Re-ranking Image Retrieval in Challenging Geographical Iconographic Heritage Collections

Emile Blettery, Valérie Gouet-Brunet

Re-ranking Image Retrieval in Challenging Geographical Iconographic Heritage Collections

Emile Blettery
LASTIG, Univ Gustave Eiffel, IGN-ENSG
Ville de Paris, DAC, DHAAP
France
emile.blettery@ign.fr

Valérie Gouet-Brunet
LASTIG, Univ Gustave Eiffel, IGN-ENSG
France
valerie.gouet@ign.fr

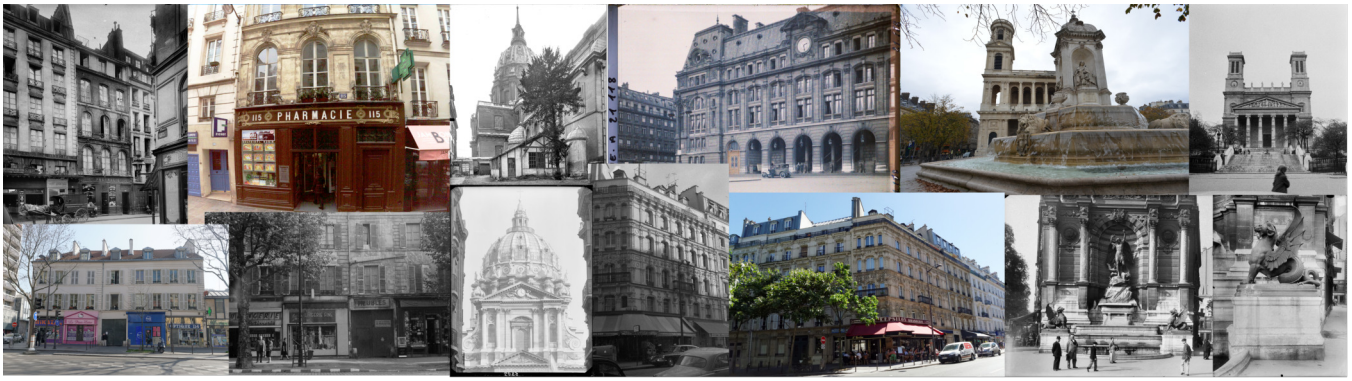

Figure 1: Examples of geographical iconographic heritage[1]

ABSTRACT

As the number of digitized geographic iconographic heritage collections increases, their global use is under-exploited by their lack of structure at large scale, which does not facilitate their access nor their understanding. Using automatic image retrieval methods appears to be the solution to bring structure by building links between contents, within and between collections. This paper presents an overview of methods for image retrieval applied to geographic iconographic heritage collections, both from the perspectives of image content description and of post-processing re-ranking. The article evaluates features and methods to identify their efficiency when faced with a challenging dataset. Moreover, new re-ranking approaches exploiting structuring information (scene geometry, metadata) are proposed to improve retrieval without having to adapt image descriptors to the specific data (retraining, fine-tuning, etc.) for every new specific collection.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Image search**; **Information retrieval**; • **Applied computing** → **Arts and humanities**.

CBMI'23, September 20–22, 2023, Orleans, France

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/3617233.3617259

KEYWORDS

Geographic iconographic heritage, Image retrieval, Re-ranking, Content interlinking

ACM Reference Format:
Emile Blettery and Valérie Gouet-Brunet. 2023. Re-ranking Image Retrieval in Challenging Geographical Iconographic Heritage Collections. In *20th International Conference on Content-based Multimedia Indexing (CBMI 2023), September 20–22, 2023, Orléans, France.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3617233.3617259

1 INTRODUCTION

In this article, we are interested in the geographical iconographic heritage, *i.e.* digitized or born-digital image collections, acquired at variable temporal periods and showing the territory and its human-made and natural visual landmarks. As a huge visual testimony of our environment, this category of contents is widespread but generally scattered in multiple provider or host institutions, such as GLAMs (Galleries, Libraries, Archives and Museums) or mapping agencies. Their exploitation embraces a large number of applications, ranging from historical and sociological studies up to mobile mapping scenarios, through digital tourism, education or landscape ecology. Some research works already try to exploit semi-automatic and automatic methods to use heritage collections for some of those applications ([3–5, 18]). As illustrated in Figure 1 with iconography

[1]From top to bottom and left to right: © Charles Lansiaux / DHAAP / Roger-Viollet; © IGN, Stereopolis; © Médiathèque du patrimoine et de la photographie; © Musée départemental Albert-Kahn; © Ville de Paris, COARC/Jean-Marc Moser; © Commission du Vieux Paris / DHAAP / Roger-Viollet; © Pascal Saussereau / DHAAP; © DHAAP / Roger-Viollet; © Donation Marcel Bovis, Médiathèque du patrimoine et de la photographie; © DHAAP / Roger-Viollet; © Marc Lelievre / DHAAP; © Charles Lansiaux / DHAAP / Roger-Viollet; © Charles Lansiaux / DHAAP / Roger-Viollet

from Paris at the street level, the visual representations associated with such contents are diverse given the various acquisition conditions (different sources, dates, viewpoints) and the evolution of landmarks over time, making their analysis still challenging today. They are usually described and indexed with metadata of variable quality and specification, making them not always easily interoperable, accessible, understandable in different contexts, and then largely under-exploited. To mitigate these issues and address new demands, one well-established alternative is to use image retrieval in order to describe, compare and link the contents directly, independently of the organization set up for the collection. Because of the specificity of such contents, in this work our aim is to evaluate image retrieval techniques for this type of data, in order to illustrate the practical feasibility of building visual links automatically and thus helping to bring structure to the collections.

The article is organized as follows: Section 2 revisits state-of-the-art descriptors and re-ranking approaches for image retrieval. Section 3 in turn presents the geographic iconographic dataset considered in the developments. We then evaluate retrieval methods, up to re-ranking, on this dataset in Section 4. Finally, in Section 5, we propose methods to improve the retrieval on such data, by exploiting the specificity and structure of the data and propagating it during re-ranking.

2 CONTENT-BASED IMAGE RETRIEVAL OVERVIEW

This section presents an overview of the state of the art on content-based image retrieval, both from the perspective of visual descriptors and the perspective of post-processing methods for re-ranking the images returned with the descriptors.

2.1 Features for image retrieval

Main improvements to image retrieval came with the advent of new image descriptors. From handcrafted to learned descriptors, the methods are numerous and current learned methods have proved to be state-of-the-art [7]. Multiple networks have first been developed for image classification and have been exploited for the more specific task of instance retrieval. Example backbones for image classification are VGG [30], ResNet [12] or ResNest [37]. State-of-the-art image descriptors are built on the features extracted by those backbones. On the one hand, the features can be aggregated in *one global feature* describing the whole image. Depending on the pooling operation used for the aggregation, different features can be extracted. Examples of global descriptors are SPoC [1], MAC [34] and RMAC [34], GeM [25] but also the more recent CVNet [15]. On the other hand, as with hand-made methods like VLAD [14], *local features* can also be extracted and aggregated to describe an image. Some learned methods also build on this paradigm, using attention mechanisms to select the most meaningful features in the image, such as DeLF [19] or How [33]. As with the visual-bag-of-words paradigm, local features are aggregated in a vector to perform similarity comparison between images. Several methods have been devised to perform the aggregation step: we can specifically mention ASMK [32] which performs highly with local features. Finally, several works attempt to combine local and global features to improve image retrieval performance, like DELG [6], DOLG [36] or CVNet [15]. However, whereas DOLG fuse both features to obtain a single descriptor, DELG and CVNet exploit the local features in a second time, as a trained re-ranking step, an essential step that will be discussed in section 2.2.

2.2 Re-ranking approaches

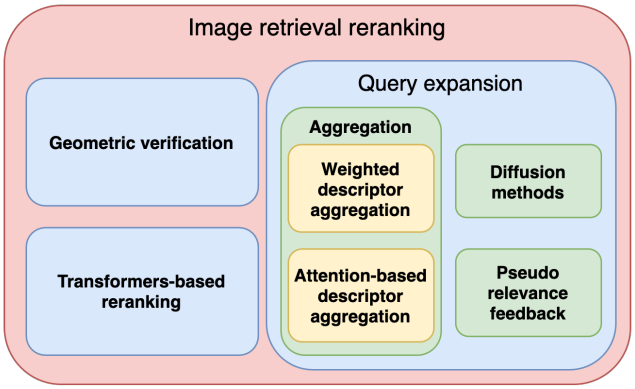

Figure 2: Re-ranking methods paradigms

.

Retrieving similar images simply based on visual descriptors and their similarities may not always yield the best results at the top of the list, because some other kinds of information, *e.g.* geometry in the image, was not encapsulated in the visual descriptor in order to be robust to the many transformations an image can undergo. Consequently, retrieval is usually considered as a two-step process: at first retrieval at large scale with descriptors, then re-ranking of the responses based on other finer or more specific criteria.

As presented in Figure 2, re-ranking methods can be divided into several categories. First, a very common re-ranking method is a geometric verification step. The idea is to match local features between the query image and each retrieved image, estimate the transformation parameters using a robust approach such as RANSAC [6, 10], and reorder the results based on the number of matches considered as inliers. Local features can be the ones used in the retrieval step as in [19] with DELF, but also other more precise features used in other computer vision tasks, such as SfM, like SuperPoint [9]. This step can also be included directly in the descriptor training process as in DOLG [36], DELG [6] or CVNet [15] which does not apply RANSAC with local features, but a dense cross-scale feature correlation to assess the coherence between images.

Another family of approaches regroups query expansion methods. The main idea is to take advantage of contextual information from the first retrieved images list by aggregating the features of the query and its most similar images to increase the meaningfulness of the query descriptor in order to improve the retrieval results. Multiple adaptations have been proposed, such as changing the aggregation weighting scheme (Average-QE [1], α-QE [25], etc.). Other approaches [16, 17] based on pseudo-relevance feedback aggregate features to be more similar to the first retrieved images and

more dissimilar to the further ranked images. More recent methods [11, 39] use an attention mechanism to select images and their weight in the aggregation process.

A specific kind of approach within query expansion re-ranking relies on diffusion, which propagates the similarity through the k-NN graph of similar images. Such solutions have achieved state-of-the-art performance on many benchmarks [2, 8, 13, 21, 29]. In the article, we focus on a representative method [38], that we describe in more detail: it first embeds a k_1-NN graph in an adjacency matrix A^* with i and j two images in the dataset and $N(i,k)$ the k most similar images to i:

$$A^*_{i,j} = \begin{cases} 1 & \text{if } j \in \mathcal{N}(i,k_1) \wedge i \in \mathcal{N}(j,k_1) \\ 0 & \text{if } j \notin \mathcal{N}(i,k_1) \wedge i \notin \mathcal{N}(j,k_1) \\ 0.5 & \text{otherwise} \end{cases} \quad (1)$$

Once the graph of k_1-NN is set as base for the GNN process, the k_2-NN graph is exploited to select the edges (e_{ij}) representing image similarity between images i and Jj that are used during the aggregation step to update the node (image) feature; k_2 is lower than k_1 (usually much more lower). The aggregation scheme, with $h_i^{(l)}$ the feature of image i at the l-th layer, is:

$$h_i^{(l+1)} = h_i^{(l)} + \sum e_{ij}^{\alpha} \cdot h_i^{(l)}, j \in \mathcal{N}(i,k_2) \quad (2)$$

This method exploits the manifold of the dataset and is very efficient because the message propagation is concurrent between all nodes. The whole dataset is re-ranked in one passage.

Finally, learned methods of re-ranking have been proposed to exploit the new paradigm of transformers. [31] proposes Re-ranking Transformers, a network that predicts the similarity of an image pair directly, provided their global and local features, as a replacement for geometric verification methods. Meanwhile, inspired by query expansion approaches, [20] proposes a transformer-based network that aggregates affinity features among the first results to enrich the representations of the images with some contextual information. [40] exploits transformers in an end-to-end fashion, both for global image description and retrieval and then for re-ranking. It exploits vision transformer tokens instead of handcrafted or CNN local features and reranks based on correlation between features rather than a pairwise geometrical verification.

3 THE GEOGRAPHICAL ICONOGRAPHIC HERITAGE COLLECTION CONSIDERED

The dataset we consider consists of more or less recent heritage content depicting Paris between 1915 and 2015 from a mostly ground-level perspective. The collections belong to eight providers:

- the Department of Architectural History of the City of Paris,
- the COARC, a service of the Department of Architectural History specialized in religious buildings,
- the mobile mapping 2015 Stereopolis dataset from the French Mapping Agency [22],
- the Planet's Archives - Paris of the Albert Kahn Museum,
- the Cité de l'Architecture et du Patrimoine,
- the Médiathèque du Patrimoine et de la Photographie,
- the Commission for the Old Paris,
- the Paris6K public benchmark [23].

In total, we assembled a dataset of 1,637 images of which an example is shown in Figure 1, divided into 31 classes depicting regular buildings, renowned monuments (*e.g.* the Panthéon), churches (*e.g.* the Saint-Sulpice church), and remarkable buildings (*e.g.* the Lavirotte building). To further challenge image retrieval in the experiments, we added 8,197 images as distractors (from the Department of Architectural History of the city of Paris), which leads to a total of **9,834 images** in the dataset.

Due to the large time period of acquisition and the multitude of providers, this dataset displays a large number of specific challenges for image retrieval:

- different techniques of acquisition, colors, etc.
- different resolution, levels of details, artisticity, etc,
- collection specificities increasing the above differences,
- changes in the scenes depicted due to the evolution of Paris throughout the century.

In addition to these images, several metadata may be available sometimes, such as an acquisition date or a location. The latter may be of various types, from an address manually provided (it is the case with some images of our dataset, *e.g.* those from the Dept. of Architectural History of the City of Paris) up to a precise pose of the camera (with the mobile mapping system Stereopolis).

4 EVALUATION OF STATE-OF-THE-ART METHODS ON OUR CHALLENGING DATA

In this section, we present the evaluation of state-of-the-art methods on the dataset presented in section 3. We first evaluate the recent literature on image descriptors and second, some of the most representative re-ranking methods.

4.1 Evaluation framework

All the experiments of the article are run on a Tesla-V100 GPU with 16 Go RAM and 10 CPU cores. We evaluate the efficiency of the approaches mostly with the mean Average Precision score (mAP); the implementation used is from [25][2]. The choice has been made to not retrain the learning networks involved, because we consider there are no training datasets existing for the data considered here, and that the constantly evolving data in digital humanities would require regular retraining which is not a realistic approach in practical use cases. All implementations are the authors' and the networks weights used are the ones provided by the authors.

4.2 Image descriptor evaluation

Four recent state-of-the-art image descriptors were evaluated: DELG [6], R101-GeM [12, 25], CVNet-Global [15], How+ASMK [32, 33]. All methods are deep detectors and descriptors. The first three ones produce a global feature per image and are trained on Google Landmarks Dataset v2 (GLDv2) [35], whereas the last one, trained on SfM120k [24], produces local features and then aggregates them using ASMK [32] for comparison between image descriptors.

A first evaluation was performed on the dataset without distractors as shown in Table 1. At this step, CVNet clearly outperforms the other three descriptors. However, when more deeply compared to How+ASMK (called How-A afterwards), it is revealed that it mainly

[2] https://github.com/filipradenovic/cnnimageretrieval-pytorch

Table 1: Score of tested image descriptors.

mAP	Dataset w/o distractors	Dataset
DELG [6]	53.2	-
R101 - GeM [12, 25]	57.9	-
CVNet-global [15]	67.3	37.1
How + ASMK [32, 33]	55.1	41.0

improves intra-collection retrieval and not between collections: indeed, when comparing the entropy of the distribution of the various providers in the true positive retrieved images, we observed that it is higher with How-A than with CVNet-Global. When including the large number of distractors in the dataset, How-A outperforms CVNet-Global; because it is a local descriptor, How-A proves to be more discriminative when the visual elements are very similar through images, which is frequent in the considered collections displaying redundant Parisian-style architectural features.

4.3 Re-ranking methods evaluation

As presented in Section 2.2, a large number of methods have been implemented to improve image retrieval results through a re-ranking step. Starting from How-A as visual descriptor, we have tested approaches from all families to evaluate what best suited our challenging data. Table 2 shows the performance of these approaches on the whole dataset, by providing an idea of the improvement in terms of mAP when exploiting re-ranking.

Table 2: Improvement of mAP with re-ranking

Approach	Order of magnitude
Weighted descriptor aggregation [1, 25]	+ 0.1
Pseudo relevance feedback [16]	< + 0.5
Transformers-based	
CSA [31], RRT [20]	- 10
CV-Net Rerank [15]	- 2
Geometric Verification [9, 26]	+ 0.5
Diffusion [29, 38]	+ 15

Descriptor aggregation methods [1, 25], as well as pseudo-relevance feedback ones [16], do not appear relevant to our problem because of the high variability of images, inducing a high variability in those multidimensional descriptors, in turn squashing the descriptor's specificity during aggregation.

Transformers based images (RRT [31], CSA [20]), and the re-ranking part of CVNet [15] all suffer from the same drawback: although reputed to be efficient, they are not here because by default trained on GLDv2, which is not suitable for specific heritage collections as ours. The solution would be to retrain or fine-tune those methods specifically to our data, but it has not been our choice as explained in Section 5.

Geometric verification is a very common step in image retrieval pipelines. In our case, we extract SuperPoint [9] local features, match them with the SuperGlue process [26] and using a classical RANSAC, we re-rank images based on their geometric coherence with the query. The mAP gain is moderate as shown in Table 2 (and confirmed further in section 5 where it serves as reference).

Finally, diffusion methods have the most impact at the re-ranking step. We tested SSR [29] and a GNN-based re-ranking method [38] (called GNN-R afterwards), where the mAP gain is substantial. Furthermore, GNN-R can be repeated multiple times to further extract similarity information and then improve retrieval; we have selected it as re-ranking method in the experiments of section 5.

5 HOW TO IMPROVE RETRIEVAL

Some existing re-ranking methods, as GNN-R, perform well on our dataset, by improving the retrieval results notably, but they do not really take into account its specificity. A commonly used solution would be fine-tuning existing state-of-the-art learned methods with our specific data. However, this solution has its drawbacks. First, it requires a certain number of annotated data possibly equally distributed among classes and providers, which is difficult to obtain in our case of collections with a sparse overlap in distribution. Secondly, fine-tuning on some specific collections may be efficient on those ones, but when confronted to their evolution (due to ongoing massive digitization, open data access, etc.), one may need to fine-tune the model all over regularly. For these reasons, we choose to study solutions that take into account the specificity of the data without having to retrain models specifically for them. Here, we propose three ways for improving retrieval, presented in sections 5.1, 5.2 and 5.3. They are evaluated in Table 3, facing several reference approaches, *i.e.* simple retrieval (How-A), retrieval with re-ranking with geometric verification (How-A + RANSAC), and several degrees of diffusion (GNN-R).

5.1 3D-based geometric verification

First, a main aspect in our dataset is the very large variation in viewpoint and level of detail, which allows to better understand the disappointing performance of the classical pairwise geometric verification such as RANSAC. To overcome this drawback, our first intuition was to use a 3D reconstruction of the scene, in order to check the geometric coherence of a result image not simply against the query image but against a more global geometry of the scene.

To do this, we reconstruct a 3D point cloud of the scene using Structure-from-Motion algorithms via the library Colmap [27, 28] based on keypoints extracted with [9, 26]. To reconstruct the scene, the query and the first ten retrieved images are used. If a scene is reconstructed successfully (at least two images including the query, using Colmap's default parameters), then the first k retrieved images are repositioned in the 3D scene through 2D-3D registration. These images are then re-ranked depending on their coherence with the scene: for each image, it is measured as the number of matches between 2D points in the image and 3D points in the point cloud exploited to compute its 3D pose.

Exploiting a 3D reconstruction encapsulates a lot more of the scene's global geometry and details than a single image, since the reconstruction belongs to several images. Thus, images very different from the query can still visually and geometrically be linked to the query image. The benefits of this approach, called R3D in the article, are quite clear when we refer to Table 3: we observe an

Table 3: mAP score with a re-ranking step on the dataset with distractors

Descriptor +	Diffusion after a first re-ranking step				Mean time
Re-ranking step	No GNN-R	GNN-R × 1	GNN-R × 2	GNN-R × 3	(k = 135)
How-A	41.0	57.2	59.3	57.0	
How-A + RANSAC	41.5	57.2	59.3	57.0	+120 s
How-A + R3D	44.4	61.9	**64.2**	61.9	+220 s
How-A + RANSAC + R3D	44.9	62.9	**65.8**	63.3	+340 s
How-A + R2D	36.4	60.0	**63.0**	60.5	+150 s
How-A + distance weighting (Stereopolis only)	41.7	58.8	61.7	59.5	+1/30s
How-A + distance weighting (all available locations)	43.3	59.4	61.8	60.0	+1/30s
How-A + R2D + distance weighting (all available locations)	36.4	60.1	63.0	60.6	+150 s

increase of 3.4% of mAP facing How-A, and of 2.9% facing How-A + RANSAC, without diffusion ("No GNN-R").

We can also note that all these results are notably improved by using several steps of diffusion up to 2 steps ("GNN-R × 2"), the results decreasing afterwards, whatever the approach. How-A + R3D reaches a mAP of 64.2% with a 2-step diffusion (an increase of 4.9%). In the Table, the "Mean time" column provides the averaged cost in terms of computation time for the first 135 images retrieved, by adding these post-processing to the simple retrieval with How-A. Not surprisingly, the R3D reconstruction takes more time than the RANSAC step.

We also observe that combining a simple RANSAC with the R3D reconstruction further improves the results (up to 65.8%) because the images used for the reconstruction process are geometrically more similar to the query, but at the expense of an even longer computation time, since both steps are performed successively.

5.2 2D geometric query expansion

Section 5.1 has demonstrated that exploiting a 3D reconstruction of the scene brings a great consistency in geometric verification. However, in some use cases, it could be considered as an operation too computationally costly. Thus, we have studied an alternative, called R2D, that tries to leverage the benefits the 3D information, by exploiting geometric information from the whole scene fully in 2D. The idea is to use the features extracted in similar images and reproject them in the query image to enrich its geometric significance and artificially enlarge the scene it encodes. It does not only encode the geometry of the scene it depicts, but also parts of the scene depicted by its most similar images.

The first step consists in creating all triplets with the query q and two images from its k most similar retrieved ones: (q, I_1, I_2) (with $k = 10$ in our experiments). Then, for each triplet:

- extract keypoints for images in triplet: sets K_q, K_{I_1}, K_{I_2}
- define matches pairwise: M_{q,I_1}, M_{q,I_2}, M_{I_1,I_2}
- define the query's **solid matches** as : $K_q^s = \left\{k \text{ if } M_{I_1,I_2} \circ M_{q,I_1}(k) = M_{q,I_2}(k), \forall k \in K_q\right\}$,
- define the query's **unsolid matches**: $K_q^u = \left\{k \text{ if } k \notin K_q^s, \forall k \in K_q\right\}$,
- if $|K_q^s| > 10$, estimate homographies $h_{I_1,q}$ and $h_{I_2,q}$,
- then reproject unmatched points of I_1 and I_2 in the query:

$$K_q^h = \left\{h_{I_1,q}(k) \text{ if } k \notin M_{q,I_1}[K_q^s], \forall k \in K_{I_1}\right\} \cup \left\{h_{I_2,q}(k) \text{ if } k \notin M_{q,I_2}[K_q^s], \forall k \in K_{I_2}\right\}$$

The three types of points are shown in the example of Figure 3. Once those steps are performed on all triplets, they are globally concatenated for each query: $K_q^a = K_q^s \cup K_q^u \cup K_q^h$.

Once the new set of points is created, all the similar images are matched pairwise and re-ranked using this new set, as follows:

- match the keypoints K_q^a and those (K_I) of image $I : M_{q,I}$
- select the subsample of solid matches S^s(matches with a solid keypoint) or the subsample of solid and unsolid matches $S^{s,u}$ if the number of solid matches is less than 5,
- estimate a transformation via RANSAC based on this subsample of matches,
- reevaluate the matches based on this transformation and keep the matches respecting this transformation up to a maximum difference of 10 pixels,
- the final score s_I of I is computed using the number of each type of match (solid: M_I^s,unsolid: M_I^u, reprojected: M_I^h) and the subsample of points used for the RANSAC:

$$s_I = \begin{cases} 10 \times |M_I^s| + 5 \times |M_I^u| + 10 \times |M_I^h| & \text{if } S^s \text{is used} \\ 10 \times |M_I^s| + 5 \times |M_I^u| + 5 \times |M_I^h| & \text{if } S^{s,u} \text{is used} \\ 10 \times |M_I^s| + 5 \times |M_I^u| + 1 \times |M_I^h| & \text{otherwise} \end{cases} \quad (3)$$

Using this 2D geometric query expansion is less efficient than the 3D reconstruction. As shown in Table 3 with How-A+R2D, the performance drops even compared to a setting without re-ranking. It can be explained by the fact that R2D relies on several homography estimations which remains a rough estimation of the scene geometry compared to the 3D reconstruction. However, the benefits of these method reveal themselves when combined with the diffusion process: the mAP is 63% after a 2-step diffusion, for a computation time comparable to RANSAC's one. An explanation for such a high impact of the diffusion is the fact that R2D increases inter-collection retrieval, which the diffusion in turn leverages very well. Indeed, we have computed the entropy of the providers distribution on the true positives images returned: it revealed that R2D's entropy is close to R3D's and higher than RANSAC's one

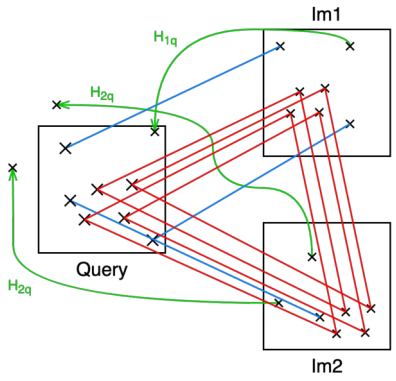

Figure 3: Query image set of keypoints with solid points in red, unsolid points in blue and reprojected ones in green

.

(and How-A's), which explains why diffusion after RANSAC leads to similar results as diffusion after How-A.

5.3 Metadata structure diffusion

As previously shown in Sections 5.1 and 5.2, exploiting structural (geometrical) information improves the retrieval performance. To continue to evaluate other tracks exploiting the specificity of the manipulated data, we chose to be interested in their metadata, starting from the observation that in practice, some metadata are present at least partially, in some of the collections. Like image retrieval, such data may provide useful links between images, that can be simply but efficiently combined with visual similarity.

To evaluate our hypothesis, we focused on the position information available for some of the images in our dataset. This is universal information, but potentially of varying nature: it can be directly available for images acquired through mobile mapping (*e.g.* Stereopolis), or the result of a geocoding of associated addresses (*e.g.* those manually provided by experts of the Dept. of Architectural History of the City of Paris). It should be noted, however, that their quality can be variable, due to potential human error (when manually added, copied and digitized), to acquisition precision (*e.g.* low-cost mapping) and to environment evolution (*e.g.* streets renamed or created through centuries). Then, based on the image location available and in addition to the visual similarity score provided by image retrieval, we define a spatial proximity score $w_{i,j}$ between two images i and j, which takes the location quality into account with a confidence rate (c_i for i and c_j for j), as follows:

$$w_{i,j} = \begin{cases} S(x_{i,j})^{\frac{1}{c_i \times c_j}} & \text{if } S(x_{i,j}) < 1 \\ S(x_{i,j})^{c_i \times c_j} & \text{otherwise} \end{cases} \quad (4)$$

$S(x_{i,j})$ is a proximity score based on the spatial Euclidean distance between i and j: $x_{i,j}$ (normalized over the diameter of Paris in our experiments). We define S as a double sigmoid function:

$$S(x_{i,j}) = a + (b-a) \times \frac{\tanh(k_1(x_{i,j} - X_1)) + 1}{2} + (c-b) \times \frac{\tanh(k_2(x_{i,j} - X_2)) + 1}{2} \quad (5)$$

with a, b, c the bottom, middle and top values of the double sigmoid's plateaux. X_1, X_2, k_1 and k_2 are respectively the values for the inflexion point and the steepness coefficient for both slopes.

$w_{i,j}$ ranges in [0,2] and equals 1 if we do not have location information for both images. Confidence rate c_i is 1 for Stereopolis, as it is precisely acquired as part of the mobile mapping, while for geocoded addresses, we set it below (0.9 for queries, 0.8 for distractors), because they are considered slightly less reliable.

We then combine the proximity and visual similarity scores between couples of images, through a simple weighting of the similarity score with the weight of equation 4, with the objective of limiting incoherent retrieval errors due to the limitations of visual descriptors. The whole process does not modify the eventual other steps of re-ranking, which are applied as previously explained.

As shown in Table 3, exploiting this spatial information improves the retrieval score more than a classic geometric verification, for a negligible online computation cost. We first exploited all possible location information available, for queries and distractors (*i.e.* 80.5% of the dataset). Then only Stereopolis locations (5% of the dataset with distractors) were used: using only a small part of the dataset is still more efficient than a simple geometric verification, especially when combined with diffusion which further propagates the structure in the retrieval process. Also, not surprisingly, the more positional information available, the better the retrieval results. However, after diffusion, the results are not much higher, suggesting that quantity of information is not as important as certainty and distribution among the dataset. Furthermore, combining a geometric re-ranking step with distance weighting on the similarity scores used in the diffusion process does not increase retrieval any further (experiment "R2D + distance weighting" in Table 3), which tends to show that data linked through R2D do not suffer from inconsistency in terms of location.

6 CONCLUSION

In this paper, we have evaluated state-of-the-art features and re-ranking methods for image retrieval with a challenging dataset of geographic heritage images. We show that diffusion-based re-ranking methods greatly improve retrieval, without considering a re-training step on the data. To further improve retrieval, we propose re-ranking approaches exploiting the structure of the dataset better: two re-ranking methods exploiting a more global geometry of the scene, and a weighting scheme using the available metadata information, here location. Once combined with diffusion-based methods, the proposed approaches improve retrieval for a computational cost on par with classical methods. As a perspective to deal with the most difficult retrieval cases, we think to continue to exploit the potential of diffusion-based methods by wisely injecting punctual manual intervention in order to propagate more structure in the dataset, without adding too much manual overhead.

ACKNOWLEDGMENTS

This work was financed by the City of Paris and the French ANRT through the Cifre grant 2019/1841. This work was carried out using HPC resources from GENCI-IDRIS (Grant 2022-AD011013510).

REFERENCES

[1] Artem Babenko and Victor Lempitsky. 2015. Aggregating local deep features for image retrieval. In *Proceedings of the IEEE international conference on computer vision*.

[2] Song Bai, Peng Tang, Philip H S Torr, and Longin Jan Latecki. 2019. Re-ranking via metric fusion for object retrieval and person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[3] Nicolas Blanc, Timothée Produit, and Jens Ingensand. 2018. *A semi-automatic tool to georeference historical landscape images*. Technical Report.

[4] Emile Blettery, Nelson Fernandes, and Valérie Gouet-Brunet. 2021. How to Spatialize Geographical Iconographic Heritage. In *Proceedings of the 3rd Workshop on Structuring and Understanding of Multimedia HeritAge Contents (SUMAC'21)*.

[5] Emile Blettery, Paul Lecat, Alexandre Devaux, Valérie Gouet-Brunet, Frédéric Saly-Giocanti, Mathieu Brédif, Laetitia Delavoipière, Sylvaine Conord, and Frédéric Moret. 2020. A spatio-temporal web application for the understanding of the formation of the parisian metropolis. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

[6] Bingyi Cao, Andre Araujo, and Jack Sim. 2020. Unifying deep local and global features for image search. In *Proceedings of ECCV*.

[7] Wei Chen, Yu Liu, Weiping Wang, Erwin M Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. 2022. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[8] Agni Delvinioti, Hervé Jégou, Laurent Amsaleg, and Michael E Houle. 2014. Image retrieval with reciprocal and shared nearest neighbors. In *2014 international conference on computer vision theory and applications (VISAPP)*.

[9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*.

[10] Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* (1981).

[11] Albert Gordo, Filip Radenovic, and Tamara Berg. 2020. Attention-based query expansion learning. In *Proceedings of ECCV*.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

[13] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Teddy Furon, and Ondrej Chum. 2017. Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[14] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. 2010. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*.

[15] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. 2022. Correlation Verification for Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Wei-Chao Lin. 2019. Aggregation of Multiple Pseudo Relevance Feedbacks for Image Search Re-Ranking. *IEEE Access* (2019).

[17] Wei-Chao Lin. 2022. Block-based pseudo-relevance feedback for image retrieval. *Journal of Experimental & Theoretical Artificial Intelligence* (2022).

[18] Ferdinand Maiwald, Jonas Bruschke, Christoph Lehmann, and Florian Niebling. 2019. A {4D} information system for the exploration of multitemporal images and maps using photogrammetry, web technologies and {VR}/{AR}. *Virtual Archaeology Review* (2019).

[19] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. 2017. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*.

[20] Jianbo Ouyang, Hui Wu, Min Wang, Wengang Zhou, and Houqiang Li. 2021. Contextual similarity aggregation with self-attention for visual re-ranking. *Advances in Neural Information Processing Systems* (2021).

[21] Shanmin Pang, Jin Ma, Jianru Xue, Jihua Zhu, and Vicente Ordonez. 2019. Deep Feature Aggregation and Image Re-Ranking With Heat Diffusion for Image Retrieval. *IEEE Transactions on Multimedia* (2019).

[22] Nicolas Paparoditis, Jean-Pierre Papelard, Bertrand Cannelle, Alexandre Devaux, Bahman Soheilian, Nicolas David, and Erwann Houzay. 2014. *Revue Française de Photogrammétrie et de Télédétection* (2014).

[23] Filip Radenović, Ahmet Iscen, Giorgios Tolias, Yannis Avrithis, and Ondřej Chum. 2018. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

[24] Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2016. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *Proceedings of ECCV*.

[25] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. 2019. Fine-tuning CNN Image Retrieval with No Human Annotation. *TPAMI* (2019).

[26] Paul Edouard Sarlin, Daniel Detone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. SuperGlue: Learning Feature Matching with Graph Neural Networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2020).

[27] Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[29] Xi Shen, Yang Xiao, Hu Shell Xu, Othman Sbai, and Mathieu Aubry. 2021. Re-ranking for image retrieval and transductive few-shot classification. In *Conference on Neural Information Processing Systems (NeurIPS)*.

[30] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015).

[31] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. 2021. Instance-level image retrieval using reranking transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[32] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. 2016. Image search with selective match kernels: aggregation across single and multiple images. *International Journal of Computer Vision* (2016).

[33] Giorgos Tolias, Tomas Jenicek, and Ond\vrej Chum. 2020. Learning and aggregating deep local descriptors for instance-level recognition. In *Proceedings of ECCV*.

[34] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. 2016. Particular Object Retrieval With Integral Max-Pooling of CNN Activations. In *ICLR 2016-International Conference on Learning Representations*.

[35] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2575–2584.

[36] Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, and Jizhou Huang. 2021. 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*.

[37] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, Mu Li, and Alexander Smola. 2022. ResNeSt: Split-Attention Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[38] Xuanmeng Zhang, Minyue Jiang, Zhedong Zheng, Xiao Tan, Errui Ding, and Yi Yang. 2020. Understanding Image Retrieval Re-Ranking: A Graph Neural Network Perspective.

[39] Xulu Zhang, Zhenqun Yang, Hao Tian, Qing Li, and Xiaoyong Wei. 2022. Indicative Image Retrieval: Turning Blackbox Learning into Grey. *arXiv preprint arXiv:2201.11898* (2022).

[40] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. 2023. R^2Former: Unified *R*etrieval and *R*eranking Transformer for Place Recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*.