International Journal of Innovative Research and Scientific Studies, 9(1) 2026, pages: 83-90

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Multi-model fusion and re-ranking for spatial image retrieval

Tran Duc Long[1*], Tran Minh Hien[1], Nguyen Duc Thanh[1], Duong Anh Tra[1]

[1]*Viettel High Technology Industries Corporation, Viettel Group, Hanoi, Vietnam.*

Corresponding author: Tran Duc Long (*Email: longtd12@viettel.com.vn*)

Abstract

Image retrieval has become a central component of large-scale visual understanding systems, particularly as real-world datasets grow in volume, diversity, and semantic complexity, and numerous methods have been proposed to improve retrieval accuracy across diverse scenarios [1]. However, the performance of individual models often varies significantly depending on the characteristics of real-world datasets, making it challenging for a single technique to consistently achieve robust results. To address this limitation, we introduce a fusion-based retrieval framework that leverages the complementary strengths of three state-of-the-art models: SALAD [2] and CliqueMining [3] MegaLoc [4]. Each model independently generates an initial ranked list, capturing different visual cues and retrieval patterns. To further enhance reliability and reduce model-specific biases, we apply a re-ranking stage using the Distribution-based Score Fusion method [5] an aggregation technique designed to normalize heterogeneous score distributions and emphasize consistent cross-model evidence. Our proposed approach provides a unified and efficient strategy for improving retrieval accuracy without requiring additional training or architectural modifications. Experimental evaluations demonstrate that the combined system consistently outperforms individual models, offering improved robustness and more stable performance across varying image domains.

Keywords: Image retrieval, Multi-model fusion, Re-ranking aggregation.

DOI: 10.53894/ijirss.v9i1.11168
Funding: This study received no specific financial support.
History: Received: 19 November 2025 / **Revised:** 25 December 2025 / **Accepted:** 31 December 2025 / **Published:** 14 January 2026
Copyright: © 2026 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Institutional Review Board Statement: Acknowledgment: The authors sincerely appreciate the support and valuable resources provided by the Modeling and Simulation Center, a division of Viettel High Technology Industries Corporation - Viettel Group, during this study. Additionally, we deeply acknowledge the leadership team, and team members at the Modeling and Simulation Center for their dedicated assistance, insightful feedback, and significant contributions to the testing and refinement of the solution examined in this project.
Publisher: Innovative Research Publishing

1. Introduction

Image retrieval is a well-established problem in computer vision, focused on identifying and ranking images that are visually or semantically relevant from large-scale collections [1]. As digital image repositories continue to grow across diverse domains such as autonomous driving, robotics, remote sensing, cultural heritage preservation, and consumer-level visual search, there is an increasing demand for retrieval systems that remain robust under varying environmental, structural, and semantic conditions [6]. Unlike conventional classification tasks, image retrieval requires capturing fine-grained relationships between images, where similarity may emerge from local structures, global layouts, or high-level semantic cues [7]. This complexity makes the problem inherently challenging and has motivated ongoing research into more expressive feature representations and more discriminative ranking strategies.

Early deep-learning-based image retrieval systems drew heavily on handcrafted descriptor pipelines such as SIFT [8] Bag-of-Words (BoW) [9] and Fisher Vectors [10]. These classical methods leveraged geometric verification, local feature correspondences, and spatial consistency checks to improve robustness, yet they often struggled under significant viewpoint changes or severe variations in illumination. The advent of convolutional neural networks (CNNs) transformed the field by enabling the learning of image representations that surpassed manually engineered features. Within this shift, aggregation-based descriptors - particularly those inspired by the VLAD [11] framework - emerged as a dominant approach. NetVLAD [12] for example, demonstrated that differentiable cluster centers combined with aggregated local CNN features could produce compact, discriminative descriptors suitable for large-scale retrieval, spurring a range of extensions and variants.

CosPlace [13] reformulates visual place recognition as a large-scale classification problem by grouping images into geographically coherent Cos-Groups and training with a margin-based cosine loss. This eliminates the need for pair or triplet mining, while a lightweight classifier attached during training is removed at inference, producing compact descriptors optimized for retrieval. Compared to traditional metric-learning pipelines, CosPlace reduces memory and training costs while maintaining strong retrieval performance. Its geographic grouping implicitly preserves spatial relationships, offering a streamlined yet effective alternative to VLAD-based or graph-dependent methods.

Despite the efficiency and strong performance of CNN-based systems such as NetVLAD and CosPlace, their reliance on convolutional receptive fields limits their capacity to model long-range interactions and capture global scene structures - capabilities that are increasingly critical for modern retrieval and localization tasks. These limitations have driven the shift toward transformer-based architectures, which naturally encode global context through self-attention mechanisms and overcome the locality bias inherent to CNNs. VLAD-BuFF [14] integrates VLAD-style aggregation with transformers, achieving high accuracy on domain-specific datasets. It requires explicit training on target-domain data, which can prolong real-world deployment due to the need for data collection, labeling, and retraining, and may limit generalization to unseen environments.

Transformer-based models have significantly advanced self-supervised retrieval tasks. SALAD [2] for instance, introduces a novel architecture that encourages consistency across augmented views, extracting discriminative features while preserving local-to-global correspondences. This enables the model to generalize well without relying on labeled data, improving robustness to domain shifts. CliqueMining [3] extends this idea by incorporating a relational mining strategy, where attention weights represent inter-patch and inter-image relationships. By constructing a graph of visually and geographically connected images, it forms clusters (cliques) that preserve fine-grained spatial distinctions and semantic coherence during training, without adding extra cost at inference time. MegaLoc [4] integrates multi-scale attention maps with structural reasoning modules, capturing both fine-grained details and broader contextual patterns. This improves geo-localization and retrieval performance, addressing the limitations of CNN-based aggregation methods and offering enhanced robustness across diverse environments.

In recent years, a number of works have explored **fusion** and **re-ranking** strategies in image retrieval to combine complementary strengths of different feature representations. Yang, et al. [15] proposed multi-feature fusion using diffusion on graphs constructed from different feature similarities, where graph weights are learned in a data-driven way without supervision Yang, et al. [15]. Yang, et al. [16] introduced DOLG, an end-to-end framework fusing orthogonal local and global features into a single compact descriptor [16]. Adaptive-weight fusion methods have also been proposed, in which unsupervised or supervised weights are assigned to each feature type depending on query-specific retrieval performance [17]. More recent works, such as Multi-FusNet, fuse features at multiple depths via self-attentive hashing coding for fine-grained retrieval [18]. These previous fusion strategies demonstrate the benefits of integrating heterogeneous representations to refine rankings or build richer embeddings.

However, most of these approaches are designed as end-to-end fusion architectures, which require joint training and significant computational resources, and are often tightly coupled to specific backbone designs. This limits their flexibility when integrating independently trained state-of-the-art retrieval models. In contrast, score-level fusion provides a more modular and lightweight alternative, enabling the combination of multiple heterogeneous models without re-training. Although existing schemes such as Reciprocal Rank Fusion [19] and Relative Score Fusion [20] have demonstrated effectiveness in this setting, they largely rely on heuristic weighting or rank-level operations that ignore the underlying score distributions. To address this limitation, we adopt Distribution-Based Score Fusion [5] which explicitly models score distributions to achieve a more principled normalization and alignment across different retrieval outputs, resulting in more stable and robust retrieval performance.

Building on this line of research, our proposed framework goes further by combining three state-of-the-art retrieval models - MegaLoc, SALAD, and CliqueMining - and employing Distribution-Based Score Fusion [5]. This approach

leverages the diversity of the individual models while mitigating their respective weaknesses, producing more robust and accurate image retrieval results.

2. Materials and Methods

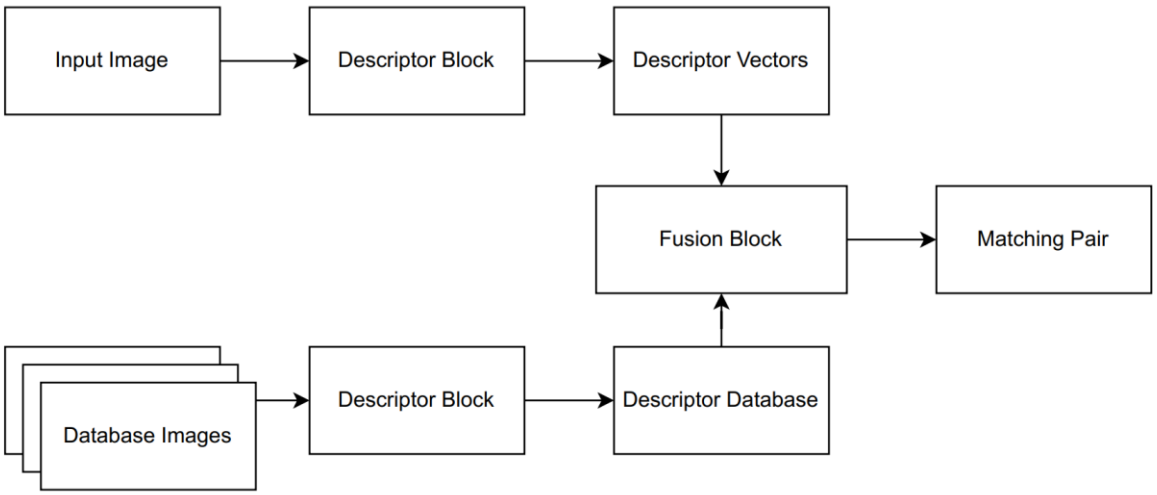

Figure 1.
Overview Proposal Image Retrieval Pipeline.

Our pipeline consists of two main components: (1) the Descriptor Block and (2) the Fusion Block, which is presented in Figure 1. Features are extracted from the input images through the Descriptor Block, and the resulting descriptor vectors are queried against a descriptor vector database. These are then combined within the Fusion Block to select the best image pair. The Descriptor Block is responsible for capturing complementary visual cues from each image, producing robust embeddings suitable for retrieval. Meanwhile, the Fusion Block integrates the outputs of multiple models or feature types, enhancing the overall retrieval accuracy by leveraging their diverse strengths.

2.1. Descriptor Block

In this work, we propose a unified Descriptor Block architecture that integrates three state-of-the-art visual retrieval models – SALAD [2], CliqueMining [3], and MegaLoc [4] – within a single computational pipeline. The design introduces a partially shared Vision Transformer (ViT) backbone and a multi-branch specialization mechanism that reduces redundant computation while preserving the distinctive strengths of each model. The overall structure of the proposed architecture is illustrated in Figure 2.

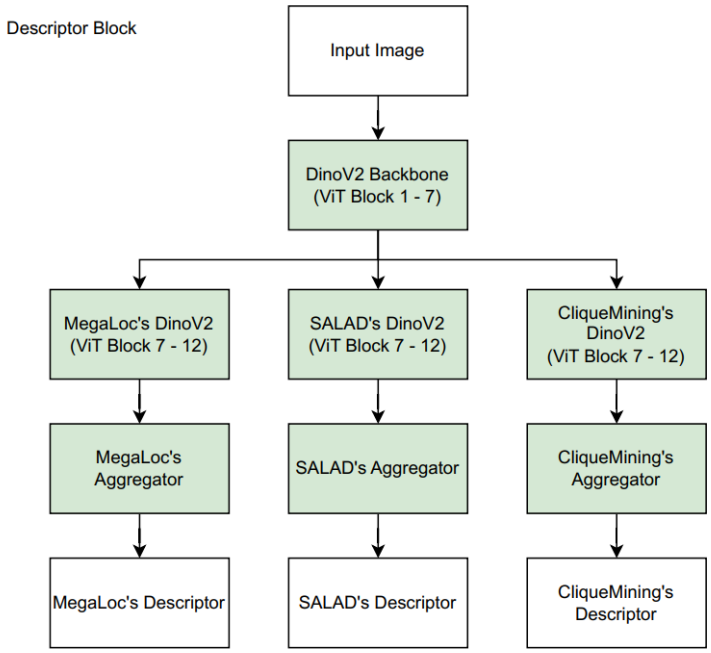

Figure 2.
Descriptor Block Pipeline.

2.2. Shared Backbone

Our architecture begins by forwarding the input image through a shared DINOv2 backbone [21] consisting of ViT Blocks 1 – 7. This shared component acts as a universal feature extractor that captures low- and mid-level visual cues, including texture patterns, local geometric structures, and coarse semantic information. Because these early layers are typically the most computationally expensive within transformer-based models, sharing them across all three branches significantly improves efficiency. Instead of executing three complete models independently, our approach performs the largest portion of the computation only once, enabling scalable multi-model fusion at inference time.

2.3. Model Specific Descriptor

Following the shared backbone, the pipeline splits into three branches, each corresponding to one of the state-of-the-art retrieval models. Although all branches operate on the same token sequence produced by Blocks 1 – 7, each branch maintains its own DINOv2 Blocks 7 – 12 and aggregator, preserving the distinctive characteristics and inductive biases of the individual models. This design allows the framework to retain the inherent strengths of each model while enabling parallel extraction of complementary visual representations.

The three branches - MegaLoc, SALAD, and CliqueMining - process the shared features independently to generate their respective descriptors. MegaLoc focuses on geometric stability and spatial coherence, SALAD emphasizes robustness against noise and illumination variations, and CliqueMining captures high-order relational patterns among image patches. By maintaining these heterogeneous representations, the framework lays the foundation for combining the strengths of all three models, aiming to produce a more robust and discriminative descriptor for image retrieval without assuming that any single branch handles a unique type of information.

2.4. Fustion Block

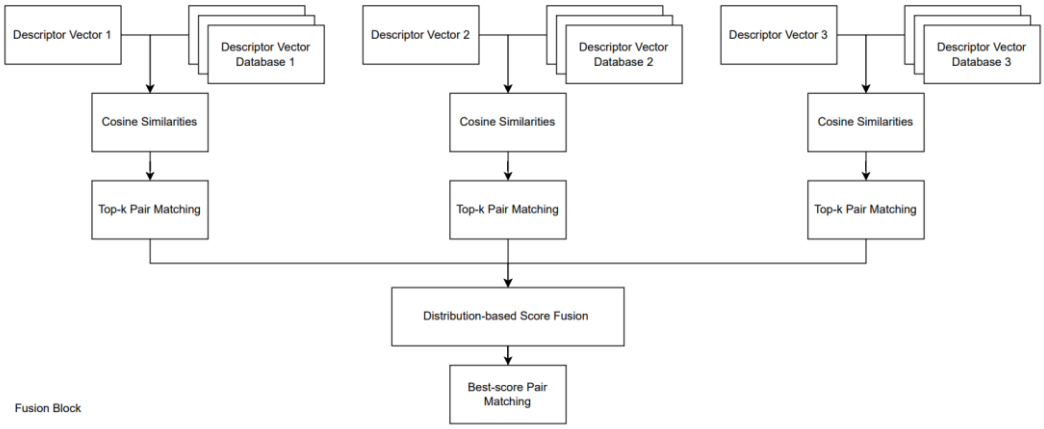

Figure 3.
Fusion Block Pipeline.

Figure 3 illustrates the fusion step of our method. Each descriptor vector is first compared against its corresponding database of descriptor vectors using cosine similarity to measure the similarity between the query and database features. For each model, the top-k most similar pairs are selected through a pair matching step. The top-k matches from all three descriptor models are then combined in a distribution-based score fusion step, which integrates the individual similarity scores to leverage complementary strengths of each model. Finally, the fused scores are used to identify the best-scoring pair matches, resulting in a more robust and accurate retrieval outcome.

2.5. Cosine Similarity Measure

In each model, the Cosine similarity measure is used to assess the likeness between the query vector q_n and all descriptor vectors $d_i \in D_n$. Cosine similarity is highly effective in high-dimensional feature spaces as it focuses on the orientation (angle) of the vectors, remaining unaffected by their magnitude. The similarity score $S_{n,i}$ for channel n is calculated as:

$$S_{n,i} = \frac{q_n \cdot d_i}{\|q_n\| \|d_i\|} \tag{1}$$

After computing $S_{n,i}$ for all vectors in the database, selects the k data pairs with the highest raw similarity scores and passes their corresponding scores ($x_{n,k}$) to the normalization stage.

2.6. Distribution-Based Score Normalization

The raw scores $x_{n,k}$ from each model are normalized independently before fusion. This method first requires transforming the raw scores $x_{n,k}$ into a unified, normalized space using a statistical min-max approach. This transformation, carried out per channel, is bounded by a statistically significant range defined by the channel's estimated

mean μ_n, standard deviation σ_n of the genuine distribution, and a statistical scaling factor α. The normalized score $S'_{n,k}$ for a raw score $x_{n,k}$ from descriptor n is calculated as:

$$S'_{n,k} = \frac{x_{n,k} - (\mu_i - \alpha\sigma_i)}{(\mu_i + \alpha\sigma_i) - (\mu_i - \alpha\sigma_i)} \quad (2)$$

2.7. Final Matching Decision

After normalization, the top-k lists from all channels (now containing normalized scores $S'_{n,k}$) are aggregated using the *Max Aggregation* technique. The final fusion score $S_{final}(d_j)$ for a candidate vector d_j that appears in one or more top-k lists is defined as the summation of its corresponding normalized scores $S'_{n,j}$ from all channels $n \in$ E = (1, 2, 3) in which it was present:

$S_{final}(d_j) = \sum_{n \in E} w_n * S'_{n,j}$ where $\sum_{n \in E} w_i = 1$ (3)

If d_j does not appear in the top-k list of channel n, $S'_{n,j}$ is considered zero for that channel. This strategy effectively prioritizes candidates that exhibit high similarity across multiple, statistically normalized descriptor views.

The fused score S_{final} represents the overall confidence of each candidate in the merged set C. The final matching decision is the *Best-score Pair Matching*, where the system selects the single candidate d_j that yields the maximum final fused score:

$$d_{best} = \max_{d_j \in C} S_{final}(d_j) \quad (4)$$

This method ensures that the final selection is the candidate with the strongest aggregated evidence across all parallel channels.

3. Results and Discussion

In this section, we describe the datasets used in our experiments. Our approach is denoted as *F(a, b, c)*, where *a*, *b*, and *c* indicate the contribution weights of SALAD, CliqueMining, and MegaLoc, respectively. The statistical scaling factor is fixed at α = 3.0, as it provides a good balance between score discrimination and numerical stability based on preliminary experiments. Additionally, we set the number of top retrieved candidates to k = 10 in order to ensure sufficient candidate diversity while maintaining reasonable computational efficiency during the re-ranking stage. We evaluate the proposed method against the three individual baselines - MegaLoc, CliqueMining, and SALAD - across all datasets.

3.1. Benchmark Dataset

3.1.1. Amster Time Dataset

AmsterTime Dataset [22] is a visual place recognition dataset containing 2,500 curated images from Amsterdam, pairing historical archival images with modern street-view images (Mapillary). The image pairs capture the same location with different cameras, viewpoints, and appearances. Evaluation includes verification and retrieval tasks, where ResNet-101 pre-trained on the Landmarks dataset achieves 84\% accuracy for verification and 24\% mAP for retrieval. A subset of images with landmark labels is also provided for classification and visual explanation tasks.

3.1.2. SPED Dataset

SPED Dataset [23] is a large-scale place recognition dataset constructed from the AMOS archive containing images from approximately 30,000 outdoor cameras worldwide. From these, 2,543 cameras were selected, and all images captured in February 2014 and August 2014 were downloaded, resulting in about 2.5 million images. Each camera provides images taken every 30 minutes across two months with strong seasonal differences, allowing evaluation under long-term variations such as day–night cycles, lighting changes, and seasonal appearance shifts. The dataset covers diverse outdoor environments, including forests, rural roads, and urban scenes.

3.1.3. MSLS Dataset

MSLS Dataset [24] is a large-scale place recognition dataset containing more than 1.6 million images collected from the Mapillary platform. The data spans 30 major cities across six continents, captured by hundreds of different cameras over a nine-year period. The sequences exhibit substantial variations in viewpoint, appearance, season, and capture time. All images are geo-located with GPS and compass information and include high-level attributes such as road type, making the dataset suitable for lifelong and large-scale place recognition benchmarks.

3.1.4. Nordland Dataset

Nordland Dataset [25] is a seasonal place recognition dataset consisting of video recordings of a 728 km train journey between Trondheim and Bodø in Norway. The full 10-hour journey was recorded four times, once in each season, providing four traversals of the same route with dramatic appearance changes. Videos were captured at 25 fps in 1920×1080 resolution using a Sony XDcam with a Canon HJ15ex8.5B KRSE-V lens. GPS readings were recorded at 1 Hz and time-synchronized to the video, enabling frame-accurate ground truth across all seasons. The dataset covers natural landscapes, varying weather conditions, and occasional urban segments.

3.1.5. Pittsburgh Dataset

Pittsburgh Dataset [26] is constructed from 10,586 Google Street View panoramas of the Pittsburgh area, from which 254,064 perspective images (640×480) are generated. Each panorama (6656×3328) is converted into 24 perspective views using 2 yaw and 12 pitch directions. For evaluation, 24,000 query images are generated from 1,000 panoramas taken in a different capture session, providing challenging variations in viewpoint, illumination, and season. All query images have ground-truth GPS positions, enabling precise recall-based place recognition evaluation.

3.1.6. Tokyo247 Dataset

Tokyo247 Dataset [27] is a large-scale place recognition dataset captured in Tokyo with 1,125 query images taken by Apple iPhone5s and Sony Xperia at 125 distinct locations. At each location, images were captured from 3 different viewing directions and at 3 times of day. Ground truth GPS coordinates were manually annotated with an estimated error below 5 meters. For evaluation, a subset of 315 query images within an area of approximately 1,600m × 1,600m is used.

3.2. Experimental

3.2.1. Evaluation Metrics

We use Recall@K (K ∈ (1, 5, 10)) as our evaluation metric, considering VPR [28] as a coarse initial retrieval step to facilitate the application of more precise metric localization and mapping techniques. For all datasets, ground-truth reference images for each query are defined within a 25 m localization radius. During inference, we resize images to 224×224 for the CliqueMining branch and to 322×322 for MegaLoc and SALAD, following the original configurations used by these models.

3.2.2. Quantitative Evaluation

Across the three evaluation settings - R@1, R@5, and R@10 - the performance comparisons reveal several important trends regarding the strengths and limitations of individual models and the significant benefits introduced by the fusion approach.

First, the standalone models (MegaLoc, SALAD, and CliqueMining) each demonstrate strong performance on specific datasets, but their behavior varies notably across conditions. MegaLoc generally performs best on structured urban datasets such as Pitts250k, Pitts30k, and Tokyo247, where its global descriptors provide high discrimination. SALAD excels on SPED, likely due to its robustness to viewpoint and appearance changes. CliqueMining shows competitive performance on MSLS and Nordland, benefiting from its mining-based feature refinement. However, none of these models consistently dominate across all datasets. Each exhibits weaknesses: SALAD underperforms on AmsterTime, CliqueMining drops significantly on SPED, and MegaLoc occasionally trails the others on more challenging cross-condition environments such as MSLS.

In contrast, the fusion models - F(0.2, 0.3, 0.5) and F(0.2, 0.4, 0.4) - show a more balanced and stable performance profile across all benchmarks. For example, on AmsterTime, the fusion variants achieve the highest scores among all methods. On Nordland and SPED, the fusion models again match or exceed the strongest single-model results, especially in Nordland. This robustness is especially notable on datasets that the individual child models were not trained on: in these cases, the fusion models consistently outperform their standalone counterparts, demonstrating better generalization and resilience to domain shifts.

Conversely, on datasets where the child models were trained, we explicitly tailor the fusion weights based on each model's relative performance - assigning higher weights to the strongest model and proportionally smaller weights to the others. This adaptive weighting strategy allows the fusion variants to retain the strengths of the best-performing model while still benefiting from complementary signals from the remaining ones. The resulting consistency is further emphasized at higher recall levels (R@10), where the fusion models frequently deliver the best or second-best performance overall.

More importantly, the qualitative implication is that the fusion strategy does not merely average the strengths of the constituent models - it actively compensates for their weaknesses. Distribution-based Score Normalization plays a central role here. By adaptively weighting each model based on their complementary characteristics, Distribution-based Score Normalization enables the fused system to recover correct matches that individual models fail to detect. This is most evident in datasets with severe visual changes (seasonal shifts, day-night variations, motion blur), where single descriptors may lose discriminative power. Distribution-based Score Normalization leverages multiple feature sources and blends them in a way that maximizes agreement where possible while allowing a stronger model to dominate when others are uncertain. As a result, the fusion system exhibits an ability to "correct" errors from its constituent models, producing predictions that none of the individual methods could achieve on their own.

Table 1.
Comparisons of various methods on popular datasets - R@1.

Dataset	AmsterTime	SPED	MSLS	Nordland	Pitts250k	Pitts30k	Tokyo247
MegaLoc	63.0	89.8	90.9	94.2	96.3	94.0	97.1
SALAD	85.2	91.8	88.1	86.0	94.9	92.2	94.0
CliqueMining	52.4	87.6	91.0	92.4	95.1	96.6	94.3
F(0.2,0.3,0.5)	63.9	91.8	92.0	97.0	96.6	94.2	97.8
F(0.2,0.4,0.4)	63.4	92.1	92.0	97.0	96.5	94.0	96.8

Table 2.
Comparisons of various methods on popular datasets - R@5.

Dataset	AmsterTime	SPED	MSLS	Nordland	Pitts250k	Pitts30k	Tokyo247
MegaLoc	84.2	95.1	94.9	97.8	98.9	97.4	99.7
SALAD	78.6	96.2	93.8	93.5	98.4	96.2	97.5
CliqueMining	75.3	94.5	94.5	97.1	98.6	96.6	98.4
F(0.2,0.3,0.5)	85.9	96.5	95.2	98.9	99.0	97.4	99.4
F(0.2,0.4,0.4)	84.6	96.9	95.2	98.9	99.0	97.3	99.0

Table 3.
Comparisons of various methods on popular datasets - R@10.

Dataset	AmsterTime	SPED	MSLS	Nordland	Pitts250k	Pitts30k	Tokyo247
MegaLoc	88.5	96.2	95.8	98.7	99.3	98.3	99.7
SALAD	83.8	96.7	95.1	95.7	99.1	97.4	97.5
CliqueMining	80.4	95.9	95.4	98.3	99.3	97.7	98.4
F(0.2,0.3,0.5)	89.3	97.2	96.0	99.4	99.4	98.3	99.4
F(0.2,0.4,0.4)	89.0	97.0	96.0	90.4	99.4	98.3	99.4

4. Conclusion

In this work, we introduced a unified multi-model fusion framework that leverages the complementary strengths of MegaLoc, SALAD, and CliqueMining to achieve more robust and accurate image retrieval. By incorporating a partially shared DINOv2 backbone and a modular Descriptor Block, our system efficiently extracts diverse feature representations without incurring the high computational cost of running three independent models. The proposed Distribution-based Score Fusion method further enhances reliability by normalizing heterogeneous similarity distributions and emphasizing candidates consistently supported across models.

Extensive experiments on a wide range of challenging benchmarks - including AmsterTime, SPED, MSLS, Nordland, Pittsburgh, and Tokyo247 - demonstrate that the fused system consistently outperforms individual state-of-the-art approaches. These improvements highlight the value of integrating diverse visual cues and model-specific inductive biases within a unified framework.

Overall, our results suggest that multi-model fusion is a practical and scalable strategy for enhancing retrieval performance, especially in scenarios characterized by substantial variations in viewpoint, illumination, and environmental conditions. Future work may explore dynamic weighting strategies, end-to-end trainable fusion mechanisms, and further optimization of shared computation to push the boundaries of large-scale visual retrieval.

References

[1] W. Chen *et al.*, "Deep learning for instance retrieval: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7270-7292, 2022. https://doi.org/10.1109/TPAMI.2022.3218591

[2] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in *Proceedings of the Ieee/cvf Conference on Computer vision and Pattern Recognition*, 2024, pp. 17658-17668.

[3] S. Izquierdo and J. Civera, "Close, but not there: Boosting geographic distance sensitivity in visual place recognition," presented at the European Conference on Computer Vision (pp. 240-257). Cham: Springer Nature Switzerland, 2024.

[4] G. Berton and C. Masone, "Megaloc: One retrieval to place them all," in *Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 2861-2867)*, 2025.

[5] D. Kim, B. Kim, D. Han, and M. Eibich, "Autorag: Automated framework for optimization of retrieval augmented generation pipeline," *arXiv preprint arXiv:2410.20878*, 2024. https://doi.org/10.48550/arXiv.2410.20878

[6] S. Lowry *et al.*, "Visual place recognition: A survey," *Ieee Transactions on Robotics*, vol. 32, no. 1, pp. 1-19, 2015. https://doi.org/10.1109/TRO.2015.2496823

[7] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262-282, 2007. https://doi.org/10.1016/j.patcog.2006.04.045

[8] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1224-1244, 2017. https://doi.org/10.1109/TPAMI.2017.2709749

[9] F. Radenović, G. Tolias, and O. Chum, "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples," presented at the European Conference on Computer Vision (pp. 3-20). Cham: Springer International Publishing, 2016.

[10] V. Chandrasekhar, J. Lin, O. Morère, H. Goh, and A. Veillard, "A practical guide to cnns and fisher vectors for image instance retrieval," *arXiv preprint arXiv:1508.02496*, 2015. https://doi.org/10.48550/arXiv.1508.02496

[11] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," presented at the In 2010 IEEE Computer society Conference on Computer Vision and Pattern Recognition (pp. 3304-3311). IEEE, 2010.

[12] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5297-5307)*, 2016.

[13] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4878-4888)*, 2022.

[14] A. Khaliq, M. Xu, S. Hausler, M. Milford, and S. Garg, "VLAD-BuFF: Burst-aware fast feature aggregation for visual place recognition," presented at the European Conference on Computer Vision (pp. 447-466). Cham: Springer Nature Switzerland, 2024.

[15] F. Yang, B. Matei, and L. S. Davis, "Re-ranking by multi-feature fusion with diffusion for image retrieval," in *2015 IEEE Winter Conference on Applications of Computer Vision (pp. 572-579). IEEE*, 2015.

[16] M. Yang *et al.*, "Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features," in *Proceedings of the IEEE/CVF International conference on Computer Vision*, 2021, pp. 11772-11781.

[17] X. Lu, J. Wang, X. Li, M. Yang, and X. Zhang, "An adaptive weight method for image retrieval based multi-feature fusion," *Entropy*, vol. 20, no. 8, p. 577, 2018. https://doi.org/10.3390/e20080577

[18] X. Cui, H. Li, L. Liu, S. Wang, and F. Xu, "Multi-FusNet: fusion mapping of features for fine-grained image retrieval networks," *PeerJ Computer Science*, vol. 10, p. e2025, 2024. https://doi.org/10.7717/peerj-cs.2025

[19] G. V. Cormack, C. L. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 758-759.

[20] S. Bruch, S. Gai, and A. Ingber, "An analysis of fusion functions for hybrid retrieval," *ACM Transactions on Information Systems*, vol. 42, no. 1, pp. 1-35, 2023.

[21] M. Oquab *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[22] B. Yildiz, S. Khademi, R. M. Siebes, and J. van Gemert, "AmsterTime: A visual place recognition benchmark dataset for severe domain shift," *arXiv. arXiv:2203.16291*, 2022.

[23] Z. Chen *et al.*, "Deep learning features at scale for visual place recognition," in *2017 IEEE international conference on robotics and automation (ICRA) (pp. 3223-3230). IEEE*, 2017.

[24] F. Warburg, S. Hauberg, M. Lopez-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[25] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons," in *Proc. of workshop on long-term autonomy, IEEE international conference on robotics and automation (ICRA) (p. 2013). Citeseer*, 2013.

[26] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013.

[27] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[28] S. Schubert, P. Neubert, S. Garg, M. Milford, and T. Fischer, "Visual place recognition: A tutorial [tutorial]," *IEEE Robotics & Automation Magazine*, vol. 31, no. 3, pp. 139-153, 2023. https://doi.org/10.1109/MRA.2023.3310859