CNNs are explainable domain-specific visual embedders

Zakhar Ostrovsky[1], Andrii Biloshchytskyi[2] and Dmytro Uhryn[3]

[1]Khmelnytskyi National University, 11, Institutes str., Khmelnytskyi, 29016, Ukraine
[2]Astana IT University, Astana 010000, Kazakhstan
[3]Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine

Abstract

Autonomous aerial systems operating in GPS-denied environments require robust visual place recognition (VPR) to ensure safety and reliability. However, creating stable visual embeddings for complex aerial imagery remains challenging due to high dimensionality, viewpoint variability, and occlusion. In this work, we propose a domain-specific, explainable pipeline for UAV geo-localisation that leverages a pre-trained Convolutional Neural Network (CNN) to extract multi-level features of buildings, followed by unsupervised outlier detection to curate distinctive landmarks. By treating landmark selection as an anomaly detection problem, we automatically build a database of visually unique, geo-tagged structures without additional training. Experiments on the cross-view VPAIR benchmark demonstrate that our method substantially outperforms typical scene features, increasing Top-1 recall from 31% to over 53% and doubling precision in visual place recognition. The resulting embeddings are not only more accurate but also highly interpretable, emphasizing salient architectural features over background clutter. These findings suggest that integrating object-centric embeddings with outlier-based distinctiveness provides a lightweight, transparent path toward reliable autonomous navigation.

Keywords

Explainable AI, visual embeddings, UAV navigation, visual place recognition, outlier detection, deep learning

1. Introduction

Autonomous aerial systems increasingly operate where GNSS is intermittent or denied, forcing reliance on onboard vision for localisation and navigation [1]. In such settings, visual place recognition (VPR), matching the current view against a prior map, becomes critical for safety and reliability [2]. Yet, turning raw pixels into stable, reusable visual embeddings is substantially harder than embedding text: images are high-dimensional, viewpoint- and illumination-variant, and rife with occlusion. Effective image embeddings must jointly encode geometry, appearance, and semantics while remaining discriminative across scenes and robust within a scene [3, 4].

A practical route to trustworthy VPR is to ground decisions in explicit, human-interpretable landmarks (e.g., distinctive buildings), as suggested in [5]. Classical robotics emphasises that a good landmark is unique and consistently observable [6]. However, many pipelines either preselect key regions by hand [7] or treat all detections of a class (e.g., "building") as equally valid, which fails in repetitive urban layouts. Deep CNNs learn hierarchical features, from textures to parts to object structure, that can support stronger embeddings [8, 9]. Global deep place recognition (e.g., NetVLAD) aggregates scene evidence effectively [10], but still tends to behave like a black box and may be distracted by background clutter. For explainable autonomy [11], we argue for object-centric, domain-specific embeddings paired with an automated notion of distinctiveness prioritising the few landmarks that genuinely anchor localisation.

The main contribution of this research is a lightweight, explainable pipeline that: (i) extracts multi-layer, object-masked CNN embeddings for buildings without extra training, (ii) applies unsupervised outlier detection to curate a compact set of distinctive, geo-tagged landmarks, and (iii) demonstrates

ExplAI-2025: Advanced AI in Explainability and Ethics for the Sustainable Development Goals, November 07, 2025, Khmelnytskyi, Ukraine

*Corresponding author.
ostrovskyiz@khmnu.edu.ua (Z. Ostrovsky); A.B@astanait.edu.kz (A. Biloshchytskyi); d.ugryn@chnu.edu.ua (D. Uhryn)
0009-0003-4644-3587 (Z. Ostrovsky); 0000-0001-9548-1959 (A. Biloshchytskyi); 0000-0003-4858-4511 (D. Uhryn)

substantial VPR gains when queries are restricted to these landmarks. The approach yields interpretable decisions ("matched this specific building") while boosting accuracy.

The structure of this paper is the following: Section 2 reviews related work on visual embeddings, object-centric retrieval, and UAV VPR. Section 3 details our segmentation-to-embedding and outlier-based landmark selection pipeline. Section 4 reports experiments and analysis. Section 5 concludes.

2. Related works

Early VPR relied on local handcrafted descriptors (e.g., SIFT) and global aggregation (e.g., BoVW) [8, 9]. These methods scale but struggle with large viewpoint changes and lack semantic abstraction. CNNs remedied much of this by learning hierarchical features; NetVLAD integrated a learnable aggregation to produce strong global scene descriptors [10], while DELF emphasised attentive local regions for retrieval [12]. Still, global embeddings can entangle background and foreground, and do not inherently tell which object grounded the match.

Another approach is object-centric retrieval and distinctiveness. Detect-to-Retrieve (D2R) reduces clutter by detecting objects first, then retrieving with object descriptors [13]. Yet most D2R variants implicitly assume all instances of a class are equally informative. In UAV navigation, surveys highlight deep learning methods tailored for GPS-denied operation and cross-view matching challenges [14], with context-enhanced models improving aerial-to-oblique alignment [15]. For detection/segmentation under compute constraints, YOLO-family models provide competitive accuracy-speed trade-offs on aerial data [16]. Manually chosen key regions can help [7], but do not scale or adapt across locales.

A better explainability can be achieved via multi-layer features and outliers. Isolation Forest efficiently surfaces unusual items in high-dimensional spaces, making it a natural tool for landmark curation in an embedding space [17]. Meanwhile, feature visualisation shows intermediate CNN layers capture interpretable patterns (textures, parts), suggesting that multi-layer, object-masked embeddings can be both discriminative and more transparent than global scene vectors [18]. Imaging XAI further supports blending low- and high-level cues to improve interpretability without sacrificing performance [4].

Guided by this landscape, we pursue the following research tasks:

1. Design domain-specific, multi-layer CNN embeddings for buildings that require no additional training but retain semantic and fine-detail cues [8, 9, 10, 12, 18].
2. Select distinctive landmarks by unsupervised outlier detection in embedding space (Isolation Forest) to down-weight repetitive structures [13, 17].
3. Evaluate the pipeline in UAV-relevant cross-view retrieval, comparing landmark-focused queries versus typical buildings to quantify gains in accuracy and explainability [14, 15, 16].

3. Materials and methods

3.1. Problem setup and notation

Let a geo-referenced collection of satellite images be $\mathcal{J} = \{I_n\}_{n=1}^{N}$. A segmentation detector produces a set of building instances.

$$\mathcal{O} = \bigcup_{n=1}^{N} \mathcal{O}(I_n), \quad \mathcal{O}(I_n) = \{(M_k, \alpha_k, \boldsymbol{p}_k)\}, \tag{1}$$

where $M_k \in \{0,1\}^{H\times W}$ is a binary mask, α_k a confidence score, and $\boldsymbol{p}_k$ the geo-coordinate (e.g., centroid) of the object.

We retain candidates with $\alpha_k \geq \theta_{\text{conf}}$ and form $\mathcal{O}_{ke} \subset \mathcal{O}$. Our goal is to learn an embedding map $\phi : \mathcal{O}_{ke} \to \mathbb{R}^d$ and curate a landmark subset $\mathcal{L} \subset \mathcal{O}_{ke}$ whose embeddings are *distinctive* within the operational area. The approach is summarised by the diagram on Figure 1.

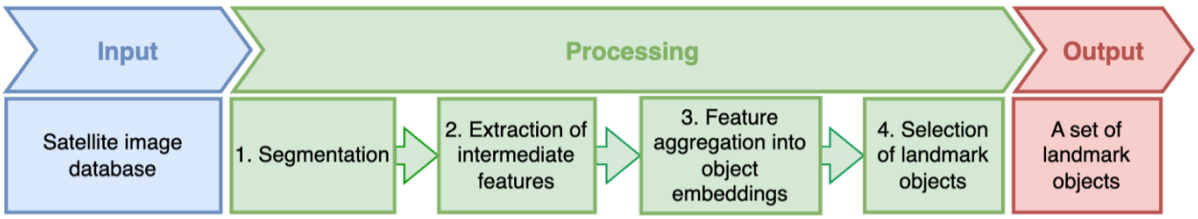


Figure 1: The proposed pipeline steps. Satellite images are processed through segmentation, feature extraction, and aggregation to form object embeddings. Unsupervised outlier detection then selects distinctive landmark objects.

3.2. Segmentation backbone and masking

We employ a lightweight YOLO-family segmentation model to detect buildings and obtain pixel-accurate masks suitable for object-level feature pooling [16]. For an input I, let the CNN backbone produce feature tensors at selected layers $\mathcal{S} = \{l_1, \ldots, l_L\}$:

$$\boldsymbol{F}^{(l)} \in \mathbb{R}^{C_l \times H_l \times W_l}, \quad l \in \mathcal{S}, \tag{2}$$

To align the object mask with each feature map, we apply resolution-aware projection

$$M_k^{(l)} = \Pi_l(M_k) \in \{0,1\}^{H_l \times W_l}, \tag{3}$$

where $\Pi_l(\cdot)$ denotes down/up-sampling consistent with the backbone strides. The masked spatial support for channel c is

$$\Omega_{k,c}^{(l)} = \{(i,j) \mid M_k^{(l)}(i,j) = 1\}. \tag{4}$$

3.3. Object-masked, multi-layer embedding

For each selected layer l and channel c, we aggregate activations over $\Omega_{k,c}^{(l)}$ using a pooling operator $A(\cdot)$:

$$z_{k,c}^{(l)} = A\left(\{\boldsymbol{F}_c^{(l)}(i,j) \mid (i,j) \in \Omega_{k,c}^{(l)}\}\right). \tag{5}$$

We consider three standard, training-free choices:

$$A_{max}(S) = \max_{x \in S} x, \quad A_{mean}(S) = \frac{1}{|S|} \sum_{x \in S} x, \quad A_{sum}(S) = \sum_{x \in S} x. \tag{6}$$

The per-layer descriptor is $\boldsymbol{z}_k^{(l)} = \left[z_{k,1}^{(l)}, \ldots, z_{k,C_l}^{(l)}\right]^\top$. The final object-masked, multi-layer embedding concatenates descriptors across layers:

$$\phi(O_k) = \boldsymbol{e}_k = \bigoplus_{l \in \mathcal{S}} \boldsymbol{z}_k^{(l)} \in \mathbb{R}^d, \quad d = \sum_{l \in \mathcal{S}} C_l. \tag{7}$$

Max pooling emphasizes salient, viewpoint-tolerant responses, mean pooling captures average appearance, and sum pooling is sensitive to object extent; we empirically compare them in Section 4, in line with prior observations about intermediate-layer semantics and interpretability [18], and global/attentive retrieval practice [10, 12].

3.4. Layer subset selection via proxy retrieval metrics

We pick which CNN layers to use by measuring how well different layer combinations help us retrieve the right building under two quick tests:

- S2S (satellite→satellite): queries and database are neighbouring/overlapping satellite crops;

- S2D (satellite→drone): queries are UAV/aircraft views; the database is satellite imagery.

For any candidate set of layers: 1. Build embeddings with those layers for all database buildings and for the query buildings. 2. Nearest neighbour match: for each query, find the most similar database embedding (Euclidean or cosine distance in the embedding space). 3. Top-1 hit rate: count how often the nearest neighbour is the correct building, and divide by the number of queries to get an accuracy score.

To select layers, we run a greedy forward search, summarised in Algorithm 1:

Algorithm 1 Greedy Algorithm for Optimal Layer Selection

Initialization: $S \leftarrow empty\_set()$.
while performance improves **do**
    Step 1. Select the best-performing single layer not yet in S
    Step 2.
    if adding this layer improves performance **then**
        add this layer to S
    else
        break
Output: S

- Start by evaluating every single layer and pick the one with the highest top-1 hit rate on the chosen proxy (S2S or S2D).
- Then, try adding each remaining layer one at a time to the current set; keep the layer that improves the score the most.
- Stop, if adding any new layer no longer improves the score.

This simple procedure consistently prefers a mix of mid-level and deep layers, the former carry texture/part details while the latter encode object shape/semantics, yielding more discriminative and robust embeddings for retrieval, in line with prior findings on global and attentive deep retrieval and intermediate-layer interpretability [10, 12, 18].

3.5. Distinctive landmark selection via Isolation Forest

With embeddings $\{\boldsymbol{e}_k\}$, we estimate distinctiveness using Isolation Forest [17]. The model $\mathcal{T}$ assigns an anomaly score $s_k = \mathcal{T}(\boldsymbol{e}_k) \in [0, 1]$, where higher indicates easier isolation (rarity). We define the landmark set by thresholding (or contamination-controlled quantile τ):

$$\mathcal{L} = \{O_k \in \mathcal{O}_{keep} \mid s_k \geq \tau\}. \tag{8}$$

This unsupervised criterion operationalizes the classical "uniqueness" property of landmarks without manual labels, while remaining computationally light.

3.6. Landmark database and UAV retrieval

After selection, we keep a compact landmark database where each entry stores: (i) the landmark's embedding (the descriptor we computed), (ii) its geo-coordinate from the map (or an ID to the source image). At runtime, the UAV captures a frame, runs the same building segmentation, and computes query embeddings with the exact same masking-and-pooling steps as during mapping. We then search only within the landmark set (not all buildings) for the nearest neighbours to each query embedding. This restriction cuts down confusion among look-alike structures and accelerates matching, both are crucial in GPS-denied flight scenarios [14, 15, 16].

There are two natural evaluation modes:

- Top-1 match, if the best match is confident, we immediately use that landmark's geo-coordinate as the position hypothesis, simple and fast for real-time control loops.
- Top-K + reranker, if we want extra reliability, we take the top-K candidates (e.g., 5) and apply lightweight checks (e.g., view consistency, simple geometry, or IMU priors) to pick the final match. This cascaded setup trades a bit of latency for robustness, which is often desirable in safety-critical flights [14, 15, 16].

To evaluate both modes with one metric, we report Recall@K. It answers: "For how many queries does the correct landmark appear within the top-K retrieved results?" Thus Recall@1 measures the strict, single-guess case (top-1 updates), while Recall@K (>1) captures pipelines that retrieve a small candidate set and then confirm the winner via a reranker or sensor fusion. This aligns directly with how the system is used in practice: instant updates when confident, or top-K shortlists when verification is enabled.

4. Results and discussion

4.1. Dataset

We evaluated the proposed method on a challenging cross-view dataset and through a series of experiments to validate each component of the pipeline. The primary dataset used is VPAIR (Visual Place Recognition, Aircraft Imagery) [19], which provides paired imagery of urban landscapes: high-altitude aerial (satellite-like) photos and low-altitude oblique photos taken from a light aircraft (simulating UAV camera views) [20].

(a) UAV-view (b) Satellite

Figure 2: The VPAIR dataset sample images, UAV-view vs satellite.

The area covered in VPAIR includes a mix of dense city blocks, industrial sites, and open areas, making it a good testbed for landmark selection. We processed the aerial images to build a landmark database, then tested localisation by using the corresponding low-altitude images as queries (Figure 2).

4.2. The qualitative properties of aggregation functions and one-layer embeddings

Why does the aggregation function matter? A building embedding is constructed by pooling activations that fall inside the object mask. The choice of pooling function determines what statistical summary of the activation multiset is preserved, in much the same way that read-out operators determine the expressive power of Graph Neural Networks (GNNs). Drawing an analogy to the analysis of multiset aggregators in the GIN framework [21], we can interpret the three functions used here as follows:

1. Sum-pooling retains multiplicities: larger or more strongly activated regions contribute proportionally more to the resulting vector. For buildings, this means that overall mass and footprint, for example, long façades or large warehouse roofs, dominate the descriptor. Thus, the obtained building embedding will largely depend on the building size. Although sum pooling has the potential to represent different buildings the most distinctly out of the trio, its dependence on the building's size may significantly influence the embedding when the building image is taken from different altitudes.
2. Average-pooling captures proportions while discarding scale. It emphasises the distribution of visual patterns and, as such, it is expected to be the most expressive in terms of texture or material, making subtle roof details (solar panels, fine tiling) prominent even on compact structures.
3. Max-pooling reduces the multiset to a simple set that records only the strongest response per channel. It highlights salient local cues, distinct corners, towers, colour patches, irrespective of object size.

Hence, we expect sum-pooling to favour geometrically atypical or very large buildings, average-pooling to surface textural oddities, and max-pooling to benefit most once deeper, semantically rich layers are consulted.

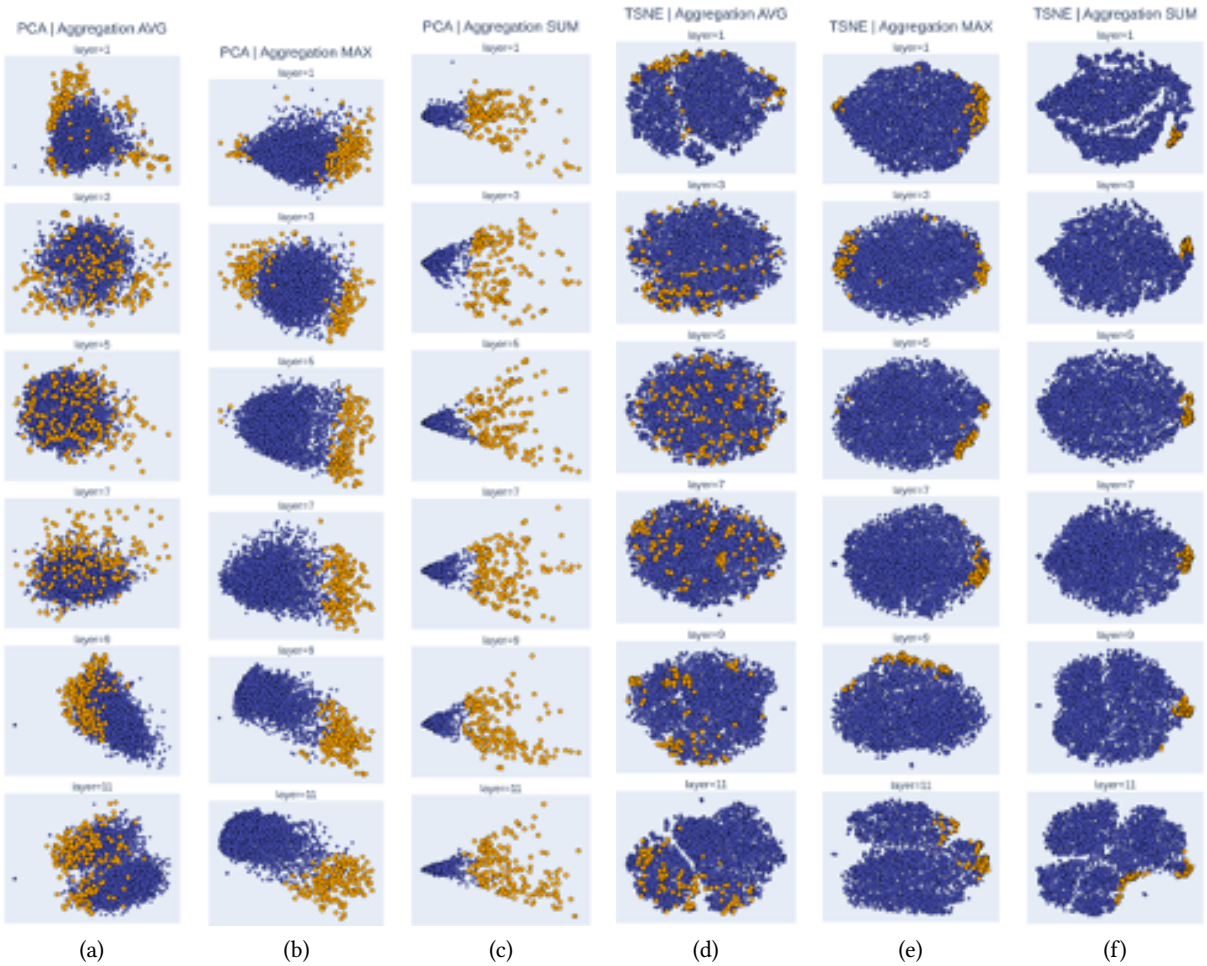


Figure 3: PCA (**a, b, c**) and t-SNE (**d, e, f**) visualisations of building embeddings obtained from single CNN layers using different aggregation functions. Each row corresponds to a specific layer index (from 1 to 11, in steps of 2), and each column shows a different aggregation method. Blue points represent typical buildings; orange-highlighted points are those selected as outliers by the Isolation Forest. Clear separation of outliers indicates more effective structural or semantic discriminativeness in the embedding space.

Thus, to make the initial assessment of different aggregation functions and feature layers, we followed

the experimental protocol described further. For each of six convolutional layers (indices 1, 3, 5, 7, 9, 11), we formed embeddings from that single layer only, applying the three aggregation functions individually. Isolation Forest (500 trees, 1% contamination) marked outliers, our candidate landmarks. The resulting embedding clouds were projected to 2-D via PCA and t-SNE to visually assess landmark separability. Figure 3 compiles the six plots per aggregator.

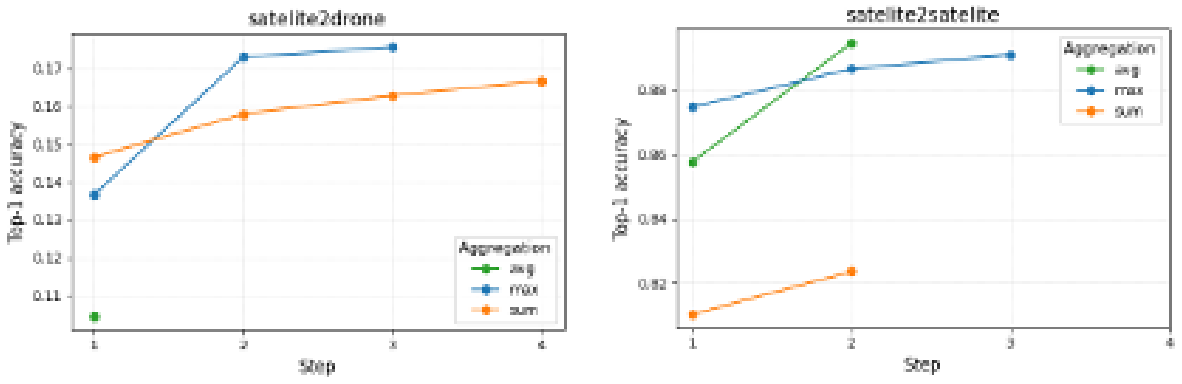
Average-pooling (Fig. 3a and 3d). Across all layers, the brown outlier points are interwoven with blue inliers, forming no isolated clusters. Manual inspection of outliers reveals many small, otherwise typical houses with atypical roof textures, precisely the fine-grained cues that average-pooling amplifies. However, the weak structural signal makes these landmarks hard to separate automatically.

Max-pooling (Fig. 3b and 3e). Layers 1–5 show partial overlap: deeper spectral features have not yet matured into strong semantic detectors. Starting at layer 7, a clear split emerges, and by layers 9–11, the outliers form a compact lobe on the right of the PCA plot and at the periphery of the t-SNE map. The selected landmarks are visually striking buildings, unusual footprints, vivid colours, confirming that max-pooling benefits from the higher-level abstractions encoded in late layers.

Sum-pooling (Fig. 3c and 3f). Separation is already pronounced at layer 1 and grows steadily. Outliers correspond to large floor-area constructions (shopping centres, factories) that accumulate high activation mass even in shallower feature maps. The wedge-shaped PCA distribution indicates that embedding magnitude acts as a proxy for object size. These behaviours support the theoretical expectations above and validate our first sub-hypothesis: the embedding space does encode semantic and structural cues, with the nature of those cues dependent on the aggregation operator and CNN depth.

4.3. Proxy criteria for the selection of the embeddings configuration

To rigorously test embedding discriminativeness, we deliberately increased the difficulty of the evaluation setup. The Isolation Forest algorithm, previously applied with a 1% contamination threshold, was adjusted to a much higher threshold of 20%. This increased threshold ensured that the pool of outliers considered as candidate landmarks was significantly larger, more diverse, and inherently less distinctive. We also strictly evaluated Top-1 accuracy instead of a more forgiving Top-5 criterion used in other experimental phases, thus enforcing stringent embedding quality demands and ensuring that our selected embeddings truly represent buildings with highly distinctive characteristics.



(a) S2D accuracy

(b) S2S accuracy

Figure 4: Proxy metrics improvement with additional layers during greedy layer selection. The left subplot demonstrates the evolution of the Satellite2Drone metric (cross-domain generalisation), while the bottom subplot illustrates the Satellite2Satellite metric (within-domain consistency). Each line represents a different aggregation strategy (average, max, sum). Steps indicate the sequential addition of CNN layers selected by the greedy algorithm. Max aggregation shows the best-balanced performance across both metrics, while average aggregation underperforms notably on the cross-domain task despite good within-domain results.

The results from this quantitative evaluation clearly demonstrate meaningful distinctions among aggregation strategies and layers. Fig. 4 illustrates the progression of Top-1 accuracy improvements

as layers are incrementally added, and Table 1 succinctly presents the final selected layers for each combination of proxy metric and aggregation function.

Table 1
Embeddings' best performance and configuration

Proxy Metric	Aggregation	Selected layers	Recall@1
S2D	max	[9, 6, 10]	**0.175**
S2D	sum	[6, 9, 10, 8]	0.166
S2D	avg	[9]	0.105
S2S	max	[9, 10, 4]	0.889
S2S	sum	[10, 9]	0.823
S2S	avg	[9, 10]	**0.895**

The analysis reveals several insights. The max aggregation function consistently benefits from combining deeper semantic layers (9, 10) with intermediate structural layers (4, 6), yielding superior generalization across domains (S2D) and consistency within domains (S2S). This aligns with our earlier theoretical speculation that max pooling preserves distinctive local features and semantic signals effectively, confirming its suitability for reliable landmark embeddings.

Sum aggregation, while offering a robust baseline, showed limited incremental gains when adding deeper layers, particularly evident in the S2S metric. Its strong initial performance, visible even with shallow layers, confirms our qualitative observation that sum pooling naturally emphasises large-scale structures. Nonetheless, additional layers proved beneficial in bridging the satellite-to-drone domain gap.

Average aggregation exhibited a striking discrepancy across the two metrics. It achieved the highest accuracy in S2S but notably underperformed in the more challenging cross-domain S2D metric. We assume this discrepancy arises because average pooling emphasises subtle textural and material cues, such as roofing patterns, that remain stable within a domain but degrade significantly when viewpoints and sensor characteristics differ dramatically, as is the case between satellite and drone imagery.

Overall, considering both theoretical arguments and these quantitative outcomes, we conclude that embeddings formed via the max aggregation method, specifically layers [9, 6, 10] for the S2D task, provide the best balance of structural, semantic, and cross-domain discriminative capabilities. Consequently, these embeddings are most suitable for robust landmark identification and subsequent UAV navigation tasks, clearly supported by both our previous theoretical speculations and current quantitative analyses.

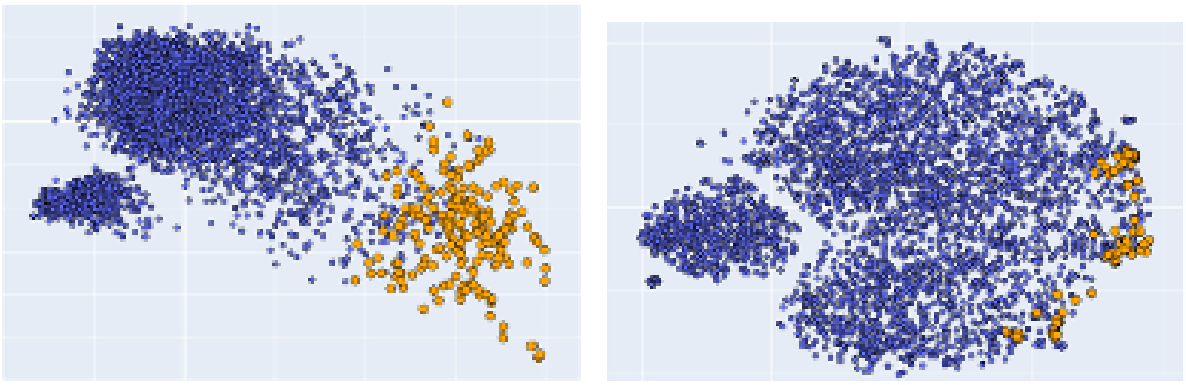
4.4. Embedding space of the top-rated configuration

The quantitative study has previously singled out the max-pooled embedding built from layers [9, 6, 10] as the most reliable representation. We now inspect the geometry of this embedding space and verify that landmark buildings occupy distinctive, well-separated regions.

PCA and t-SNE dimensionality reduction algorithms were employed to inspect the global structure. Figure 5 juxtaposes two-dimensional projections of all satellite-image embeddings produced by the best configuration. In each plot, orange dots denote buildings flagged as landmarks by the Isolation Forest (contamination = 0.2), while blue dots correspond to typical buildings.

In the PCA view (Fig. 5a), the majority of embeddings form a compact cloud to the upper left. From this dense core, a sparse, elongated branch extends down-right, ending in a clearly detached cluster of orange points. The continuous transition from core to tail suggests a spectrum of visual distinctiveness: small, repetitive residences populate the high-density nucleus, whereas progressively more unusual structures migrate towards the periphery.

The t-SNE map (Fig. 5b) echoes this picture with higher non-linear fidelity. It displays several tight islands of nearly identical embeddings; most lie inside the blue core, but a pronounced orange enclave



(a) (b)

Figure 5: Embedding projections for the best-performing configuration. (**a**) PCA; (**b**) t-SNE. Orange, landmarks, Blue, typical buildings.

appears on the right fringe. Because t-SNE preserves local neighbourhoods, such edge clustering indicates that landmark embeddings are indeed far from typical ones in the high-dimensional space, not merely artefacts of linear projection.

To illustrate how these projections relate to concrete urban scenes, Figure 6 enlarges two annotated areas from the PCA plot.

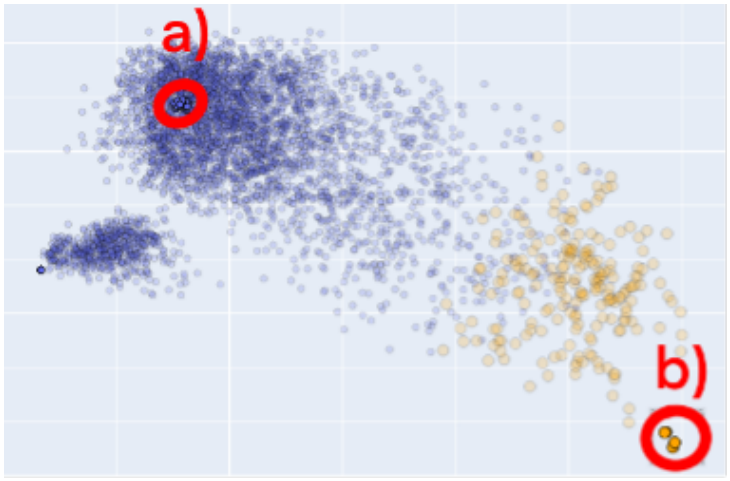


Figure 6: PCA close-ups. Region **a**) encloses densely packed typical houses; Region **b**) isolates a landmark cluster.

Region a) sits deep inside the blue core. It contains a high concentration of points whose embeddings are almost indistinguishable. Visual examination confirms that these correspond to small, rectangular family houses with homogeneous grey roofs, by far the most frequent pattern in the dataset (examples in Fig. 7).

Region b) lies at the extreme tip of the orange branch. The highlighted cluster comprises only landmark points. The corresponding buildings (Fig. 8) are large, architecturally irregular complexes, shopping malls, sports halls, L-shaped blocks, whose footprints and textures deviate strongly from suburban norms. Their separation validates the outlier-based landmark criterion.

A similar analysis on the t-SNE embedding (Fig. 9) yields consistent conclusions. Within the blue nucleus we find a miniature cluster, again marked a), that gathers houses partially cropped by image borders (Fig. 10). On the opposite flank, marker b) points to the same striking structures already observed in PCA (Fig. 11). The fact that both linear and non-linear projections isolate identical landmark sets reinforces the robustness of the learned representation.

An encouraging observation is that many orange points occur in small, self-contained groups of

Figure 7: Examples from Region a. Compact single-family dwellings with uniform roofs.

Figure 8: Examples from Region b. Large, irregular buildings, prime candidates for navigation landmarks.

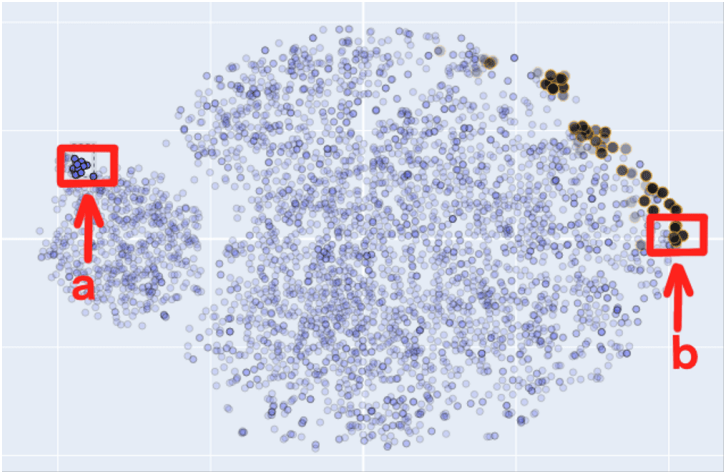


Figure 9: t-SNE close-ups. Region **a**) aggregates partially cropped houses; Region **b**) groups the same landmarks as in PCA.

Figure 10: Examples from t-SNE Region a. Typical houses located at image boundaries.

two or three. Manual inspection shows these groups have repeated detections of the same landmark building in consecutive frames. The tight clustering of their embeddings indicates strong invariance to minor changes in viewing angle, as was investigated and proved in [22], illumination, and partial occlusions, an essential property for reliable UAV localisation.

Conversely, blue points that stray into low-density outskirts often correspond to buildings that are visually similar yet geographically distant from each other. Their presence cautions that, although our method suppresses most ambiguity, truly fool-proof disambiguation requires either a larger landmark

Figure 11: Examples from t-SNE Region b. Distinctive landmark buildings match those in Figure 9.

pool or an additional geometric consistency check, an avenue we explore in future work.

In summary, the qualitative evidence aligns with earlier quantitative findings: the max-pooled, multi-layer embedding carves out a well-structured space where visually distinctive buildings occupy separable, easily identifiable regions.

4.5. Comparison of retrieval accuracy for typical and landmark buildings

For quantitative evaluation, it was necessary to establish a manual benchmark set due to the lack of ground truth correspondences in VPAIR. This set comprised 200 manually annotated buildings: 100 landmark and 100 typical buildings. To measure retrieval effectiveness from UAV-captured buildings back to satellite imagery, embeddings for each of the 200 manually annotated UAV buildings were compared to all satellite-derived embeddings using the L2 norm. Retrieval performance was quantified using the metrics Recall@1 and Recall@5, computed independently for landmark and typical buildings, thus objectively demonstrating the relative advantage of landmark selection. The results for the best-performing embedding configuration are summarised in Table 2.

Table 2
Retrieval performance

		Recall@K	
Metric		K=1	K=5
Buildings	Landmark	**0.53**	**0.70**
	Typical	0.31	0.51

A clear gap appears: searches that target the automatically selected landmark set succeed almost twice as often as searches for ordinary buildings. In particular, Recall@1 rises from 0.30 for typical structures to 0.53 for landmarks, while Recall@5 climbs from 0.51 to 0.70. The latter figure suggests that a lightweight re-ranking of the top-5 candidates could push single-shot accuracy close to 0.70 without altering the core pipeline.

4.6. Limitations

The present study is confined to a single public dataset, VPAIR, whose drone imagery was captured under favourable daylight and near-nadir conditions. Consequently, the learned embeddings have not yet been stress-tested against seasonal changes, low-sun shadows, or highly oblique UAV views. A second constraint is the reliance on YOLOv11-nano for building segmentation. Although qualitative checks confirm good cross-dataset generalisation, occasional mask errors reveal that downstream performance is ultimately bounded by segmentation quality. As the exact landmark embeddings are model-dependent, it is important that the same CNN is used both for the UAV on-board camera and the landmarks preparation pipeline.

Evaluation, too, is approximate. In the absence of authoritative building-to-building correspondences, we circumvent these limitations with (i) manually labelled pairs for the final benchmark and (ii) proxy metrics that exploit index proximity. While the latter proved effective for layer selection, they assume

both sufficient image overlap and uniform flight speed during the dataset construction and do not guarantee the global optima of the solution. Finally, Isolation Forest employs a fixed contamination rate; adapting this hyper-parameter to scenes with markedly different object density remains an open problem.

5. Conclusions

We introduced an end-to-end framework that automatically discovers visually distinctive urban landmarks and harnesses them for UAV localisation when GNSS is unreliable. The core idea is simple yet powerful: extract multi-layer CNN features for each object, aggregate them into a semantically rich embedding, and treat the landmark objects as outliers in the resulting embedding space as a natural way to select the most distinctive objects automatically. We proposed a lightweight approach, which uses a greedy search and two proxy retrieval metrics, to guide the selection process of the optimal embedding parameters without the requirement of ground-truth labels. The selected embedding, max pooling over layers 9, 6, 10, doubles Recall@1 compared with typical buildings (0.53 vs 0.31) and achieves 0.70 Recall@5, demonstrating that true landmarks are indeed easier to recover. Qualitative visualisations confirm that these embeddings carve out well-separated clusters for architecturally unique structures while remaining stable under viewpoint shifts, segmentation noise, and large rotations. Taken together, the results validate both the theoretical intuition that max pooling preserves salient cues and the practical viability of outlier-based landmark selection. While future work will address dynamic graph structures and geometric verification to filter residual false positives, the method currently stands as a potent, drop-in module for robust GPS-denied navigation.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

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, Artificial Intelligence Review 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, Applied Sciences 13 (2023) 9. doi:10.3390/app13095521.

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

[5] E. Manziuk, W. Wojcik, O. V. Barmak, I. V. Krak, A. Kulias, V. A. Drabovska, V. M. Puhach, S. Sundetov, A. Mussabekova, Approach to creating an ensemble on a hierarchy of clusters using model decisions correlation, Przegląd Elektrotechniczny 96 (2020) 108–113. doi:10.15199/48.2020.09.23.

[6] S. Se, D. Lowe, J. Little, Global localization using distinctive visual features, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2002, pp. 226–231. doi:10.1109/IRDS.2002.1041393.

[7] M. Karnes, J. Riffel, A. Yilmaz, Key-region-based UAV visual navigation, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLVIII-2 (2024) 173–179. doi:10.5194/isprs-archives-XLVIII-2-2024-173-2024.

[8] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110. doi:10.1023/B:VISI.0000029664.99615.94.

[9] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1–8. doi:10.1109/CVPR.2007.383172.

[10] R. Arandjelović, 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), 2016, pp. 5297–5307. doi:10.1109/CVPR.2016.572.

[11] P. Radiuk, O. Barmak, E. Manziuk, I. Krak, Explainable deep learning: a visual analytics approach with transition matrices, Mathematics 12 (2024) 1024. doi:10.3390/math12071024.

[12] H. Noh, A. Araujo, J. Sim, T. Weyand, B. Han, Large-scale image retrieval with attentive deep local features, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3456–3465. doi:10.1109/ICCV.2017.373.

[13] C. H. Song, J. Yoon, T. Hwang, S. Choi, Y. H. Gu, Y. Avrithis, On train-test class overlap and detection for image retrieval, arXiv preprint arXiv:2306.02484 (2024).

[14] O. Y. Al-Jarrah, A. S. Shatnawi, M. M. Shurman, O. A. Ramadan, S. Muhaidat, Exploring deep learning-based visual localization techniques for UAVs in GPS-denied environments, IEEE Access 12 (2024) 113049–113071. doi:10.1109/ACCESS.2024.3440064.

[15] Y. Xu, M. Dai, W. Cai, W. Yang, Precise GPS-denied UAV self-positioning via context-enhanced cross-view geo-localization, arXiv preprint arXiv:2502.11408 (2025).

[16] C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7464–7475. doi:10.1109/CVPR52729.2023.00721.

[17] F. T. Liu, K. M. Ting, Z.-H. Zhou, Isolation forest, in: Proceedings of the IEEE International Conference on Data Mining (ICDM), 2008, pp. 413–422. doi:10.1109/ICDM.2008.17.

[18] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision – ECCV 2014, Springer, 2014, pp. 818–833. doi:10.1007/978-3-319-10590-1_53.

[19] M. Schleiss, F. Rouatbi, D. Cremers, VPAIR – Aerial visual place recognition and localization in large-scale outdoor environments, 2022. doi:10.48550/arXiv.2205.11567.

[20] S. Javaid, M. A. Khan, H. Fahim, B. He, N. Saeed, Explainable AI and monocular vision for enhanced UAV navigation in smart cities: prospects and challenges, Frontiers in Sustainable Cities 7 (2025) 1561404. doi:10.3389/frsc.2025.1561404.

[21] K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks?, in: Proceedings of the International Conference on Learning Representations (ICLR), 2019, pp. 1–17. URL: https://openreview.net/forum?id=ryGs6iA5Km.

[22] O. Barmak, I. Krak, E. Manziuk, Diversity as the basis for effective clustering-based classification, in: Proceedings of the 9th International Conference on Information Control Systems & Technologies (ICST 2020), volume 2711, CEUR-WS.org, Aachen, 2020, pp. 53–67. URL: https://ceur-ws.org/Vol-2711/paper5.pdf.