{I_1, I_2, \ldots, I_N\}$, covering a specific geographic area, and a fixed set of landmark object types, defined as $LandmarkTypes=\{Type_1, Type_2, \ldots, Type_t\}$ potentially comprising multiple object categories. Each satellite image I_n may contain several objects from the set $LandmarkTypes$, represented as $\{O_n^1, O_n^2, \ldots, O_n^{K_n}\}$. O – designates the set of all objects from all images I.

The proposed approach employs convolutional neural networks (CNNS) specialised for image segmentation tasks. These CNNs are trained to recognise object types from the $LandmarkTypes$ set. The output of this network for each recognized object O_n^k is the corresponding segmentation mask M_n^k.

For each object O_n^k, it is necessary to apply a mapping function f to a d-dimensional real-valued embedding vector space R^d:

$$f\left(I_n, M_n^k\right)=e_n^k, e_n^k \in R^d \tag{1}$$

The obtained vector representation of an object should possess the following properties:

- *uniqueness* – e_n^k must emphasise distinctive visual features of object O_n^k, enabling reliable differentiation from others. Let us denote by S the set of visually similar objects to O_n^k, and by D the set of visually distinct objects. The following inequality must hold:

$$\forall e^+ \in S, \forall e^- \in D: d\left(e_n^k, e^+\right) < d\left(e_n^k, e^-\right) \tag{2}$$

where $d(\cdot, \cdot)$ is a chosen distance metric (e.g., Euclidean distance).

- *robustness* – under minor transformations of the object, such as changes in viewpoint, lighting conditions, or partial occlusions, the vector representation e_n^k remains practically unchanged.

Suppose T is a transformation modelling changes in imaging conditions, and $e\left(T\left(O_n^k\right)\right)$ is the object embedding after applying transformation T. Stability is ensured if:

$$\left\| e\left(T\left(O_n^k\right)\right) - e_n^k \right\| < \epsilon, \tag{3}$$

where ϵ is a small constant defining the permissible deviation level, and $\|.\|$ denotes a vector norm (for instance, Euclidean).

After obtaining all embeddings e_n^k within the dataset, the task arises to automatically identify instances that stand out significantly in the embedding vector space. Formally, let $\{e_1, e_2, \ldots, e_M\}$ denote all object embeddings. A criterion based on outlier detection algorithms is introduced, distinguishing objects with distinctive features from those forming dense clusters. Objects thus identified are labelled as landmarks. The set L of such landmark objects represents the final output of the landmark detection task.

These landmark objects, identified through the described method, form the basis for UAV route planning. In scenarios lacking GPS signals, a UAV can determine its location by recognising selected stable and unique landmarks on the terrain.

3.2. Hypothesis

An object database along the flight trajectory is prepared before the mission to facilitate UAV navigation based on visual features. These objects are extracted from satellite images, each associated with geolocation markers. During flight, the UAV identifies objects from onboard camera images and searches for matching objects in this pre-formed database. Upon finding a match, the UAV uses the object's geolocation marker to determine its current position.

Generally, it is reasonable to assume that if an object identified in a satellite image acts as a landmark, standing out from surrounding objects due to unique features, this distinction will persist in images captured by UAV cameras. Based on these assumptions, the core hypothesis can be articulated as follows: if a mapping into vector space preserving semantic and structural features is applied to the set of objects in an image, landmark objects will distinctly differ from other objects in the embedding space, regardless of whether they originate from UAV camera images or satellite imagery. Consequently, when receiving images from onboard cameras, the UAV can significantly more accurately identify precisely those objects recognised as landmarks by the approach proposed in this study.

Thus, the essence of the proposed method lies in the specific utilisation of convolutional neural networks to create an embedding vector space that preserves semantic and structural characteristics. It is well-known that convolutional neural networks learn hierarchical feature representations, with initial layers capturing local textures and edges, and deeper layers encoding higher-level forms or semantic features. An embedding with hierarchical features of the specific object is effectively obtained by constructing a vector from activations corresponding to a particular object from various hidden layers.

3.3. Method steps

The main steps of the proposed approach are illustrated in Fig. 1.

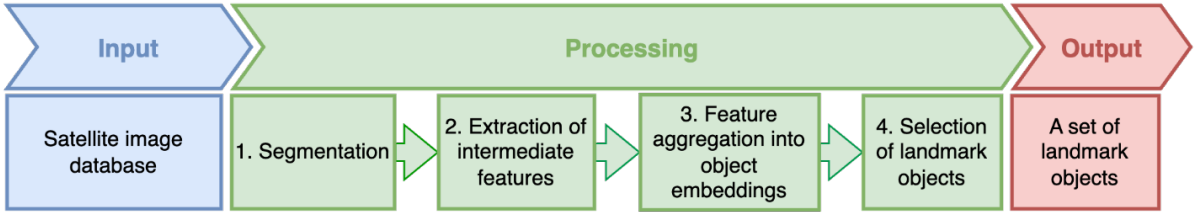


Figure 1: The main steps of the proposed approach

The input information of the approach (Fig. 1) consists of a set of satellite images of a specific territory I and a fixed set of object types that potentially serve as landmarks, referred to as $LandmarkTypes$.

In the first step of the proposed approach, the set O of potential landmark objects is formed by segmenting images from the satellite imagery database. A CNN-based segmentation model, trained to recognise objects from the $LandmarkTypes$ set, is applied to each image I_n, resulting in $\{M_n^k\}$ – a set of masks and corresponding confidence scores. Only objects with a confidence coefficient above a threshold θ are selected for reliability.

In step 2, intermediate features of objects are extracted from the hidden layers of the convolutional neural network. To describe this process, denote the feature map of a CNN at layer $\boldsymbol{I}$ as $\boldsymbol{F_l} \in R^{C_l \times H_l \times W_l}$, where C_l is the number of channels at layer $\boldsymbol{I}$, and $H_l ma W_l$ represent the height and width, in pixels, of the feature maps at layer $\boldsymbol{I}$, respectively. Let $l_1, l_2, \ldots, l_L$ be parameters corresponding to the backbone CNN layers used for embedding formation. During segmentation in step 1, feature maps F_l are extracted from the backbone CNN layers $l_1, l_2, \ldots, l_L$. Each layer l is associated with a stride ρ_l, defining the reduction in spatial resolution of feature maps compared to the original image. Accordingly, each mask M_n^k is resized to dimensions $\widetilde{M}_n^{kl}$ based on the corresponding stride ρ_l of the feature map F_l. This alignment ensures pixels in the mask correspond precisely to positions on the feature map, thus isolating only the area corresponding to the detected object O_n^{kl}.

Step 3 involves aggregating intermediate features to form the final object embeddings. Since each object may vary in size, occupying different-sized regions on the activation maps, an aggregation function must be employed to obtain a fixed-dimensional vector. Generally, the aggregation function can be a parameterised function trainable via backpropagation, such as a graph neural network [16].

For each layer l and object k, the aggregation is computed as:

$$z_c^l = \mathrm{agg}\left(\left\{F_l[c,u,v] \middle| (u,v) \in \widetilde{M}_n^{kl}\right\}\right) \tag{4}$$

where the aggregation function can be, for example, max pooling or average pooling.

Max pooling:

$$z_c^l = \max_{(u,v)\in\widetilde{M}_n^{kl}} F_l[c,u,v] \tag{5}$$

Average pooling:

$$z_c^l = \frac{1}{\left|\widetilde{M}_n^{kl}\right|} \sum_{(u,v)\in\widetilde{M}_n^{kl}} F_l[c,u,v] \tag{6}$$

Values z_c^l, computed for all C_l channels across selected convolutional layers $l_1, l_2, \ldots, l_L$, are concatenated to form the final embedding $e_n^k \in R^d$:

$$e_n^k = \left[z_c^l \middle| c \in 1..C_l, l \in l_1, l_2, \ldots, l_L \right] \tag{7}$$

Thus, the embedding dimension d is determined by the total number of channels in convolutional layers $l_1, l_2, \ldots, l_L$, where each dimension effectively represents the presence of patterns detected by corresponding convolutional filters:

$$d = \sum_{l \in l_1, l_2, \ldots, l_L} C_l \tag{8}$$

Step 4 identifies landmark objects. Initially, all embeddings $\{e_1^1, e_1^2, \ldots\}$ are combined into the set E. Since each embedding dimension corresponds to a specific convolutional filter trained to recognize particular image structures, embeddings implicitly represent visual features of objects. Based on the initial assumption, objects with atypical visual characteristics yield embeddings with atypical values. Consequently, the final stage entails differentiating "typical" points in the embedding space from those with rare features. Theoretically, this problem class corresponds to outlier detection methods aimed at identifying objects statistically deviating from the majority.

Generally, the outlier detection task can be formulated as follows: let $E = \{e_1, e_2, \ldots, e_M\} \subset R^d$, where each vector e_i is an object embedding. Suppose that for most points, the feature distribution approximates a “typical” (“normal”) subset E_{norm}, while a few points $e_j \in E_{\text{out}}$ significantly deviate from this distribution. Formally, an evaluation function is assumed:

$$s: R^d \to R, \tag{9}$$

which returns the deviation from the typical distribution for each e_i. If $s(e_i)$ exceeds a threshold s_{thr}, e_i is considered an outlier (anomaly). In our context, objects with such embeddings possess distinctive visual characteristics and can serve as stable landmarks. Hence, applying an outlier detection algorithm to set E forms the final landmark object set $L = \{e_i \vee s(e_i) > s_{thr}\}$.

The landmark object set L is the output of the proposed approach.

3.4. Evaluation metrics

The UAV localisation problem considered in this study is classically framed as a retrieval task. Consequently, literature conventionally evaluates UAV localisation methods using the Recall@N metric [17], [18]. This metric considers a retrieval result as a true-positive for a given query if the corresponding image from the database appears among the top N retrieved images:

$$\text{Recall@N} = \frac{M_Q}{N_Q}, \tag{10}$$

where N_Q is the total number of query images, and M_Q is the number of queries with at least one correct match within the top-N results.

This metric is popular within computer vision communities and suits applications employing post-processing to eliminate false-positive matches.

4. Results and discussion

4.1. Dataset

The VPAIR dataset [19] was selected for conducting experiments – a dataset designed explicitly for evaluating visual place recognition tasks and UAV localisation based on images from onboard cameras. Data collection occurred on October 13, 2020, during a flight of a light aircraft at altitudes ranging from 300 to 400 meters above ground, covering an area between Bonn, Germany, and the Eifel Mountain range, with a total route length of 107 km. The dataset includes imagery captured perpendicular to the Earth's surface and high-precision pose/orientation data obtained using GNSS/INS systems. The VPAIR dataset contains 2,788 aerial photographs paired with corresponding satellite images and *does not* provide any annotations about the objects in the images. The satellite images were gathered from Geobasis NRW, a state-funded geodata repository under a permissive open data license. It provides comprehensive coverage of the entire state of Nordrhein-Westfalen, Germany. During image capture, the aircraft maintained a speed of 150 km/h and a frame rate of 1 Hz, resulting in approximately 41.7 meters between consecutive image centres.

Figure 2: Examples of images from the VPAIR dataset. Left – aerial images captured from aircraft; right – corresponding satellite images

4.2. Experiment description

The YOLOv11 segmentation convolutional neural network [20], pre-trained for building segmentation in satellite images, was utilised in the experiments. It is important to emphasise that the proposed method *uses the pre-trained CNN* that segments the objects of interest and *requires no additional training* on the target dataset.

For outlier detection – specifically to identify landmark objects – Isolation Forest [21], a tree-based algorithm, was chosen. The selection of this algorithm was motivated by three main reasons: tree-based algorithms are robust against variations in feature value ranges and thus do not require normalisation; they operate rapidly; and they only need two primary parameters that are easily adjustable (the proportion of objects considered as outliers and the number of trees). This straightforward algorithm facilitated focusing on hypothesis verification and proved sufficient to confirm it. Subsequent experiments present results obtained with Isolation Forest configured with 500 trees and 1% outliers.

To validate the proposed hypothesis, the following sub-hypotheses must be tested:

1) The proposed embedding generation approach encodes structural and semantic information about objects.
2) The accuracy of landmark building retrieval from UAV images is significantly higher than that of typical (non-landmark) buildings.

It should be noted that the VPAIR dataset contains no specific annotations for buildings; thus, the set of buildings used in this study was obtained using the YOLOv11 segmentation model.

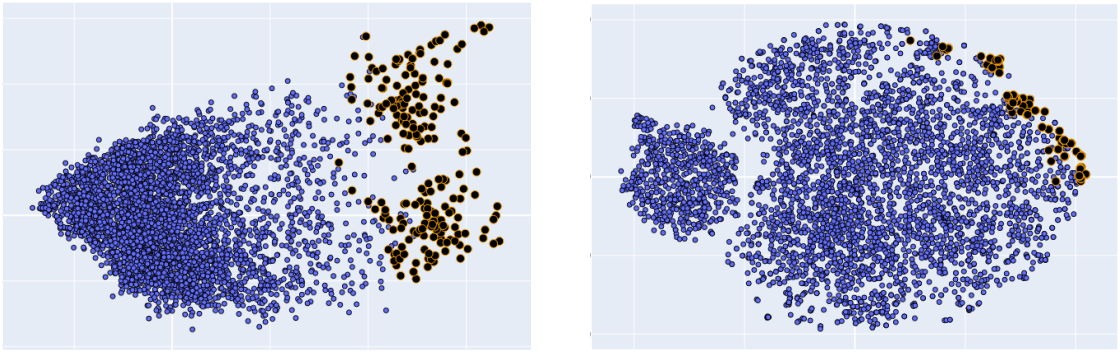
The absence of building ground-truth annotations in the dataset makes it impossible to quantify misclassifications, false positives, and false negatives in the object detection process on the VPAIR dataset.

However, for the *pre-trained* YOLOv11 used in the experiments, the following metrics are reported by its developers: 18,794 true positives, 8,462 false positives, and 5,628 false negatives. At the same time, true-negative background pixels are undefined for segmentation. Across seven random splits the model attained mAP 0.754, precision 0.771, recall 0.680 and F1 0.722. The reported values establish a realistic error bound when the model is applied to the VPAIR dataset, and the manual inspection of the predictions confirms its high performance and generalisation to this dataset.

To verify the first hypothesis, visualisation of the building embeddings—obtained from segmented satellite images—was conducted using two dimensionality reduction methods: PCA for analysing linear dependencies and t-SNE for non-linear dependencies. Researchers then visually inspected the proposed method and provided qualitative assessments.

Validation of the second hypothesis required manual data labelling to create a benchmark set, as the VPAIR dataset contains no building annotations. Given corresponding satellite and UAV images and buildings previously segmented by YOLOv11, matching identical buildings across UAV and satellite images was necessary. Considering the time-intensive nature of manual labelling, a random, non-repetitive sample of 100 landmark buildings and 100 typical buildings was selected for annotation. For the embeddings of each of the selected 200 UAV buildings, the five nearest embeddings from satellite images were identified using the L2 norm. The metrics Recall@1 and Recall@5 were calculated separately for landmark and typical buildings.

4.3. Analysis of the obtained embedding space



a) b)

Figure 3: Visualisation plots of building embeddings from satellite images using dimensionality reduction methods: left a) – PCA; right b) – t-SNE. Black points represent landmark buildings, and blue points represent typical buildings.

Visualisation results of building embeddings obtained via dimensionality reduction methods (Fig. 3) demonstrate that the embedding space is structured.

The PCA plot shows that most buildings concentrate on the left side, with the remaining points forming an elongated, sparse tail. It is logical to hypothesise that the dense concentration corresponds to numerous typical buildings, while the progressively extending tail represents buildings with increasing visual uniqueness. Visual inspection of points in these areas (Fig. 4) confirms this assumption (Fig.5 and Fig.6). Thus, the selection of buildings at the tail end of this distribution by the outlier detection algorithm as landmarks aligns with expectations, as these points correspond to the most distinctive structures.

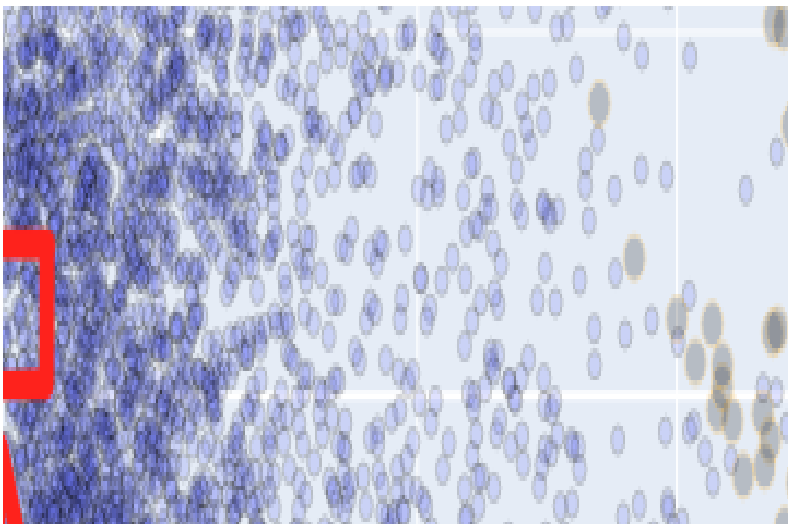


Figure 4: Visualisation plot of building embeddings from satellite images using PCA dimensionality reduction. Segment a) highlights a region with a high density of points corresponding to typical small buildings (Fig. 5), while segment b) represents a cluster of landmark buildings (Fig. 6)

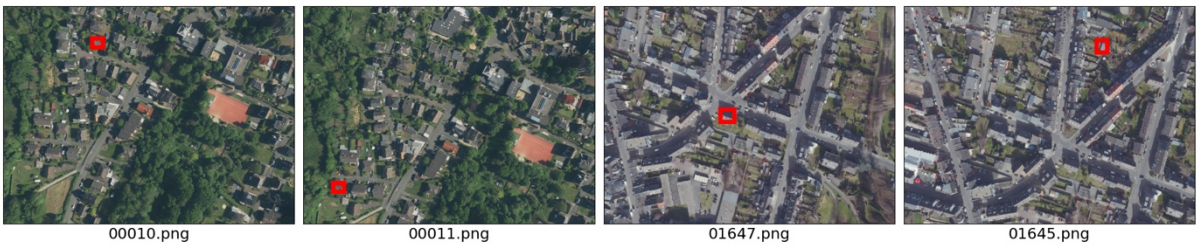


Figure 5: Examples of typical small buildings, outlined with red rectangles, corresponding to selected points in Fig. 4a.

Figure 6: Examples of landmark buildings, outlined with red rectangles, corresponding to points from the highlighted cluster in Fig. 4b.

The t-SNE visualisation, which reveals non-linear relationships, displays multiple small clusters grouping visually similar buildings or identical buildings from adjacent frames. The fact that landmark buildings cluster at the edges of the point cloud, rather than being dispersed throughout, indicates good embedding space structure. A particularly notable cluster emerges distinctly in the left region of the t-SNE plot. Visual inspection revealed that this cluster corresponds to small buildings with typical structures positioned at image boundaries (so buildings partially extend beyond the frame edge). Examples of these buildings and their corresponding points in the t-SNE visualisation are illustrated in Fig.7-9.

Thus, the proposed method effectively distinguishes landmark buildings from typical ones within the embedding space. Selected landmark buildings exhibit unique characteristics, often large or irregular shapes. Visually similar buildings in size, colour, and form have close embeddings. The neighbourhoods around embeddings situated in regions of greater uniqueness mostly contain embeddings of the same buildings from adjacent frames, indicating stability of the vector representation across different viewpoints. However, as uniqueness decreases, the neighbourhoods increasingly include buildings that, although visually similar, originate from spatially distant locations.

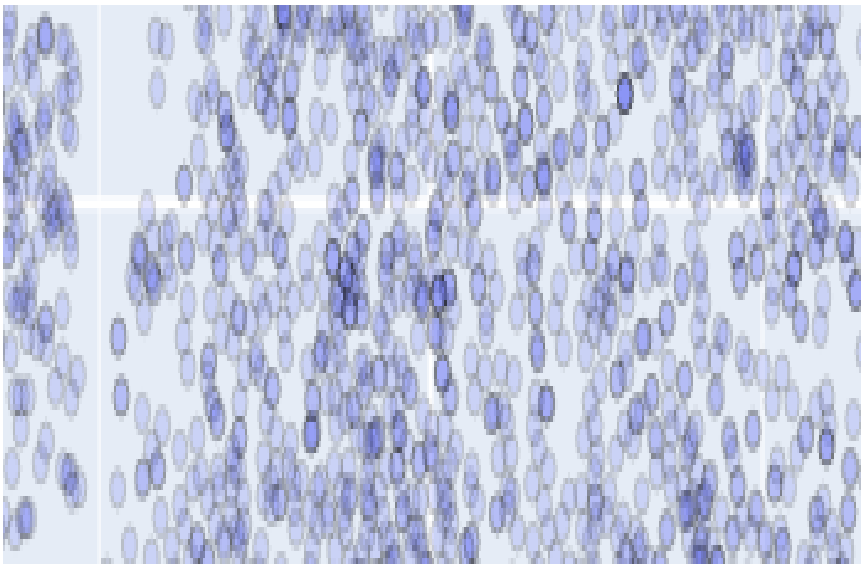


Figure 7: Visualisation plot of building embeddings from satellite images using t-SNE dimensionality reduction. On the segment a) – a region with a high density of points corresponding to typical small buildings is highlighted (Fig. 8); on the segment b) – a cluster of landmark buildings is selected.

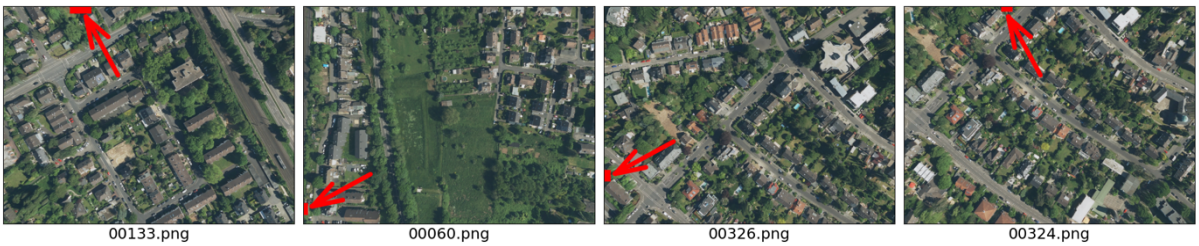


Figure 8: Examples of typical small buildings, outlined with red rectangles, corresponding to selected points in Fig. 7a. These buildings have the distinctive characteristic of being located at the image boundaries, partially extending beyond the image frame

Figure 9: Examples of landmark buildings, outlined with red rectangles, corresponding to points from the highlighted cluster in Fig. 7b.

4.4. Comparison of retrieval accuracy for typical and landmark buildings

The quantitative measurements presented in Table 1 demonstrate that the accuracy of UAV-based searches for landmark buildings nearly doubles compared to searches for typical buildings, thus confirming the efficacy of the proposed approach. The Recall@5 value indicates that incorporating a post-filtering stage for the top-5 most similar buildings could potentially increase the current implementation's Recall@1 up to 0.66.

Several special cases were observed during the evaluation phase. For example, YOLOv11 may detect the same building twice at the building segmentation stage from images. Still, in one instance, YOLO might merge the building with an adjacent one, as illustrated in cases a) and b) in Figure 10. However, the proposed approach effectively generates embeddings that robustly encode semantic and structural information, rendering these representations resilient to CNN segmentation errors; in both cases, the correct matching building was identified successfully.

Table 1
Metrics for the retrieval of typical buildings and landmark buildings

Metric		Recall@K	
		K=1	K=5
Buildings	Landmark	**0.51**	**0.66**
	Typical	0.28	0.46

While cases a) and b) focused on searches of landmark buildings, case c) involved a building classified as typical, characterised by medium size and a visually distinct angular shape. Despite the corresponding satellite image building being rotated by more than 90°, it was accurately identified as the top-1 match, demonstrating the embeddings' robustness to object rotations. Case d) involved a typical building—a long residential structure with an orange roof. Such buildings are numerous in the dataset, and semantically retrieved buildings were correct, matching the elongated rectangular shape and roof colour. However, none matched the UAV-captured building, emphasising the importance of selecting truly unique buildings for search accuracy.

Figure 10: Examples demonstrating the performance of the proposed approach in edge cases. The first row shows the UAV-captured building for which a match is searched from the satellite-based building database. The second row presents the corresponding satellite-based reference building. The subsequent rows illustrate the most similar buildings based on L2 distance embeddings. For compactness, we show only the top-2 most similar samples. In cases a) and b), the same landmark building was segmented twice by YOLOv11—case a) shows correct segmentation, while case b) merges the building with an adjacent one into a single segment. Despite this segmentation discrepancy, the correct corresponding satellite building was successfully identified within the top 5

in both cases, demonstrating the robustness of embeddings against CNN segmentation errors. Case c) involves a uniquely shaped yet classified as a typical building that was rotated more than 90° in the satellite imagery. Despite this rotation, the correct corresponding building was identified as the top-1 match, indicating robustness of embeddings to object rotations. In case d), a typical elongated residential building with an orange roof is considered. Although the retrieved buildings are semantically correct, being elongated rectangles with similar roof colours, none precisely match the UAV-based query building. This emphasises the importance of selecting unique landmark buildings to ensure retrieval accuracy.

4.5. Limitations

The experimental validation in this study was conducted within an urban environment, using buildings as landmarks. Both satellite and UAV images were captured during daylight from the same vertical, top-down perspective.

Significantly, the specific set of landmark objects for a given set of satellite images depends on the convolutional neural network used. Different neural networks might segment the same image differently, potentially failing to identify buildings or merging multiple adjacent buildings into a single segment. Additionally, object embeddings generated by these networks may differ, resulting in variations in the final landmark object set.

The calculation of Recall@1 and Recall@5 metrics for matching accuracy between UAV and satellite images required human labelling of search results. Due to the time-intensive nature of this task, the test dataset size was limited to 200 unique buildings.

4.6. Future work

Future directions for improvement include extending this approach to other types of urban landmarks (e.g., intersections, roads, sports fields) and different environments (e.g., forests, fields).

An interesting research aspect involves the impact of aggregation functions on the embedding space and the objects forming the final landmark set. Combining graph neural networks, trainable via backpropagation, with Contrastive Learning methods [22] could enhance the invariance of object embeddings to variations in viewing angles or lighting conditions.

To create landmark sets without the strict requirement for a fixed number (as seen in Isolation Forest), and considering additional practical constraints, future improvements may involve more flexible outlier detection algorithms. An alternative approach could replace outlier detection with clustering algorithms that do not require a fixed cluster count. Here, landmarks could be represented by objects in tiny clusters or those lying outside of any cluster, with the introduction of supplementary constraints, such as a maximum allowable distance between neighbouring landmarks.

5. Conclusions

This work proposes an approach for identifying unique landmark objects by analysing embeddings obtained from convolutional neural networks. The study aims to enhance UAV localisation by isolating distinctive landmark buildings within an embedding space encoding structural and visual features.

Experimental results successfully met this goal, revealing landmark building identification accuracy nearly twice as high as typical building recognition (Recall@1 = 0.51 and Recall@5 = 0.66 versus 0.28 and 0.46, respectively).

Nevertheless, the current implementation has limitations, notably its application to navigation within urban and suburban environments under good lighting conditions, and dependency on a particular segmentation model.

Future work aims to broaden this approach to various object and terrain types and investigate more adaptable feature aggregation and anomaly detection methods. Such enhancements could expand the system's applicability and navigational accuracy. Ultimately, refining this approach could enable fully automated UAV route planning based on visual features in GPS-denied environments. The only parameters needed would include satellite surface imagery, specific landmark set constraints derived from UAV technical specifications, the selected convolutional neural network deployed on the UAV, and defined start and end route points.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] C. Masone, B. Caputo, A Survey on Deep Visual Place Recognition, IEEE Access 9 (2021) 19516–19547. doi:10.1109/ACCESS.2021.3054937.

[2] A. Ayala, L. Portela, F. Buarque, B. J. T. Fernandes, F. Cruz, UAV control in autonomous object-goal navigation: a systematic literature review, Artif. Intell. Rev. 57 (2024) 125. doi:10.1007/s10462-024-10758-7.

[3] J. Maurício, I. Domingues, J. Bernardino, Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review, Appl. Sci. 13 (2023) 9. doi:10.3390/app13095521.

[4] L. Rundo, C. Militello, Image biomarkers and explainable AI: handcrafted features versus deep learned features, Eur. Radiol. Exp. 8 (2024) 130. doi:10.1186/s41747-024-00529-y.

[5] R. Shwartz Ziv, Y. LeCun, To Compress or Not to Compress—Self-Supervised Learning and Information Theory: A Review, Entropy 26 (2024) 3. doi:10.3390/e26030252.

[6] S. Raxit, S. B. Singh, A. Al Redwan Newaz, YoloTag: Vision-based Robust UAV Navigation with Fiducial Markers, in: Proceedings of the 2024 33rd IEEE International Conference on Robot and Human Interactive Communication, ROMAN '24, IEEE, 2024, pp. 311–316. doi:10.1109/RO-MAN60168.2024.10731319.

[7] V. Lepetit, F. Moreno-Noguer, P. Fua, EPnP: An Accurate O(n) Solution to the PnP Problem, Int. J. Comput. Vis. 81 (2009) 155–166. doi:10.1007/s11263-008-0152-6.

[8] Q. Huang, J. DeGol, V. Fragoso, S. N. Sinha, J. J. Leonard, Optimizing Fiducial Marker Placement for Improved Visual Localization, IEEE Robot. Autom. Lett. 8 (2023) 2756–2763. doi:10.1109/LRA.2023.3260700.

[9] R. Martins, D. Bersan, M. F. M. Campos, E. R. Nascimento, Extending Maps with Semantic and Contextual Object Information for Robot Navigation: a Learning-Based Framework Using Visual and Depth Cues, J. Intell. Robot. Syst. 99 (2020) 555–569. doi:10.1007/s10846-019-01136-5.

[10] A. Babenko, V. Lempitsky, Aggregating Local Deep Features for Image Retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV '15, IEEE Computer Society, 2015, pp. 1269–1277.

[11] G. Tolias, R. Sicre, H. Jégou, Particular object retrieval with integral max-pooling of CNN activations, 2016. arXiv:1511.05879. doi:10.48550/arXiv.1511.05879.

[12] Y. Kalantidis, C. Mellina, S. Osindero, Cross-Dimensional Weighting for Aggregated Deep Convolutional Features, in: G. Hua, H. Jégou (Eds.), Computer Vision – ECCV 2016 Workshops, Springer International Publishing, Cham, 2016, pp. 685–701. doi:10.1007/978-3-319-46604-0_48.

[13] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, J. Sivic, NetVLAD: CNN Architecture for Weakly Supervised Place Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR '16, IEEE Computer Society, 2016, pp. 5297–5307.

[14] T. Qiu, et al., CLDA-YOLO: Visual Contrastive Learning Based Domain Adaptive YOLO Detector, 2024. arXiv:2412.11812. doi:10.48550/arXiv.2412.11812.

[15] Y. Chang, Y. Cheng, U. Manzoor, J. Murray, A review of UAV autonomous navigation in GPS-denied environments, Robot. Auton. Syst. 170 (2023) 104533. doi:10.1016/j.robot.2023.104533.

[16] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, A Comprehensive Survey on Graph Neural Networks, IEEE Trans. Neural Netw. Learn. Syst. 32 (2021) 4–24. doi:10.1109/TNNLS.2020.2978386.

[17] M. Zaffar, et al., VPR-Bench: An Open-Source Visual Place Recognition Evaluation Framework with Quantifiable Viewpoint and Appearance Change, Int. J. Comput. Vis. 129 (2021) 2136–2174. doi:10.1007/s11263-021-01469-5.

[18] O. Rainio, J. Teuho, R. Klén, Evaluation metrics and statistical tests for machine learning, Sci. Rep. 14 (2024) 6086. doi:10.1038/s41598-024-56706-x.

[19] M. Schleiss, F. Rouatbi, D. Cremers, VPAIR-Aerial Visual Place Recognition and Localisation in Large-scale Outdoor Environments, 2022. URL: https://github.com/AerVisLoc/vpair.

[20] G. Jocher, J. Qiu, A. Chaurasia, Ultralytics YOLO, 2023. URL: https://github.com/ultralytics/ultralytics.
[21] F. T. Liu, K. M. Ting, Z.-H. Zhou, Isolation Forest, in: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, ICDM '08, IEEE, 2008, pp. 413–422. doi:10.1109/ICDM.2008.17.
[22] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, F. Makedon, A Survey on Contrastive Self-Supervised Learning, Technologies 9 (2021) 1. doi:10.3390/technologies9010002.