Where Am I in the Museum? Metadata-Driven Visual Localization via Image Retrieval from 360-degree LiDAR scans

Gioele Litrico[1], Roberto Rizza[1], Mattia Litrico[1], Dario Allegra[1], Anna Maria Gueli[2], and Filippo Stanco[1]

Department of Mathematics and Computer Science, University of Catania, Italy[1]
Department of Physics and Astronomy "Ettore Majorana", University of Catania, Italy[2]
{gioele.litrico, mattia.litrico}@phd.unict.it,
{roberto.rizza, dario.allegra, anna.gueli, filippo.stanco}@unict.it

Abstract. Indoor localization remains a challenging problem due to the limitations of Global Navigation Satellite Systems (GNSS) in enclosed environments and the practical constraints of infrastructure-based solutions such as Bluetooth, WiFi, Ultra Wide Band (UWB), and Radio-Frequency (RF) systems. Recent vision-based approaches have shown promising results, but they often require manually annotated pose data, computationally expensive reconstruction pipelines, or additional sensor calibration and alignment procedures.

In this paper, we propose a lightweight vision-based framework for indoor visitor localization in museum environments using frames extracted from smartphone-recorded videos. We approach localization as a content-based image retrieval task, matching query frames against a reference database generated from 360-degree LiDAR acquisitions captured with a Matterport scanning system. By directly leveraging the spatial metadata and acquisition graph natively provided by Matterport scans, the framework automatically associates visual observations with spatial coordinates without requiring any manual annotation.

We evaluate the proposed framework at the Museum of Archaeology of the University of Catania (MAUC), demonstrating strong localization performance in a real-world cultural heritage environment. Experimental results show that the proposed approach achieves up to 88.54% exact viewpoint and 91.56% tolerance-aware localization accuracy, while operating in real time at only 80 ms per frame on lightweight hardware.

Keywords: Indoor Museum Localization · Content-Based Image Retrieval · LiDAR-based visual database

1 Introduction

Positioning or localizing refers to the determination of the absolute or relative placement of an entity in a 3D space. In outdoor environments, the localization is typically obtained through analysis of radio signals from Global Navigation

Satellite Systems (GNSS), often yielding accuracy on the order of 10 meters [22]. However, GNSS systems face significant limitations in indoor environments due to multiple technological challenges, including multipath propagation [13], signal attenuation caused by building materials [18], insufficient satellite visibility [3, 30], and signal distortion from electronic devices in enclosed spaces [8].

To address these limitations, mainstream indoor localization systems have predominantly adopted Bluetooth, WiFi, Ultra Wide Band (UWB) [14], and Radio-Frequency (RF)-based technologies [24]. However, bluetooth localization is highly affected by signal reflection and refraction in indoor environments, resulting in reduced transmission stability and localization accuracy [38]. WiFi-based localization need for multiple access points, drastically increasing costs when the size of the environment scales [14]. Similarly, UWB technology have hardware-specific requirements that limit its adoption in large-scale deployment scenarios [4, 6]. Instead, RF-based solutions are significantly affected by multipath propagation and signal interference in enclosed environments [7].

Vision-based indoor localization has recently emerged as a promising alternative that relies only on visual information without requiring dedicated positioning infrastructures. Early methods were based on handcrafted feature matching and spatially annotated image databases [19, 37], while later approaches introduced Visual SLAM (Simultaneous Localization and Mapping) systems to jointly perform localization and environment reconstruction [23,26,33]. However, these approaches require pre-built and spatially annotated image databases and computationally expensive reconstruction pipelines. More recent deep learning methods leverage convolutional neural networks to enable real-time localization on mobile devices [17, 31, 36], and further improvements have been achieved through semantic-based methods [12, 28], multimodal sensing [11, 20, 25], and domain adaptation strategies [15, 21]. However, despite their effectiveness, these approaches often rely on supervised pose annotations, additional sensor calibration, or complex preprocessing and alignment stages, limiting their scalability and practical deployment in large indoor environments. To address these limitations, we propose a vision-based framework for visitor localization in museum environments using frames extracted from smartphone-recorded videos. Localization is performed through a content-based image retrieval approach, matching the query frames against a reference database generated from 360-degree LiDAR acquisitions captured at strategically selected viewpoints throughout the museum using a Matterport scanning system. Each acquisition viewpoint is associated with spatial metadata providing its spatial coordinates, avoiding the necessity of manual annotations.

We test our framework at the Museum of Archaeology of the University of Catania (MAUC), demonstrating that the proposed retrieval-based localization pipeline achieves up to 88.54% exact viewpoint accuracy and 91.56% under tolerance-aware evaluation, while maintaining real-time performance at only 80 ms per frame on lightweight hardware. Extensive ablation studies further show that simple brightness and contrast normalization effectively mitigates the domain gap between smartphone query images and Matterport renderings, and

that lightweight architectures such as MobileNetV3-Large outperform deeper ResNet-based models both in accuracy and computational efficiency. These results highlight the potential of image-retrieval based localization as scalable solution for indoor cultural heritage navigation and visitor assistance systems.

The contributions of the paper can be summarized as follows:

- We propose a lightweight vision-based framework for indoor visitor localization in museum environments based on content-based image retrieval from smartphone video frames.
- We introduce an annotation-free localization pipeline that directly leverages the spatial metadata and acquisition graph generated by Matterport LiDAR scans, eliminating the need for manual pose annotation, reconstruction stages, or external alignment procedures.
- We build a viewpoint-aware reference database for the MAUC museum from 360-degree LiDAR acquisitions, enabling automatic association between visual observations and spatial coordinates.
- We evaluate the proposed framework on a real-world scenario at the MAUC museum, achieving 88.54% and 91.56% in exact viewpoint and tolerance-aware localization performance while operating in real time on lightweight hardware.

2 Related Works

2.1 Vision-based Indoor Localization

Indoor localization is challenged by GNSS limitations in enclosed environments, where signal attenuation and multipath degrade positioning accuracy. This has driven research toward image-based approaches, which leverage visual information alongside advances in computer vision and deep learning. Early works rely on handcrafted features: Liu et al. [19] use local descriptors to match query images against a reference database, while [37] improve efficiency by combining ORB features with locality-sensitive hashing. Although computationally efficient, these methods remain sensitive to illumination changes, viewpoint variations, and repetitive textures. To address these limitations, Visual SLAM methods enable joint localization and mapping: [33] combine point and line features in a visual-inertial framework, [26] leverage RGB-D data for improved geometric accuracy, and [23] propose a lightweight visual odometry solution, though prone to drift over time. More recently, deep learning has been integrated into SLAM pipelines, e.g. to filter outliers [34] or combine semantic understanding with depth estimation [32]. Beyond SLAM, several works propose end-to-end deep learning approaches for direct camera localization, including multi-scale feature aggregation [31], refined CNN architectures [36], temporal modeling across frames [1], and lightweight architectures for real-time mobile applications [17]. A further advancement is represented by semantic-based methods, which incorporate high-level scene understanding into the localization process. Uygur et al. [28] exploited semantic cues extracted from panoramic images to improve robustness,

while [12] combined semantic and contextual information with image matching to enhance localization performance in complex environments. In parallel, retrieval and feature matching-based approaches have been developed, where localization is achieved by comparing query images with a database of reference images. Wang et al. [29] proposed a hybrid method that combines learned feature representations with key-frame selection to improve efficiency and scalability, reflecting a shift toward learned descriptors and efficient matching strategies. Recent research has also addressed robustness in challenging scenarios, such as low-light environments and perceptual degradation. The authors of [35] proposed a system based on thermal imaging combined with geometric constraints, enabling reliable localization even in the absence of visible light. Finally, several works explore multimodal approaches, combining visual data with additional sources of information to enhance localization performance. The work of [20] integrated visual and geometric features within a deep learning framework, while [25], combined multiple sensors to improve robustness. Similarly, [11] proposed a system that integrates deep visual features with sensor data to achieve accurate and real-time indoor localization.

2.2 Structured 3D Scanning for Indoor Reference Map Construction

The construction of structured visual databases from 3D scans has emerged as a foundational step for image-based indoor localization. Chang et al. [5] introduced Matterport3D, a large-scale RGB-D dataset collected using the Matterport scanning platform, comprising 10,800 panoramic views from 194,400 RGB-D images across 90 building-scale scenes. The dataset provides camera poses, surface reconstructions, and semantic annotations, enabling a variety of scene understanding tasks including camera relocalization. Crucially, the Matterport platform generates structured metadata, including viewpoint identifiers and precise 3D coordinates, automatically during acquisition, without requiring manual annotation. Building on the availability of such structured reference databases, Taira et al. [27] proposed InLoc, an indoor visual localization framework that uses panoramic 3D scans as a reference database against which smartphone-captured query images are matched. The database images are perspective cutouts extracted from panoramic scans acquired at discrete positions, and localization is achieved through image retrieval followed by dense feature matching and pose verification. This work establishes a paradigm closely related to ours: a structured scan-derived reference database is queried at inference time using images captured by a handheld device, with no need for real-time sensor infrastructure. Differently from InLoc, which requires manual annotation of query poses for evaluation, our framework fully exploits the automatic metadata provided by the Matterport platform, including viewpoint identifiers, room assignments, and 3D coordinates, eliminating any manual labeling overhead. Furthermore, the spatial topology of viewpoints, encoded as a graph of 3D positions, is leveraged to define tolerance-based evaluation metrics that reflect the sub-meter precision requirements of museum navigation applications.

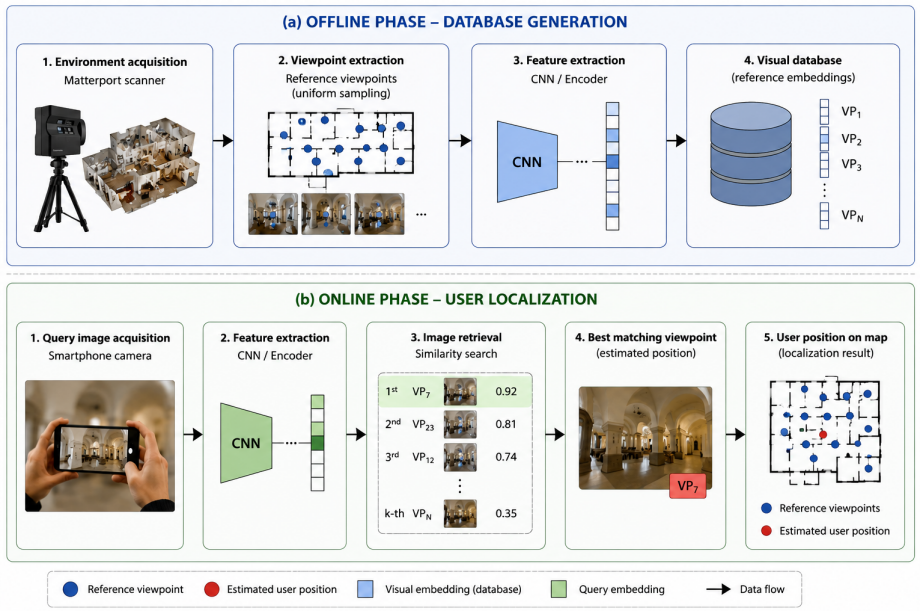


Fig. 1: The figure shows the methodology behind our framework. In the offline phase, a visual database is generated using a Matterport Pro3 LiDAR scanner. In the online phase, frames extracted from user-recorded videos are matched against the visual database to recognize the position.

3 Proposed framework

In this work, we propose a computer vision–based framework aimed at estimating the position of a user in an indoor environment without requiring human annotations (fig. 1). Specifically, the proposed approach relies on image retrieval to identify the location by leveraging a set of predefined viewpoints, used as reference positions and acquired through a Matterport scanner. The framework uses frames extracted from a video recorded by the user with its smartphone camera, and compares them with the viewpoints. This process identifies the most likely viewpoint corresponding to the user's position, effectively preventing the need for additional sensors or infrastructure.

3.1 Data acquisition

We collected two datasets from the MAUC: (i) the reference dataset containing LiDAR scans of museum halls acquired from pre-determined viewpoints; (ii) a testing dataset with videos simulating a visitor which navigates the museum.
Reference Dataset. This dataset includes 1,492 images covering the five exhibition rooms of the museum, acquired using a Matterport Pro3 LiDAR scanner. The device captures full 360° panoramic scans (295° vertically) in less than 20

Fig. 2: MAUC museum images acquired from the Matterport scanner, used to create our visual database.

Fig. 3: MAUC museum images acquired using a smartphone camera while walking throughout its halls.

seconds per sweep, featuring a 20 MP image sensor with ultra-wide angle lens, an output panoramic resolution of 134.2 MP, and a depth resolution of approximately 1.5 million points per scan at a rate of 100,000 points per second, with a spatial accuracy of ± 20 mm at 10 m range. During the acquisition phase, the Matterport device was positioned at a set of 38 predefined locations, referred to as viewpoints, from which full 360-degree panoramic images were captured. Subsequently, the dataset is constructed by extracting perspective images from each panoramic view at regular angular intervals of 20 degrees along the horizontal axis. For each horizontal direction, three images are generated by varying the pitch angle: one at 0°, one tilted downward by 10°, and one tilted upward by 10°. Examples of the acquired images are shown in fig. 2. Since the viewpoint positions and adjacency relationships are automatically provided by the Matterport platform as structured metadata, it does not require any manual annotation. Specifically, for each viewpoint the platform exports metadata, such as a unique identifier (`sid`/`uuid`), the 3D position coordinates (x, y, z) derived from the LiDAR sensor of the Matterport Pro3, the camera rotation, and the list of adjacent viewpoints (`neighbors`). These metadata are organized as a spatial graph in which each viewpoint is connected to its three nearest neighbors. This graph structure is later leveraged to define the tolerance-based evaluation metrics described in Section 4.

Testing Dataset. This dataset includes 1001 images of the museum obtained from videos recorded using a smartphone camera, simulating a visitor navigating the museum. Videos are acquired at eye-level to simulate a realistic usage scenario in which a visitor naturally explores the environment while holding a smartphone. Each frame has been then labeled with the nearest viewpoint. Examples of the acquired images are shown in fig. 3.

3.2 Image-based indoor localization

Extracting Features Representation. For each frame extracted from the user video, we identify the most similar viewpoint to infer the spatial localization of the visitor. With this aim, we extract feature representations $z_q = f(q)$ and $z_i = f(x_i)$ from query frames and reference images respectively, where $f(\cdot)$ is a CNN, q is the query and $\{x_i\}_{i=1}^{N}$ are the reference images. We then ℓ_2-normalize these features vectors before using them in the image retrieval process.

Retrieving the nearest viewpoints. We build upon a Content-Based Image Retrieval (CBIR) approach [16] that predicts the visitor position by matching images extracted from a user-recorded video against a reference database. Let $q \in \mathbb{R}^{3\times H\times W}$ denote the query image and $\mathcal{D} = \{x_i \in \mathbb{R}^{3\times H\times W}\}_{i=1}^{N}$ the reference dataset. The goal is to retrieve the K nearest reference images with respect to the query. To efficiently perform the retrieval, we employ Annoy [2], an open-source library designed for approximate nearest neighbor search with low memory footprint. Given the query z_q and the reference images embeddings z_i, nearest neighbors are retrieved according to the angular distance computed in the embedding space. Formally, the set of nearest neighbors is defined as:

$$\mathcal{N}_K(z_q) = \{z_{(i)}\}_{i=1}^{K} \tag{1}$$

where $z_{(i)}$ represents the i-th nearest neighbor retrieved ranked by similarity to the query embedding z_q.

Weighted Voting Approach. To reduce the effect of possible incorrect retrieved neighbors, we use a weighted voting approach that prioritizes neighbors that exhibit high similarity with the query. Specifically, for the i-th nearest neighbor, let d_i denote the angular distance returned by Annoy. We convert this distance to a weight as $w_i = 1.0 - d_i$, ensuring that neighbors with smaller angular distances (i.e., more similar to the query) contribute more strongly to the final decision, having higher weights. For a given prediction task, let $\ell_i \in \{\ell_1, \ell_2, \ldots, \ell_m\}$ denote the label associated with the i-th nearest neighbor (where m is the number of classes). We accumulate weights across all top-K neighbors belonging to the same label:

$$W_\ell = \sum_{i=1}^{K} w_i \cdot \mathbf{1}[\ell_i = \ell] \tag{2}$$

where $\mathbf{1}[\cdot]$ is the indicator function. The final prediction $\hat{\ell}$ is then determined as the label with the maximum accumulated weight, i.e. $\hat{\ell} = \arg\max_\ell W_\ell$.

Temporal Voting for Sequence Smoothing. Since we process video sequence captured by the user, we leverage temporal voting to smooth predictions across consecutive frames to improve prediction stability and temporal consistency. Specifically, given a sequence of predictions for frames extracted from the video $\{\hat{\ell}_1, \hat{\ell}_2, \ldots, \hat{\ell}_T\}$ (where T is the total number of frames), we use a sliding window of size w and we compute the majority vote within each window, as follows:

$$\tilde{\ell}_t = \text{mode}\{\hat{\ell}_{t-w+1}, \hat{\ell}_{t-w+2}, \ldots, \hat{\ell}_t\} \tag{3}$$

where mode$(\cdot)$ denotes the majority voting operator, returning the most frequently occurring label in the window, thereby providing the final viewpoint prediction. The final visitor localization is finally inferred by assigning the spatial position of the predicted viewpoint. Since we retrieve one viewpoint for each sliding window, the user position is temporally updated accordingly providing a real-time estimation. The hyperparameter w provides us with a trade-off between position estimation frequency and computational requirements. In fig. 5a, we show performance with different values of w.

4 Proof-of-Concept and Experiments

In this section, we report and discuss the results of the experimental evaluation carried out as a proof of concept to assess the effectiveness of the proposed framework to identify the position of visitors in the MAUC museum.

4.1 Implementation Details

As feature extractor, we adopted and compared the performance of four different pre-trained convolutional neural network, namely MobileNetV3 Large [10], Resnet18, Resnet50 and Resnet101 [9]. For the retriever, we build the Annoy index using 300 trees and we set the number of retrieved neighbors $k = 3$ and the Annoy search parameter to 20000. For the temporal smoothing, we use a sliding window $w = 7$ frames. To reduce the domain shift between query (obtained from smartphone cameras) and reference images (obtained from rendered LiDAR scans), we perform data enhancement on query images through brightness and contrast normalization with Matterport statistics. Both query and reference images are resized to 320×320 and 512×512, when using the MobileNetV3 and ResNet architectures, respectively. Note that no enhancement is applied to the reference images, preserving the clean render quality of Matterport scans. Our framework is implemented in Python and evaluated on a computer equipped with 12 GB of RAM, an Intel Core i5-8265U CPU, and an integrated Intel UHD Graphics 620 GPU.

4.2 Evaluation Metrics

As each viewpoint can be treated as a single class, to evaluate the performance of our framework we use the accuracy score. Specifically, we define *Exact Viewpoint*

Table 1: Viewpoint accuracy (%) comparison: exact, tolerance-1, and tolerance-2 configurations across all tested models. Best results in bold.

Model	VP Exact	VP Tol-1	VP Tol-2
ResNet18	78.39	85.43	90.35
ResNet50	69.95	75.98	78.29
ResNet101	72.46	80.60	84.02
MobileNetV3-Large	**88.54**	**91.56**	**93.57**

Accuracy (VP Exact) as the percentage of query frames for which the predicted viewpoint exactly matches the ground-truth viewpoint, with no margin of error allowed. Moreover, to account for spatial ambiguity in viewpoint localization, we introduced a tolerance strategy defined in two levels as follows: (i) *Tolerance-1 (Tol-1)*, where a prediction is correct if it matches the ground-truth viewpoint or its nearest spatial neighbor, and (ii) *Tolerance-2 (Tol-2)*, where the prediction is correct if it falls within the two nearest neighbors. We also evaluated the computational efficiency of our framework, assessed in terms of peak RAM usage (MB) and average inference time per image (seconds).

4.3 Experimental Results

This section presents experimental results obtained on the MAUC museum. The experiments evaluate the effectiveness of the proposed framework in estimating visitor position within the museum, by processing images extracted from user-recorded videos, and the impact of tolerance-based assessment criteria.

Viewpoint Identification. We evaluate our framework on both the exact viewpoint identification and the tolerance-based localization, where predictions are considered correct if they fall within the first or the two nearest neighboring viewpoints. We report results in table 1. As expected, all models exhibit a consistent improvement under tolerance-based evaluation, achieving approximately 92% accuracy in the *Tol-1* setting and 94% in the *Tol-2* setting, compared to the exact localization reaching an accuracy of approximately 89%. While the tolerance-based strategy increases performance, it also reduce spatial precision, as the predicted positions spans multiple viewpoints. To quantitatively evaluate this effect, we leverage the Matterport spatial graph to compute the average distance between viewpoints. As shown in fig. 4, *Tol-1* corresponds to an average spatial uncertainty of 1.317 m ($\pm$0.317 m, range: 0.810–2.096 m), indicating that the nearest neighbor is typically within 1.3 meters, a distance that can still be considered acceptable for indoor navigation. Conversely, *Tol-2* increases this uncertainty to 1.717 m ($\pm$0.374 m, range: 1.101–2.631 m). Therefore, considering both localization accuracy and spatial error, the *Tol-1* configuration is the best trade-off between accuracy and spatial tolerance.

On architectural comparison, table 1 shows MobileNetV3 Large as the best-performing model, achieving approximately 89% accuracy for the exact viewpoint prediction and improving to 92% and 94% under Tol-1 and Tol-2, respectively. These results significantly outperform ResNet-based architectures, which

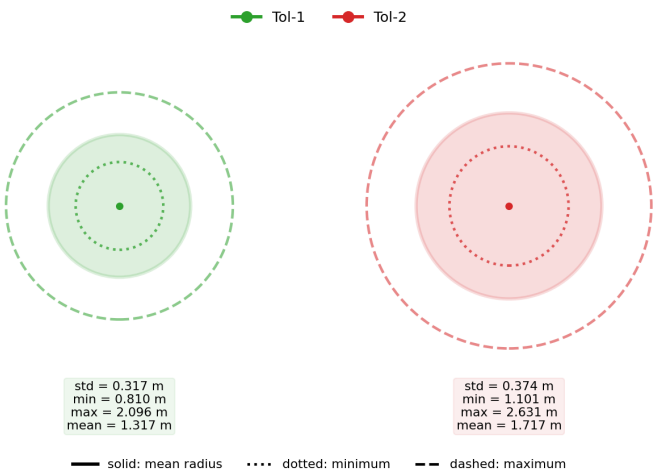


Fig. 4: Metric errors under *Tolerance-1 (Tol-1)* and *Tolerance-2 (Tol-2)* settings.

Table 2: Computational resources requirements: per-frame execution time (sec.) and peak RAM (MB) consumption for each model tested. Best results in bold.

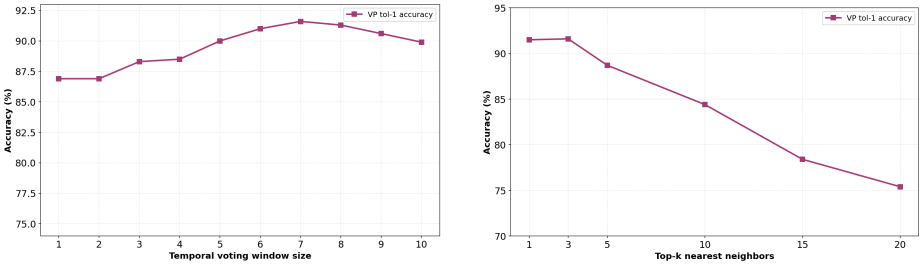
Model	Time	RAM
MobileNetV3-Large	**0.08**	**597.85**
ResNet18	0.24	637.62
ResNet50	0.43	717.68
ResNet101	0.71	771.77

exhibit drops of up to 20% and rarely exceed 90%. Notably, bigger architectures such as ResNet50 and ResNet101 perform worse than smaller ones. This is probably due to larger features vectors that are more affected by domain shift.

Computational Requirements. Our framework can enable future indoor navigation systems on mobile devices, which are limited by computational cost and inference time. Results in table 2 show that MobileNetV3-Large achieves optimal performance, processing each frame in 80 ms while consuming only 598 MB of RAM. This is approximately 3× faster than ResNet18 (244 ms), 6× faster than ResNet50 (443 ms), and 9× faster than ResNet101 (711 ms). At the same time, the lightweight memory footprint ($<$ 800 MB across all models) confirms the framework is suitable for deployment on mobile and edge devices with constrained computational resources.

5 Additional Analysis

Here we present additional analyses on our framework, performed with the MobileNetV3 Large. Specifically, we analyze: (i) temporal voting window size w, (ii)



(a) Temporal voting window size sweep showing accuracy values varying with different sizes.

(b) Top k nearest neighbors sweep showing accuracy values varying the number of neighbors retrieved.

Fig. 5: Ablation studies on temporal voting window size and top-k nearest neighbors for localization performance.

the number of retrieved nearest neighbors k, (iii) the impact of query image enhancement strategies, and (iv) performance with hierarchical-based predictions.
Temporal Voting Window. Figure 5a shows performance with different values for the temporal window size from $w = 1$ (no smoothing) to $w = 10$ frames. Performance shows a general improvement with increasing window size, reflecting the benefits of temporal smoothing. However, they reach best accuracy at window size 7 (91.6%), while reducing performance with higher values.
Top-k Nearest Neighbors. Figure 5b shows performance with different values for the number of retrieved reference images k. Best performance is achieved at $k = 3$, reaching an accuracy of 91.6%. Beyond this point, performance degrades monotonically, with the worst results observed at $k = 20$. This behavior reflects the properties of the weighted voting scheme with angular distance metrics, where the top 3 neighbors provide the most discriminative information, while additional candidates introduce noise and reduce localization accuracy.
Image Enhancement Strategies. To reduce the domain gap between smartphone query frames and Matterport reference scans, we compare 6 sets of image enhancements:

- **E0**: No enhancements applied.
- **E1**: Pixel brightness and contrast normalization to match Matterport statistics.
- **E2**–**E5**: Progressive addition of bilateral denoising, unsharp masking, white balance, and CLAHE.

As shown in table 3, E1 represents the optimal configuration, consistently improving over the baseline E0 with a gain of +3.02% in accuracy. This suggests that simple lighting normalization is sufficient to reduce the domain gap between smartphone captures and Matterport reference scans. Conversely, other enhancements (E2–E5) produce substantial performance degradation, revealing that multi-stage preprocessing degrades the discriminability of CNN features.

Table 3: Viewpoint with 1-neighbor tolerance accuracy (%) when varying the image enhancement configurations. Best result in bold.

Enhancement	VP Tol-1
E0 - No enhancement	88.5
E1 - Brightness/contrast norm	**91.6**
E2 - E1 + Bilateral denoising	78.7
E3 - E2 + Unsharp mask	73.7
E4 - E3 + White balance	72.9
E5 - E4 + CLAHE	67.7

Table 4: Viewpoint (VP) accuracy (%) comparison between Global and Room-conditioned settings under different evaluation criteria. Best results in bold.

Setting	VP Exact	VP Tol-1	VP Tol-2
Global	**88.54**	**91.56**	**93.57**
Room-conditioned	84.32	87.74	89.55

Room-based Hierarchical Prediction. Here, we evaluate the accuracy values obtained using a room-based hierarchical prediction approach where we first predict in which room the query has been acquired and then we perform retrieval only among viewpoints associated with the predicted room. Despite this hierarchical approach reduces the search space and requires rooms human annotations, results in table 4 reveal a performance degradation with accuracy dropping about 4% across the evaluated settings. These results further demonstrate the effectiveness of our framework which obtains high localization performance without requiring any human annotation, making it suitable for real-world applications.

6 Conclusions

In this work, we presented a lightweight, annotation-free vision-based framework for indoor visitor localization in museum environments using frames extracted from smartphone-recorded videos. The proposed approach formulates localization as a content-based image retrieval problem, matching query images against a structured reference database built from Matterport Pro3 LiDAR scans. By directly exploiting the spatial metadata and acquisition graph provided by the scanning system, our method removes the need for manual annotation, addressing key limitations of existing vision-based indoor localization approaches.

Experimental results on the MAUC museum dataset demonstrate that the proposed framework achieves strong localization performance, reaching up to 88.54% exact viewpoint accuracy and 91.56% under tolerance-aware evaluation while operating in real time on lightweight hardware. We also investigated the impact of different CNN architectures and found that lightweight models such as MobileNetV3-Large consistently outperform deeper ResNet-based architectures in both accuracy and efficiency. Overall, the results highlight the effectiveness of

retrieval-based localization as a scalable and practical alternative to traditional infrastructure-based, SLAM-based, and fully supervised learning approaches for indoor environments. Future work will focus on extending the approach to larger and more diverse environments, improving robustness under stronger domain shifts through adaptive or self-supervised feature learning, and integrating the proposed localization framework into end-to-end navigation and visitor assistance systems for cultural heritage applications.

7 Acknowledgement

The research activity was funded by European Union (NextGeneration EU), through the MUR-PNRR project SAMOTHRACE (ECS00000022) "SiciliAn MicronanOTecH Research And Innovation CEnter" - Ecosistema della innovazione (PNRR, Mission 4, Component 2 Investment 1.5, Avviso n. 3277 del 30-12-2021), Spoke 1_Università di Catania - Work Package 6 Cultural Heritage.

We sincerely thank the Scientific Coordinator of the MAUC, Professor S. Todaro, for letting us acquire images of the museum environment.

References

1. Alam, M., Mohamed, F., Hossain, A.K.M.: Integrating recurrent neural networks and loss function optimization for efficient indoor camera positioning. Computer Science Journal of Moldova **33**, 68–90 (2025). https://doi.org/10.56415/csjm.v33.04
2. Bernhardsson, E.: Annoy: Approximate Nearest Neighbors in C++/Python (2018), python package version 1.13.0
3. Cao, S., Lu, X., Shen, S.: Gvins: Tightly coupled gnss–visual–inertial fusion for smooth and consistent state estimation. IEEE Transactions on Robotics **PP**, 1–18 (01 2022). https://doi.org/10.1109/TRO.2021.3133730
4. Cha, K.J., Lee, J.B., Ozger, M., Lee, W.H.: When wireless localization meets artificial intelligence: Basics, challenges, synergies, and prospects. Applied Sciences **13**(23) (2023). https://doi.org/10.3390/app132312734, https://www.mdpi.com/2076-3417/13/23/12734
5. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. pp. 667–676 (10 2017). https://doi.org/10.1109/3DV.2017.00081
6. Che, F., Ahmed, Q.Z., Lazaridis, P.I., Sureephong, P., Alade, T.: Indoor positioning system (ips) using ultra-wide bandwidth (uwb)—for industrial internet of things (iiot). Sensors **23**(12) (2023). https://doi.org/10.3390/s23125710
7. Geok, T., Aung, K., Aung, M., Soe, M., Abdaziz, A., Liew, C., Hossain, F., Tso, C., Wong, H.: Review of indoor positioning: Radio wave technology. Applied Sciences **11**, 279 (12 2020). https://doi.org/10.3390/app11010279
8. Guo, Y., Zheng, J., Di, S., Xiang, G., Guo, F.: A beacons selection method under random interference for indoor positioning. Remote Sensing **14**(17) (2022). https://doi.org/10.3390/rs14174323
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

10. Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., Le, Q.: Searching for MobileNetV3 . In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1314–1324. IEEE Computer Society, Los Alamitos, CA, USA (Nov 2019)
11. Hu, T., Liao, Q.: Real-time camera localization with deep learning and sensor fusion. In: ICC 2021 - IEEE International Conference on Communications. pp. 1–7 (2021). https://doi.org/10.1109/ICC42927.2021.9500770
12. Jia, S., Ma, L., Yang, S., Qin, D.: Semantic and context based image retrieval method using a single image sensor for visual indoor positioning. IEEE Sensors Journal **21**, 18020–18032 (2021)
13. Jiang, W., Cao, Z., Cai, B., Li, B., Wang, J.: Indoor and outdoor seamless positioning method using uwb enhanced multi-sensor tightly-coupled integration. IEEE Transactions on Vehicular Technology (09 2021). https://doi.org/10.1109/TVT.2021.3110325
14. Jing, X., Yang, H., Chen, J., Chen, H., Li, Y., Tang, Z.: Visual indoor localization in complex scenes based on image semantic and spatial features. Journal of Building Engineering **111** (2025). https://doi.org/https://doi.org/10.1016/j.jobe.2025.113172
15. Kwon, O., Jung, D., Kim, Y., Ryu, S., Yeon, S., Oh, S., Lee, D.: Wayil: Image-based indoor localization with wayfinding maps. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 6274–6281 (2024). https://doi.org/10.1109/ICRA57147.2024.10610480
16. Latif, A., Rasheed, A., Sajid, U., Ahmed, J., Ali, N., Ratyal, N.I., Zafar, B., Dar, S.H., Sajid, M., Khalil, T.: Content-based image retrieval and feature extraction: A comprehensive review. Mathematical Problems in Engineering **2019**(1) (2019)
17. Lee, K., Haeyun, L., Hwang, J.: Son: Selective optimal network for smartphone-based indoor localization in real-time. Expert Systems with Applications **272** (2025). https://doi.org/10.1016/j.eswa.2025.126639
18. Liu, T., Li, B., Chen, G., Yang, L., Jing, Q., Chen, W.: Tightly coupled integration of gnss/uwb/vio for reliable and seamless positioning. IEEE Transactions on Intelligent Transportation Systems **PP**, 1–13 (01 2023). https://doi.org/10.1109/TITS.2023.3314836
19. Liu, X., Huang, H., Hu, B.: Indoor visual positioning method based on image features. Sensors and Materials (2022)
20. Liu, Y., Xie, K., Huang, H.: Vgf-net: Visual-geometric fusion learning for simultaneous drone navigation and height mapping. Graphical Models **116** (2021). https://doi.org/10.1016/j.gmod.2021.101108
21. Mauro, D., Furnari, A., Signorello, G., Farinella, G.: Unsupervised domain adaptation for 6dof indoor localization. pp. 954–961 (01 2021). https://doi.org/10.5220/0010333409540961
22. Merry, K., Bettinger, P.: Smartphone gps accuracy study in an urban environment. PLOS ONE **14**(7), 1–19 (07 2019). https://doi.org/10.1371/journal.pone.0219890
23. Patruno, C., Colella, R., Nitti, M., Renò, V., Mosca, N., Stella, E.: A vision-based odometer for localization of omnidirectional indoor robots. Sensors **20**(3) (2020). https://doi.org/10.3390/s20030875
24. Şentürk, M., Demiral, E., Sevinç, H.K., Karaş, İ., Ujang, U.: Image-based indoor positioning: Current status and challenges. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XLVIII-4/W19-2025**, 119–126 (03 2026). https://doi.org/10.5194/isprs-archives-XLVIII-4-W19-2025-119-2026

25. Shuai, Z., Yu, H.: Multi-sensor fusion for autonomous positioning of indoor robots. In: Proceedings of the International Conference on Computer and Communication Systems (ICCCS). pp. 105–112 (2021)
26. Sun, X., Zhao, Y., Wang, Y., Li, Z., He, Z., Wang, X.: Upl-slam: Unconstrained rgb-d slam with accurate point-line features for visual perception. IEEE Access **PP** (2024). https://doi.org/10.1109/ACCESS.2024.3524465
27. Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A.: Inloc: Indoor visual localization with dense matching and view synthesis. pp. 7199–7209 (06 2018). https://doi.org/10.1109/CVPR.2018.00752
28. Uygur, I., Miyagusuku, R., Pathak, S., Moro, A., Yamashita, A., Asama, H.: Robust and efficient indoor localization using sparse semantic information from a spherical camera. Sensors **20**(15) (2020). https://doi.org/10.3390/s20154128
29. Wang, C., Bi, K., Zhao, B., Li, M., Chen, Y., Tao, S., Yang, J.: Lightweight indoor positioning system based on multiple self-learning features and key frame classification. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences pp. 373–379 (2024). https://doi.org/10.5194/isprs-annals-X-4-2024-373-2024
30. Wang, Y., Pan, H., Qiu, L., Zhong, L., Liu, J., Ma, R., Chen, Y.C., Xue, G., Ren, J.: Gpms: Enabling indoor gnss positioning using passive metasurfaces. pp. 1424–1438 (12 2024). https://doi.org/10.1145/3636534.3690702
31. Xie, T., Dai, K., Wang, k., Li, R., Wang, J., Tang, X., Zhao, L.: A deep feature aggregation network for accurate indoor camera localization. IEEE Robotics and Automation Letters **7** (2022). https://doi.org/10.1109/LRA.2022.3146946
32. Yang, C., Chen, Q., Yang, Y., Zhang, J., Wu, M., Mei, K.: Sdf-slam: A deep learning based highly accurate slam using monocular camera aiming at indoor map reconstruction with semantic and depth fusion. IEEE Access **10** (2022). https://doi.org/10.1109/ACCESS.2022.3144845
33. Zhang, T., Liu, C., Li, J., Pang, M., Wang, M.: A new visual inertial simultaneous localization and mapping (slam) algorithm based on point and line features. Drones **6** (2022). https://doi.org/10.3390/drones6010023
34. Zhang, Y., Leonard, J.J.: A front-end for dense monocular slam using a learned outlier mask prior. 2021 IEEE International Conference on Robotics and Automation (ICRA) pp. 11732–11738 (2021)
35. Zhou, B., Xiao, Y., Li, Q., Sun, C., Wang, B., Pan, L., Zhang, D., Zhu, J., Li, Q.: Darkloc+: Thermal image-based indoor localization for dark environments with relative geometry constraints. IEEE Transactions on Geoscience and Remote Sensing **PP** (2024). https://doi.org/10.1109/TGRS.2024.3350043
36. Zhou, T., Ku, J., Lian, B., Zhang, Y.: Indoor positioning algorithm based on improved convolutional neural network. Neural Comput. Appl. **34**(9), 6787–6798 (2022). https://doi.org/10.1007/s00521-021-06112-5
37. Zhou, X., Meng, B., Dong, Y., Huang, X., Zhang, K.: An efficient image-based indoor positioning approach using orb and lsh. pp. 2985–2989 (10 2021). https://doi.org/10.1109/CAC53003.2021.9727606
38. Zhuang, Y., Zhang, C., Huai, J., Li, Y., Chen, L., Chen, R.: Bluetooth localization technology: Principles, applications, and future trends. IEEE Internet of Things Journal **9**(23), 23506–23524 (2022). https://doi.org/10.1109/JIOT.2022.3203414