Global-to-Local or Local-to-Global? Enhancing Image Retrieval with Efficient Local Search and Effective Global Re-ranking

Anonymous ICCV submission

Paper ID *****

Abstract

The dominant paradigm in image retrieval systems today is to search large databases using global image features, and re-rank those initial results with local image feature matching techniques. This design, dubbed global-to-local*, stems from the computational cost of local matching approaches, which can only be afforded for a small number of retrieved images. However, emerging efficient local feature search approaches have opened up new possibilities, in particular enabling detailed retrieval at large scale, to find partial matches which are often missed by global feature search. In parallel, global feature-based re-ranking has shown promising results with high computational efficiency. In this work, we leverage these building blocks to introduce a* local-to-global *retrieval paradigm, where efficient local feature search meets effective global feature re-ranking. Critically, we propose a re-ranking method where global features are computed on-the-fly, based on the local feature retrieval similarities. Such re-ranking-only global features leverage multidimensional scaling techniques to create embeddings which respect the local similarities obtained during search, enabling a significant re-ranking boost. Experimentally, we demonstrate unprecedented retrieval performance on the Revisited Oxford and Paris datasets, setting new state-of-the-art results.*

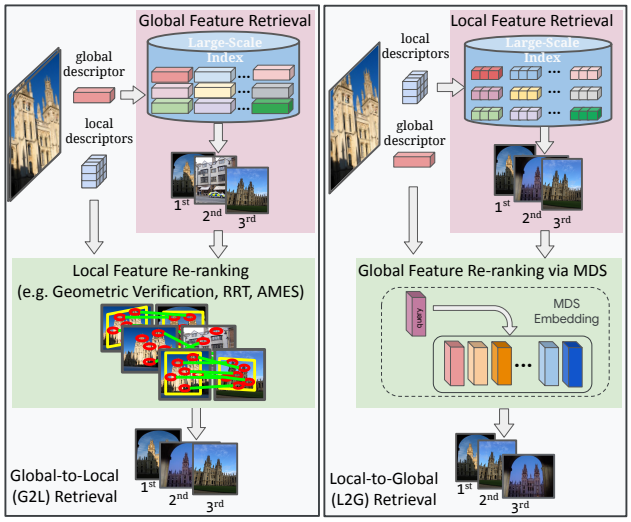

Figure 1. In contrast to conventional Global-to-Local (G2L) image retrieval systems (left), where global features are used for the initial search followed by re-ranking with local features, we introduce a new **Local-to-Global (L2G)** paradigm (right). In L2G, efficient retrieval with local features meets effective re-ranking with global features. Critically, we propose a **novel global feature re-ranking stage** leveraging multidimensional scaling (MDS) to create query-specific re-ranking embeddings, which are sensitive to localized similarities. Experimentally, our system improves upon the state of the art significantly.

1. Introduction

Searching vast image databases efficiently with a query picture enables a number of multimodal applications, e.g. visual shopping [14, 21, 30], fine-grained entity identification [12, 39, 40], knowledge-based visual question answering [5, 10, 16], among others. Today, such image retrieval systems are generally designed leveraging a *global-to-local* paradigm [4, 13, 31, 32], where *global* image features are used in a first search stage and *local* image features are used to re-rank the initial list of retrieved candidates via detailed matching. This approach benefits from the discriminative power and compactness of global representations for large-scale similarity computation, combined with the localized similarity verification capabilities of local representations.

Despite the success of this framework, a significant concern is the lack of localized search capabilities at large scale, which lead to recall losses at the global feature search stage. For example, when the query image only has partial matches with relevant database images, the system is usually unable to return pertinent results due to the limitations of global feature similarity estimation. Besides, while the local feature-based re-ranking stage helps refining the initial matches, it generally does not leverage information across the shortlisted database images to enhance the refinement

process, which ends up limiting the final performance.

In this work, we address these challenges by proposing a *local-to-global* (L2G) retrieval system, illustrated in Fig. 1. For the initial search stage, we build on top of recent advances in scalable *local* feature retrieval [1], which enables localized search at large scale, enhancing the initial set of retrieved candidates. We then develop a novel *global* feature re-ranking process, which allows information sharing between the shortlisted images, based on detailed local similarities. More precisely, we propose to leverage multidimensional scaling (MDS) [26] to create a global feature for re-ranking purposes, on-the-fly, and to feed it into an effective re-ranking process [27]. MDS enables us to compute global features which respect the detailed local feature similarities. In summary, the key advantage of our proposed system is that it can efficiently retrieve and re-rank based on localized and detailed similarities – in particular, the re-ranking process can refine relevance scores with a customized embedding that approximates rich local similarities between a given query and its most relevant database images.

Contributions. In summary, this paper makes three main contributions:

(1) We propose a new *local-to-global* (L2G) image retrieval paradigm, which flips the conventional script, leveraging local feature search and global feature re-ranking. Our system enables image retrieval based on localized and precise similarities, which is generally difficult to achieve with previous methods.

(2) A critical component to make the L2G paradigm work effectively is a new re-ranking method, which creates global features on-the-fly at query time, respecting local similarities. Leveraging multidimensional scaling, this process creates re-ranking embeddings specific to a given query, allowing us to jointly process the shortlisted images and reorder them effectively.

(3) We showcase strong performance on the conventional Revisited Oxford and Paris datasets [24], with $2-3\%$ gain with respect to previous work, setting a new state of the art.

2. Related Work

Image retrieval and re-ranking systems have a long history in computer vision, even before deep learning techniques dominated the field. Initial promising results in this area were dominated by *local-to-local* methods, where local feature search was used to find candidates and re-ranking employed geometric verification [15, 18, 22]. While these traditionally leveraged hand-crafted local features [3, 15], they were later revisited in the deep learning era with deep image features [1, 19]. Little by little, direct local search techniques gave way to methods which aggregated hand-crafted local features into a global feature for search [11, 29, 34], instantiating the first *global-to-local* systems. These had the advantages of much simpler and lighter search mechanisms, also delivering improved recall. Such aggregation techniques were also shown effective with deep learned features [33, 35, 37]. Most of the deep learning work for image retrieval, though, has been focused on enhancing global features [2, 9, 17, 25, 38], which generally surpassed the performance of local aggregation techniques. This also led to a deep learned version of the *global-to-local* paradigm, with both global and local features being extracted in the same model [4, 13], possibly with additional learnable modules for local similarity estimation [31, 32]. More recently, researchers demonstrated that global features can also be used effectively for the re-ranking stage [27], which introduced a *global-to-global* system. Our work goes beyond these existing directions to introduce the *local-to-global* paradigm, which presents significant advantages compared to previous ones. We build on top of recently-proposed techniques for efficient local feature search and global feature re-ranking, which enables localized search at large scale and effective re-ranking that merges information across the query and shortlisted images.

Global feature-based re-ranking is a recent idea to improve retrieval systems, as introduced by [27], to efficiently reorder the shortlisted images found in the initial search stage. Its key insight is to leverage pairwise similarities among all shortlisted images, which can guide an aggregation process that refines the global features for re-ranking purposes. In this work, we go beyond to introduce the usage of multidimensional scaling [26] to create a new global feature at re-ranking time. This can help leverage detailed pairwise local feature-based similarities between the images, by converting them on-the-fly into an embedding space which respects those similarities. Such a re-ranking embedding can then be used in the procedure introduced in [27], to refine all of the embeddings based on pairwise similarities, enabling a significantly improved final list of shortlist images.

3. Local-to-Global Image Retrieval

The initial approaches to image retrieval relied heavily on hand-crafted local features. However, extracting and matching a large number of these features can be computationally expensive. With the advent of deep learning, global features, which capture the overall content and semantics of an image in a single vector representation, gained prominence. While global features offer efficiency, local features provide finer granularity and robustness to changes in viewpoint or occlusions. Some methods try to combine the strengths of both. For example, they might use global features for an initial retrieval and then re-rank results using local feature matching.

Recent advances in efficient algorithms have enabled the use of local features even in the initial retrieval stage, not

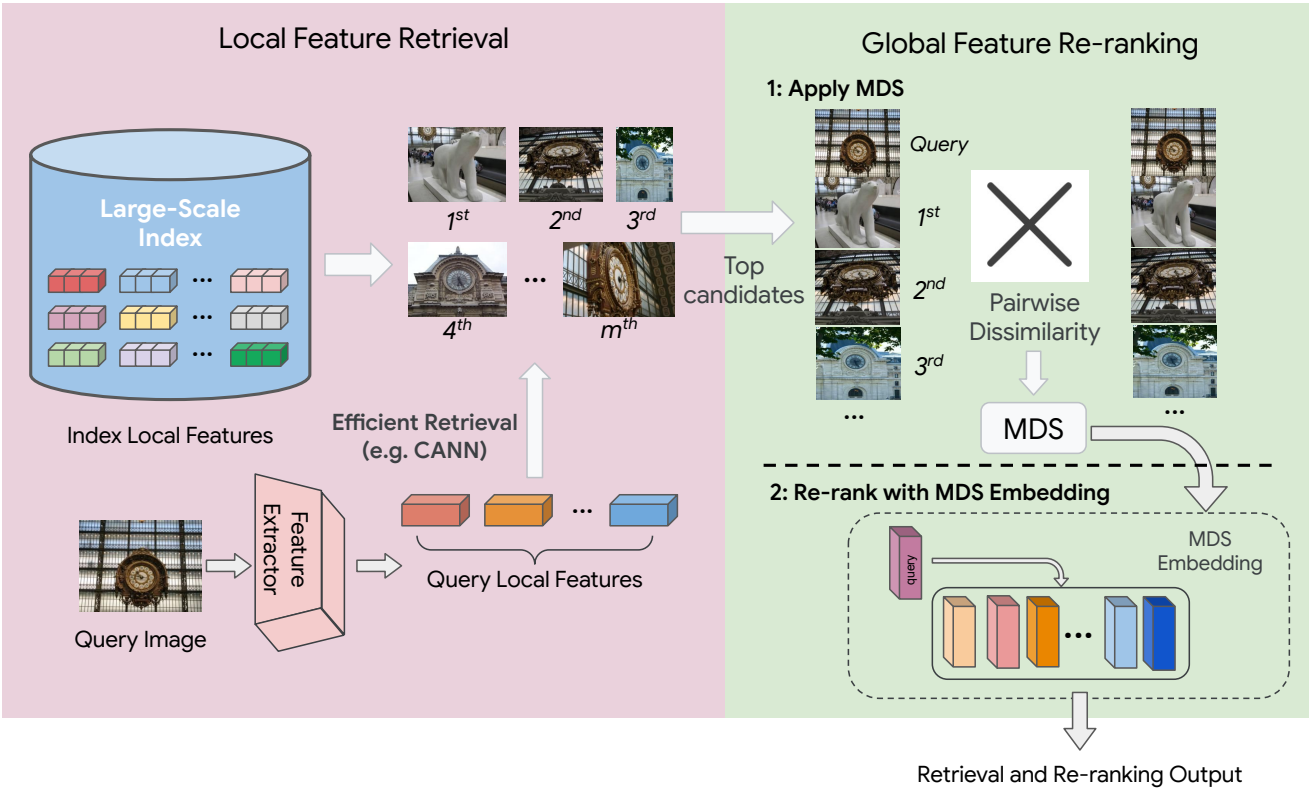

Figure 2. **Block diagram of the proposed Local-to-Global (L2G) retrieval system.** Left: given a query image, we extract local features which are used for efficient search via CANN [1]. Right: the top-ranked candidates from the local feature search stage undergo a re-ranking process leveraging embeddings computed on-the-fly via multidimensional scaling (MDS), based on the pairwise dissimilarities of the query and shortlisted database images.

just as a post-processing step. One such approach is Constrained Approximate Nearest Neighbors (CANN) [1], initially proposed for visual localization. CANN employs a novel nearest neighbor search strategy that efficiently finds the best matches in both appearance and geometry space using only local features and (asymmetric) Chamfer similarity. As a byproduct, the authors demonstrated the potential of CANN for efficient image retrieval using local features in the first stage. They showed that a simple weighted average of rankings obtained from both global and local features significantly improves retrieval quality. Another approach, MUVERA [7], utilizes multiple vector embeddings. It can also be used for image retrieval for efficient local feature-based image retrieval using the Chamfer similarity.

In this work, we present a novel and more effective method for merging local and global features, illustrated in Fig. 2. Leveraging local features for the initial search stage and global features for re-ranking, our Local-to-Global (L2G) method delivers effective retrieval performance. We also introduce a natural way to integrate the re-ranking technique proposed by Shao et al. [27] with any (dis)similarity, not necessarily a metric.

3.1. Re-ranking Global Features

Shao et al. [27] introduced a novel image re-ranking method that refines global features for improved retrieval performance. This method, designed to be plugged into any existing retrieval system, operates solely on global features, offering a significant efficiency advantage over conventional re-ranking techniques that rely on computationally expensive local features. Notably, it was the first solution to address both retrieval and re-ranking using only global image features.

Our work builds upon this concept but deviates from the convention of relying solely on global features. We leverage recent advances in efficient local feature-based retrieval, specifically CANN [1], to incorporate local information into the re-ranking process. To achieve this, we convert the ranking produced by CANN, which is based on non-metric similarity, into points in an embedding space. These points can then be treated as "global features" representing the local feature information. This transformation allows us to seamlessly integrate local and global features of any similarity. By merging these "globalized" local features with existing global features, we can effectively utilize the re-ranking

method of Shao et al. [27], resulting in improved retrieval accuracy.

3.2. Enhanced Re-ranking with MDS

Consider the following well known problem: Given pairwise dissimilarities, reconstruct a set of points that preserves pairwise distances. Multidimensional scaling (MDS) [26] is a family of widely used techniques for mapping data from a high-dimensional to a lower-dimensional space and for visualizing data. It is also a known method to reconstruct a set of points in high dimensional space from their pairwise distances. Given a set of dissimilarities, one can ask whether these values are distances and, moreover, whether they can even be interpreted as Euclidean distances. Given a dissimilarity (distance) matrix $D = (d_{ij})$, MDS seeks to find $x_1, \ldots, x_n \in R^p$ such that $d_{ij} \approx ||x_i - x_j||$ as close as possible. For certain cases, for some large p, there exists a configuration $x_1, \ldots, x_n$ with exact distance match $d_{ij} = ||x_i - x_j||$.

In such a case the distance involved is called Euclidean. There are, however, cases where the dissimilarity is distance, but there exists no configuration in any p with perfect match $d_{ij} = ||x_i - x_j||$, for some i, j. Such a distance is called non-Euclidean. Classical MDS is the case where we have Euclidean distance matrix $D = (d_{ij})$. In this case, we have a globally optimal solution (non-unique since any rigid transformation of it is always a solution) at some dimension p and there are efficient methods to find it. Metric MDS is where we are given a dimension p and a monotone function f, and we seek to find an optimal configuration $X \subset R^p$ that gives $f(d_{ij}) \approx \hat{d}_{ij} = ||x_i - x_j||_2$ as close as possible. In many applications of MDS, dissimilarities are known only by their rank order, and the spacing between successively ranked dissimilarities is of no interest or is unavailable. This is the Non Metric MDS where we are given a dimension p, and we seek to find an optimal configuration $X \subset R^p$ that gives $f(d_{ij}) \approx d^*_{ij} = ||x_i - x_j||_2$ as close as possible. Different approaches exist for Non-Metric MDS, including stress minimization techniques like SMACOF, which aim to minimize a stress function that quantifies the discrepancy between the disparities and the distances in the embedded space. Unlike metric MDS, here f is much general and is only implicitly defined. $f(d_{ij}) = d^*_{ij}$ are called disparities, which only preserve the order of d_{ij} , i.e.,

$$d_{ij} < d_{kl} \iff f(d_{ij}) \leq f(d_{kl}) \iff d^*_{ij} \leq d^*_{kl}$$

Our goal is to find an embedding in a high-dimensional space for the dissimilarity matrix of index and query images This method requires a complete set of pairwise distances between all images (both index and query) to perform nearest neighbor operations (for averaging) within a metric space.

However, our data presents two key challenges: (1) Non-metric dissimilarities: The initial ranking we obtain from local feature matching is non-metric. This means it may not satisfy the properties of a distance function, such as the triangle inequality. Directly applying the re-ranking method, which assumes a metric space, would lead to inconsistencies. (2) Incomplete distances signify that the relationships between certain pairs of data points are unknown. This absence of information can lead to a distorted representation of the data's overall structure in the embedding. In our case, obtaining pairwise distances between all pairs of index images is impractical, as it would require computation that grows quadratically with the dataset size. To maintain efficiency and scalability, we instead utilize a sparse distance matrix. In this sparse representation, each image only stores distances to its nearest neighbors, determined using an existing retrieval system.

To address these challenges, we turn to Multidimensional Scaling (MDS). While various MDS methods exist, we found the classic landmark MDS [28, 36], which relies on eigenvalue decomposition and Nyström approximation, less effective in our case. This is likely because landmark MDS generally requires metric distances and struggles with the non-metric nature of our dissimilarities.

Instead, we employ SMACOF (Scaling by MAjorizing a COmplicated Function), an iterative optimization algorithm first introduced in [6]. SMACOF minimizes a stress function that quantifies the discrepancy between the given dissimilarities and the distances in the reconstructed configuration. This iterative approach is well-suited for handling non-metric dissimilarities and incomplete data.

Specifically, SMACOF allows us to: (i) Handle non-metric dissimilarities: Effectively address the non-metric nature of our ranking data. (ii) Complete the dissimilarity matrix: Infer the missing entries to construct a complete distance matrix required for the re-ranking method. (iii) Weight dissimilarities: Assign weights to different dissimilarities based on their reliability and importance, potentially improving the embedding quality.

Complexity The computational complexity of MDS varies significantly depending on the specific method and the size of the data. Here's a breakdown of the complexity for some common MDS approaches: Classical MDS (Eigenvector-based) has complexity of $O(N^3)$ where N is the number of data points. This is dominated by the eigenvalue decomposition of the $N \times N$ dissimilarity matrix. SMACOF complexity is $O(N^2)$ per iteration and the number of iteration in our case is fairly small ($\approx$ 5-10). Landmark MDS complexity can be significantly lower than classical MDS for large datasets. By using a smaller set of "landmark" points (say, L landmarks), the complexity can be reduced to approximately $O(NL^2 + L^3)$. This makes

it more scalable for large N when $L << N$. FastMap [8] complexity is $O(Nk)$ where N is the number of data points and k is the desired dimensionality of the embedding.

3.3. On-the-fly MDS per query

Similarity embedding using MDS can be viewed as generating "global features" derived from a specific similarity measure, as opposed to generic embeddings trained on separate data. We take a similarity matrix and compute a global representation of the points that best preserves these similarities.

We assume that for an appropriate similarity (e.g., Chamfer), we have an efficient method (CANN, MUVERA) to index a large dataset and retrieve the top-k most similar images for any given query. We denote this method as EFF-INDEX-QUERY. Using EFF-INDEX-QUERY, we can pre-compute the top-k nearest index images for each image in the index offline. This results in a sparse set of pairwise distances, denoted as INDEX-SPARSE-DISTANCES, with a size linear in the number of data points since k is constant.

One approach is to apply MDS to the entire index at index time, creating and storing the embeddings just as we would store global features. This can be achieved in approximately linear time using fast MDS methods. At query time, we obtain the distances from the query to the top index images using EFF-INDEX-QUERY and then compute the query image embedding using the same methods used in landmark MDS, which requires only a constant number of images. With all embeddings computed, we can proceed with re-ranking in a metric space.

However, we propose a more efficient alternative that avoids applying MDS to the entire index. Instead, we compute MDS specifically for each query image and its top-k ranked index images, retrieved initially using EFF-INDEX-QUERY. This localized approach generates an embedding for only $k+1$ points, significantly reducing the computational burden.

For each query image, we use EFF-INDEX-QUERY to retrieve the top-ranked images and obtain their pairwise distances from INDEX-SPARSE-DISTANCES. If the distance between a pair of images is not present in INDEX-SPARSE-DISTANCES, we set it to 1 (the maximum possible distance). We then apply standard MDS to the $(k+1)\times(k+1)$ similarity matrix to obtain an Euclidean embedding of the query and its top-k neighbors. This computation takes $O(k^2)$ time, where k is typically a constant. While this can be further improved to $O(k)$ using landmark MDS or other fast approximate MDS methods, we opted for standard MDS in our experiments due to its efficiency for moderate k and our focus on demonstrating the core concept.

This localized embedding strategy allows us to apply efficient re-ranking techniques, similar to those in [27], without requiring index-time embedding. While the embedding is recomputed for each query, the overall computational cost remains manageable due to the small number of points involved.

4. Experiments

4.1. Experimental Setup

Our experiments are conducted on well-established benchmarks. Concretely, we use Oxford [22] and Paris [23] with revisited annotations, referred to as $\mathcal{R}$Oxf and $\mathcal{R}$Par, respectively. There are 4993 (6322) database images in the $\mathcal{R}$Oxf ($\mathcal{R}$Par) dataset, and each dataset contains a query set with 70 images. Large-scale results are further reported with the $\mathcal{R}$1M distractor set [24], which contains 1M database images.

We report the effectiveness of SMACOF MDS [6] re-ranking when combined with local feature retrieval. We leverage the FIRE [37] image features, and to ensure high retrieval efficiency, we follow the algorithm proposed by the CANN paper [1] and utilize the official implementation[1] and tune it on the $\mathcal{R}$Oxford dataset. For re-ranking with global features, we employ a weighted average between the MDS embeddings and SuperGlobal global features, where the MDS embeddings are obtained directly using the pairwise FIRE Chamfer similarities. We further tune the following hyperparameters on the $\mathcal{R}$Oxford dataset in order to obtain optimized re-ranking performance.

- ϵ which controls the convergence threshold for the MDS algorithm. A smaller ϵ generally leads to a more accurate embedding but requires more iterations.
- p (power), the modulation parameter that adjusts the influence of small and large distances in the Chamfer similarity metric. Higher values of p emphasize larger distances.
- w (weight) that determines the relative importance of the SuperGlobal features and the MDS embeddings when combining them for re-ranking. A higher w gives more weight to the MDS embeddings.
- k (top ranked for MDS) that specifies the number of top-ranked images from the initial retrieval that are used for MDS embedding. It's important to note that this is distinct from the M parameter used in the re-ranking stage, which determines the number of top-ranked images considered for neighborhood analysis.

In our experiments, re-ranking is always conducted among the top 1600 candidates. We use the standard mean Average Precision (mAP) as the evaluation metric.

4.2. Results

We compare our results with state-of-the-art models in Table 1. The results are split into four settings: (1) Global feature retrieval. (2) Global feature retrieval + Local feature

[1] https://github.com/google-research/google-research/tree/master/cann

	Medium				Hard			
Method	$\mathcal{R}$Oxf	$\mathcal{R}$Oxf+1M	$\mathcal{R}$Par	$\mathcal{R}$Par+1M	$\mathcal{R}$Oxf	$\mathcal{R}$Oxf+1M	$\mathcal{R}$Par	$\mathcal{R}$Par+1M
(1) Global feature retrieval								
RN50-DELG [4]	73.6	60.6	85.7	68.6	51.0	32.7	71.5	44.4
RN101-DELG [4]	76.3	63.7	86.6	70.6	55.6	37.5	72.4	46.9
RN50-DOLG [38]	80.5	76.6	89.8	80.8	58.8	52.2	77.7	62.8
RN101-DOLG [38]	81.5	77.4	91.0	83.3	61.1	54.8	80.3	66.7
RN50-CVNet [13]	81.0	72.6	88.8	79.0	62.1	50.2	76.5	60.2
RN101-CVNet [13]	80.2	74.0	90.3	80.6	63.1	53.7	79.1	62.2
RN50-SuperGlobal (No re-ranking) [27]	83.9	74.7	90.5	81.3	67.7	53.6	80.3	65.2
RN101-SuperGlobal (No re-ranking) [27]	85.3	78.8	92.1	83.9	72.1	61.9	83.5	69.1
(2) Global feature retrieval + Local feature re-ranking								
RN50-DELG (GV re-rank top 100) [4]	78.3	67.2	85.7	69.6	57.9	43.6	71.0	45.7
RN101-DELG (GV re-rank top 100) [4]	81.2	69.1	87.2	71.5	64.0	47.5	72.8	48.7
RN50-CVNet (Re-rank top 400) [13]	87.9	80.7	90.5	82.4	75.6	65.1	80.2	67.3
RN101-CVNet (Re-rank top 400) [13]	87.2	81.9	91.2	83.8	75.9	67.4	81.1	69.3
(3) SuperGlobal retrieval + Re-ranking								
RN50-SuperGlobal (Re-rank top 400) [27]	88.8	80.0	92.0	83.4	77.1	64.2	84.4	68.7
RN101-SuperGlobal (Re-rank top 400) [27]	90.9	84.4	93.3	84.9	80.2	71.1	86.7	71.4
AMES (600,600) (Re-rank top 1600) [31]	**93.6**	88.2	95.3	90.1	**84.8**	77.7	90.7	82.0
(4) Local feature retrieval + Global feature re-ranking								
L2G CANN-FIRE + MDS re-ranking (**Ours**)	92.9	**90.5**	**97.1**	**92.1**	83.0	**79.8**	**91.7**	**83.4**

Table 1. **Comparison to the state of the art.** Results of our L2G approach, compared to state-of-the-art methods on $\mathcal{R}$Oxford and $\mathcal{R}$Paris, on their base and extended "+1M" versions. Best results per dataset in **bold**, second-best underlined.

re-ranking, corresponding to the Global-to-Local (G2L) paradigm. (3) SuperGlobal retrieval + Re-ranking. (4) Local feature retrieval + Global feature re-ranking, corresponding to the proposed Local-to-Global (L2G) method. All the comparisons in (3) assume using SuperGlobal for retrieval. Moreover, we present `AMES(600,600) (re-rank top 1600)` for using 600 DINOv2 [20] local features for both query and candidate sides and re-ranking with AMES similarity, with numbers directly taken from [31] (this is also a G2L method).

We would like to highlight that our L2G approach, though applied to the FIRE feature which was proposed in 2022, already achieves state-of-the-art performance in most of the evals when combined with MDS, exceeding the performance of modern approaches such as SuperGlobal (proposed in 2023) and AMES (proposed in 2024). In particular, our L2G approach is extremely effective in large-scale retrieval settings and it achieves 79.8% in $\mathcal{R}$Oxf+1M Hard and 83.4% in $\mathcal{R}$Par+1M Hard, beating the best AMES results by 2.1% and 1.4%, respectively.

We conduct a study on $\mathcal{R}$Paris about the total number of correct images retrieved at different K values for all the 70 queries, as is shown in Fig. 3. We plot K within [400, 1600] since that is range for re-ranking. It shows that local feature retrieves more correct images at top K when K is no larger than 1200, but with the increase of K, the gap between local and global retrieval closes. For top 1600, local and global retrieves about the same number of correct images. It reveals that MDS re-ranking is critical, since the number of correct images we get from local feature or global feature retrieval is the same. But MDS re-ranking outperforms SuperGlobal when conducting re-ranking on top 1600, showing the importance of using an embedding that is sensitive to localized similarities.

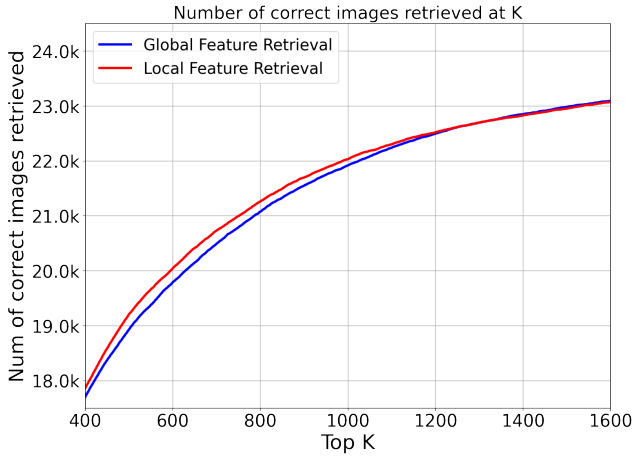

Figure 3. Number of correct images retrieved at top K via local feature retrieval and global feature retrieval on the $\mathcal{R}$Paris dataset.

In Fig. 4, we provide qualitative results between the top results from (1) SuperGlobal retrieval and re-ranking and (2) our L2G approach, with images taken from $\mathcal{R}$Oxford and $\mathcal{R}$Paris. The ranking positions are selected such that L2G retrieves matching images (highlighted in green boxes) while SuperGlobal doesn't (highlighted in red boxes). We

Configuration	$\mathcal{R}$Paris	$\mathcal{R}$Oxford
Full Model	91.7	83.0
- Without SuperGlobal re-ranking process	84.6	73.3
- Without merging final similarities with SuperGlobal	89.4	81.6
- Replace MDS re-ranking by SuperGlobal re-ranking	86.4	72.4
- Replace FIRE local similarity by AMES	93.6	85.2

Table 2. Ablation Study: mAP for $\mathcal{R}$Oxford and $\mathcal{R}$Paris Hard with various components of the pipeline. See the text for details.

observe that L2G resolves many failure causes of SuperGlobal.

4.3. Ablation Study

We conducted an ablation study to analyze the impact of different components and configurations on our method's performance. The resulting mAP values are presented in Table 2. Each row in the table represents a different configuration:

Without SuperGlobal re-ranking process. This baseline configuration uses local FIRE/Chamfer similarity for retrieval and re-ranks only using the computed MDS embeddings, without the embedding refinement update proposed by SuperGlobal.

Without merging final features with SuperGlobal. Here, we only apply MDS re-ranking to the local features but without merging with SuperGlobal features (w=1.0), demonstrating the effectiveness of local re-ranking in isolation compared with the previous row. Considering the performance of the full model, we would also like to highlight the synergy between local features and SuperGlobal features, where without SuperGlobal features in re-ranking we observe regressions in the metrics.

Replace MDS re-ranking by SuperGlobal re-ranking. This configuration explores the direct use of CANN-FIRE for retrieval and then re-ranking with SuperGlobal features. The poor performance indicates the incompatibility of using a non-metric similarity for retrieval within the metric space of SuperGlobal features.

Replace FIRE local similarity by AMES. In this configuration, during re-ranking, MDS is applied to the improved local similarity from AMES [31] instead of the FIRE/Chamfer similarity. This configuration achieves the best overall performance, demonstrating the power of combining a strong local similarity measure with MDS re-ranking.

While AMES may not be computationally efficient for direct retrieval, last configuration in our ablation study underscores the crucial role of MDS in enabling effective re-ranking with local similarities. Without MDS, as in [31] local similarity would have to be applied after re-ranking, potentially limiting its impact. Our approach, by embedding the local similarity into a metric space, allows for a more integrated and synergistic combination of local and global information.

5. Conclusions

Our work presents a new *local-to-global* image retrieval system, leveraging local image features for the initial large-scale search and global image features, induced by the MDS of the local similarity, for the re-ranking stage. This is a significant departure from the today's conventional *global-to-local* paradigm, helping overcome issues with partial matches at large scale and insufficient local information for re-ranking a short list of images. Notably, we introduce a novel global feature re-ranking process which can effectively leverage local similarities by converting these similarities into a new embedding space which respects those. Leveraging multidimensional scaling, these re-ranking embeddings significantly boost performance. Our experiments showcase state-of-the-art results in conventional image retrieval datasets.

Future work. This work with MDS, particularly its fast variants, opens exciting possibilities for using it with diverse similarity measures, including learned ones like those in AMES [31]). The key observation that embedding from pairwise distances requires only a constant number of pairs suggests a novel and efficient approach to building indexing and query systems for any generic similarity. This research direction could lead to significant advancements in similarity search, enabling more efficient and accurate retrieval across diverse domains and applications.

Limitations. The local feature retrieval process is more expensive than the global feature one, however the use of CANN makes it very efficient and competitive, while at the same time providing better results. Our re-ranking technique is more expensive than SuperGlobal as it requires the multidimensional scaling step to compute the re-ranking embeddings – however, this can be efficiently computed, as previously discussed, while at the same time enhancing the accuracy of the system.

Figure 4. **Qualitative results.** Examples comparing our L2G method with FIRE [37] local features against SuperGlobal [27] retrieval and re-ranking, on representative queries from the $\mathcal{R}$Oxf and $\mathcal{R}$Par datasets.

References

[1] D. Aiger, A. Araujo, and S. Lynen. Yes, we CANN: Constrained Approximate Nearest Neighbors for Local Feature-Based Visual Localization. In *Proc. ICCV*, 2023. 2, 3, 5

[2] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural Codes for Image Retrieval. In *Proc. ECCV*, 2014. 2

[3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-Up Robust Features (SURF). *CVIU*, 2008. 2

[4] B. Cao, A. Araujo, and J. Sim. Unifying Deep Local and Global Features for Image Search. In *Proc. ECCV*, 2020. 1, 2, 6

[5] Y. Chen, H. Hu, Y. Luan, H. Sun, S. Changpinyo, A. Ritter, and M.-W. Chang. Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? In *Proc. EMNLP*, 2023. 1

[6] Jan De Leeuw. Applications of convex analysis to multidimensional scaling. In J. R. Barra, F. Brodeau, G. Romier, and B. Van Cutsem, editors, *Recent Developments in Statistics*, pages 133–145. North-Holland Publishing Company, 1977. 4, 5

[7] Laxman Dhulipala, Majid Hadian, Rajesh Jayaram, Jason Lee, and Vahab Mirrokni. Muvera: Multi-vector retrieval via fixed dimensional encodings. In *Advances in Neural Information Processing Systems*, 2023. 3

[8] Christos Faloutsos and King-Ip Lin. Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, pages 163–174, 1995. 5

[9] A. Gordo, J. Almazan, J. Revaud, and D. Larlus. End-to-end Learning of Deep Visual Representations for Image Retrieval. *IJCV*, 2017. 2

[10] Z. Hu, A. Iscen, C. Sun, Z. Wang, K.-W. Chang, Y. Sun, C. Schmid, D. Ross, and A. Fathi. REVEAL: Retrieval-Augmented Visual-Language Pre-Training with Multi-Source Multimodal Knowledge Memory. In *Proc. CVPR*, 2023. 1

[11] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating Local Image Descriptors into Compact Codes. *PAMI*, 2012. 2

[12] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3D Object Representations for Fine-Grained Categorization. In *Proc. ICCV Workshops*, 2013. 1

[13] S. Lee, H. Seong, S. Lee, and E. Kim. Correlation Verification for Image Retrieval. In *Proc. CVPR*, 2022. 1, 2, 6

[14] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *Proc. CVPR*, 2016. 1

[15] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 2004. 2

[16] T. Mensink, J. Uijlings, L. Castrejon, A. Goel, F. Cadar, H. Zhou, F. Sha, A. Araujo, and V. Ferrari. Encyclopedic VQA: Visual Questions About Detailed Properties of Fine-Grained Categories. In *Proc. ICCV*, 2023. 1

[17] T. Ng, V. Balntas, Y. Tian, and K. Mikolajczyk. SOLAR: Second-Order Loss and Attention for Image Retrieval. In *Proc. ECCV*, 2020. 2

[18] D. Nistér and H. Stewenius. Scalable Recognition with a Vocabulary Tree. In *Proc. CVPR*, 2006. 2

[19] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-Scale Image Retrieval with Attentive Deep Local Features. In *Proc. ICCV*, 2017. 2

[20] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P. Huang, H. Xu, V. Sharma, S. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024. 6

[21] J. Peng, C. Xiao, and Y. Li. RP2K: A Large-Scale Retail Product Dataset for Fine-Grained Image Classification. *arXiv:2006.12634*, 2021. 1

[22] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object Retrieval with Large Vocabularies and Fast Spatial Matching. In *Proc. CVPR*, 2007. 2, 5

[23] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In *Proc. CVPR*, 2008. 5

[24] F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking. In *Proc. CVPR*, 2018. 2, 5

[25] J. Revaud, J. Almazan, R. S. Rezende, and C. R. Souza. Learning With Average Precision: Training Image Retrieval With a Listwise Loss. In *Proc. ICCV*, October 2019. 2

[26] N. Saeed, H. Nam, M. Haq, and D. Saqib. A Survey on Multidimensional Scaling. *ACM Comput. Surv.*, 2018. 2, 4

[27] S. Shao, K. Chen, A. Karpur, Q. Cui, A. Araujo, and B. Cao. Global Features are All You Need for Image Retrieval and Reranking. 2023. 2, 3, 4, 5, 6, 8

[28] V. d. Silva and J. B. Tenenbaum. Sparse multidimensional scaling using landmark points. *Technical Report (Stanford University)*, 2004. 4

[29] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*, 2003. 2

[30] H. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep Metric Learning via Lifted Structured Feature Embedding. In *Proc. CVPR*, 2016. 1

[31] P. Suma, G. Kordopatis-Zilos, A. Iscen, and G. Tolias. AMES: Asymmetric and Memory-Efficient Similarity Estimation for Instance-level Retrieval. In *Proc. ECCV*, 2024. 1, 2, 6, 7

[32] F. Tan, J. Yuan, and V. Ordonez. Instance-level Image Retrieval using Reranking Transformers. In *Proc. ICCV*, 2021. 1, 2

[33] M. Teichmann, A. Araujo, M. Zhu, and J. Sim. Detect-to-Retrieve: Efficient Regional Aggregation for Image Search. In *CVPR*, 2019. 2

[34] G. Tolias, Y. Avrithis, and H. Jegou. Image Search with Selective Match Kernels: Aggregation Across Single and Multiple Images. *IJCV*, 2015. 2

[35] G. Tolias, T. Jenicek, and O. Chum. Learning and Aggregating Deep Local Descriptors for Instance-Level Recognition. In *ECCV*, 2020. 2

[36] Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Global versus local methods in nonlinear dimensionality reduction. *Neural Networks*, 23(1):125–136, 2010. 4

[37] P. Weinzaepfel, T. Lucas, D. Larlus, and Y. Kalantidis. Learning Super-Features for Image Retrieval. In *ICLR*, 2022. 2, 5, 8

[38] M. Yang, D. He, M. Fan, B. Shi, X. Xue, F. Li, E. Ding, and J. Huang. DOLG: Single-Stage Image Retrieval with Deep Orthogonal Fusion of Local and Global Features. In *Proc. ICCV*, 2021. 2, 6

[39] N.-A. Ypsilantis, K. Chen, B. Cao, M. Lipovský, P. Dogan-Schonberger, G. Makosa, B. Bluntschli, M. Seyedhosseini, O. Chum, and A. Araujo. Towards Universal Image Embeddings: A Large-Scale Dataset and Challenge for Generic Image Representations. In *Proc. ICCV*, 2023. 1

[40] N.-A. Ypsilantis, N. Garcia, G. Han, S. Ibrahimi, N. Van Noord, and G. Tolias. The Met Dataset: Instance-level Recognition for Artworks. In *Proc. NeurIPS Datasets and Benchmarks Track*, 2021. 1