



Structuring heritage iconographic collections : from automatic interlinking to semi-automatic visual validation

Emile Blettery

► To cite this version:

Emile Blettery. Structuring heritage iconographic collections : from automatic interlinking to semi-automatic visual validation. Databases [cs.DB]. Université Gustave Eiffel, 2024. English. NNT : 2024UEFL2001 . tel-04550089

HAL Id: tel-04550089

<https://theses.hal.science/tel-04550089>

Submitted on 17 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Structuring heritage iconographic collections: from automatic interlinking to semi-automatic visual validation

Thèse de doctorat de l'Université Gustave Eiffel

École doctorale n° 532, Mathématiques et STIC (MSTIC)

Spécialité de doctorat: Sciences et Technologies de l'Information Géographique

Unité de recherche : LaSTIG - Laboratoire en Sciences et Technologies de l'Information Géographique

**Thèse présentée et soutenue à l'Université Gustave Eiffel,
le 09/01/2024, par :**

Emile BLETTERY

Composition du Jury

Ewa KIJAK

Maitresse de Conférence, HDR, Université Rennes 1

Rapportrice

Camille KURTZ

Professeur des universités, Université Paris Cité

Rapporteur

Livio DE LUCA

Directeur de recherche, UMR 3495 MAP CNRS/MCC

Président du jury

Florian NIEBLING

Professor, Technische Hochschule Köln, Germany

Examineur

Encadrement de la thèse

Valérie GOUET-BRUNET

Directrice de recherche, Univ. Gustave Eiffel, IGN, LaSTIG

Directrice de thèse

Laurent FAVROLE

Chef de département du DHAAP, Ville de Paris

Tuteur en entreprise

"A Marie et Belle-Maman. Et 1, et 2 et 3 docteurs !"

Abstract

This thesis explores automatic and semi-automatic structuring approaches for iconographic heritage contents collections. Indeed, exploiting such contents could prove beneficial for numerous applications. From virtual tourism to increased access for both researchers and the general public, structuring the collections would increase their accessibility and their use. However, the inherent "in silo" organization of those collections, each with their unique organization system hinders automatic structuring approaches and all subsequent applications.

The computer vision community has proposed numerous automatic methods for indexing (and structuring) image collections at large scale. Exploiting the visual aspect of the contents, they are not impacted by the differences in metadata structures that mainly organize heritage collections, thus appearing as a potential solution to the problem of linking together unique data structures. However, those methods are trained on large, recent datasets, that do not reflect the visual diversity of iconographic heritage contents. This thesis aims at evaluating and exploiting those automatic methods for iconographic heritage contents structuring.

To this end, this thesis proposes three distinct contributions with the common goal of ensuring a certain level of interpretability for the methods that are both evaluated and proposed. This interpretability is necessary to justify their efficiency to deal with such complex data but also to understand how to adapt them to new and different content.

The first contribution of this thesis is an evaluation of existing state-of-the-art automatic content-based image retrieval (CBIR) approaches when faced with the different types of data composing iconographic heritage. This evaluation focuses first on image descriptors paramount for the image retrieval step and second, on re-ranking methods that re-order similar images after a first retrieval step based on another criterion. The most relevant approaches can then be selected for further use while the non-relevant ones provide insights for our second contribution.

The second contribution consists of three novel re-ranking methods exploiting a more or less global spatial information to re-evaluate the relevance of visual similarity links created by the CBIR step. The first one exploits the first retrieved images to create an approximate 3D scene of the scene in which retrieved images are positioned to evaluate their coherence in the scene. The second one simplifies the first while extending the classical geometric verification setting by performing geometric query expansion, that is aggregating 2D geometric information from retrieved images to encode more largely the scene's geometry without the costly step of 3D scene creation. Finally, the third one exploits a more global location information, at dataset-level, to estimate the coherence of the visual similarity between images with regard to their spatial proximity.

The third and final contribution is a framework for semi-automatic visual validation and manual correction of a collection's structuring. This framework exploits on one side the most suited automatic approaches evaluated or proposed earlier, and on the other side a graph-based visualization platform. We exploit several visual clues to focus the expert's manual intervention on impacting areas. We show that this guided semi-automatic approach has merits in terms of performance as it solves mistakes in the structuring that automatic methods can not, these corrections being then largely diffused throughout the structure, improving it even more globally.

We hope our work will provide some first insights on automatically structuring heritage iconographic content with content-based approaches but also encourage further research on guided semi-automatic structuring of image collections.

Keywords: Content-based image retrieval, Re-ranking, Visual and spatial collection structuring, Graph-based visualization and structuring, Deep learning.

Résumé

Cette thèse explore des approches de structuration automatique et semi-automatique pour les collections de contenus iconographiques patrimoniaux. La structuration et l'exploitation de tels contenus pourrait s'avérer bénéfique pour de nombreuses applications, du tourisme virtuel à un accès facilité pour les chercheurs et le grand public. Cependant, l'organisation "en silo" inhérente à ces collections entrave les approches de structuration automatique et toutes les applications subséquentes.

La communauté de la vision par ordinateur a proposé de nombreuses méthodes automatiques pour l'indexation (et la structuration) de collections d'images à grande échelle. Exploitant l'aspect visuel des contenus, elles fonctionnent indépendamment des structures de métadonnées qui organisent principalement les collections patrimoniales, apparaissant ainsi comme une solution potentielle au problème de liage entre les structures uniques des différentes collections. Cependant, ces méthodes sont généralement entraînées sur de grands jeux d'images récentes ne reflétant pas la diversité visuelle des contenus patrimoniaux. Cette thèse vise à évaluer et à améliorer ces méthodes automatiques pour la structuration des contenus iconographiques patrimoniaux.

Pour cela, cette thèse apporte trois différentes contributions avec l'objectif commun d'assurer une certaine explicabilité des méthodes évaluées et proposées, nécessaire pour justifier de leur pertinence et faciliter leur adaptation à de nouvelles acquisitions.

La première contribution est une évaluation des approches automatiques de recherche d'images basée sur le contenu, confrontées aux différents types de données du patrimoine iconographique. Cette évaluation se concentre d'abord sur les descripteurs d'images de l'étape de recherche d'images, puis sur les méthodes de ré-ordonnement qui réorganisent ensuite les images similaires en fonction d'un autre critère. Les approches les plus pertinentes peuvent alors être sélectionnées pour la suite tandis que celles qui ne le sont pas fournissent des informations inspirant notre deuxième contribution.

La deuxième contribution consiste en trois nouvelles méthodes de ré-ordonnement exploitant des informations spatiales plus ou moins globales pour réévaluer les liens de similarité visuelle créés par l'étape de recherche d'images. La première exploite les premières images retrouvées pour créer une scène 3D approximative dans laquelle les images retrouvées sont positionnées pour évaluer leur cohérence dans la scène. La deuxième simplifie la première avec une expansion de requête géométrique, c'est-à-dire en agrégeant des informations géométriques 2D issues des images récupérées pour encoder plus largement la géométrie de la scène sans la reconstruire (ce qui est coûteux en temps de calcul). Enfin, la troisième exploite des informations de position plus globales, à l'échelle du jeu d'images, pour estimer la cohérence entre la similarité visuelle entre images et leur proximité spatiale.

La troisième et dernière contribution est un processus semi-automatique de validation visuelle et de correction manuelle de la structuration d'une collection. Ce cadre exploite les approches automatiques les plus adaptées et une plateforme de visualisation basée sur une représentation en graphes. Nous utilisons plusieurs indices visuels pour orienter l'intervention manuelle de l'expert sur les zones impactantes. Cette approche semi-automatique guidée présente des avantages certains, car elle résout des erreurs de structuration qui échappent aux méthodes automatiques. Ces corrections étant ensuite largement diffusées dans toute la structure, l'améliorant globalement.

Nous espérons que notre travail apportera quelques perspectives sur la structuration automatique de contenus iconographiques patrimoniaux par des approches basées sur le contenu, tout en ouvrant la porte à davantage de recherches sur la structuration semi-automatique guidée de collections d'images.

Mots-Clés: Recherche d'images par contenu visuel, Ré-ordonnancement des résultats, Structuration visuelle et spatiale des collections, Structuration et visualisation basée graphe, Apprentissage profond.

Acknowledgments

I would like here to express my gratitude to all the people who, directly or indirectly, have helped me complete this thesis.

To my supervisor, Valérie Gouet-Brunet, thank you for everything. These past three years have not been a smooth journey, and I would not have succeeded without you. Thank you for your guidance, your high standards, and your support when things got tougher than expected. I will miss our preliminary discussions, ranging from interior design to politics, along with the latest interesting exhibitions and trips to the other side of the world.

To Laurent Favrole, my supervisor at the Ville de Paris, thank you for your occasional presence, which allowed me to ground my work in a practical context with feedback from potential users of my research.

To Livio De Luca, a special thank you for offering me the opportunity to join the Notre-Dame project after my thesis, allowing me to finish it with relative serenity. And thank you for the statement that my work was valid and could be defended during the first presentation of my findings; it meant a lot during the writing process.

I would also like to thank all the colleagues with whom I have had the chance to exchange ideas during these years at LaSTIG. I cannot mention anecdotes for each one of you, but please know that they are etched in my memory.

Firstly, Nathan and Dimitri, who preceded me in this field of research and provided initial guidance to get started. Then, Anatole, Yanis, Vivien, Luc, Mehdi, who preceded me as doctoral students and whose experience was invaluable when I found myself in their shoes.

Thanks to the permanent staff at the laboratory with whom I exchanged more or less, making my stay these past four years more enjoyable: Alexandre, Ana-Maria, Arnaud, Bénédicte, Bruno, Catherine, Clément, Ewelina, Gérald, Jacques, Julien, Lâman, Laurent, Laurence, Loïc, Manchun, Marc, Maria-Jesus, Mathieu, Nathalie, Paul, Sébastien, Sidonie.

Thanks to the interns and other young researchers who crossed my path, with whom I shared coffee breaks, foosball, moments of laughter, and more serious discussions, making it easier to talk about research questions: Azelle, Bérénice, Chahine-Nicolas, Charly, Christelle, Florent, Grégoire, Guillaume, Karim, Lu-Lin, Laura, Martin, Matilde, Mohamed, Mouhamadou, Nelson, Quentin, Raphaël, Salomé, Sami, Samuel, Wu Teng, and I'm sure I'm forgetting some.

A special thank you to Florent, Melvin, Solenn, and Camelvin, who brightened up my office in the past year. I'm glad to know that I will continue to share office space with each of you even when you're 800 kilometers apart.

Another very special thank you to Michelle, you can not know how much our "real-life" talks every morning helped me get through tough times just by knowing that my first interaction at work would be a pleasant one with you. Many thanks also to Aurore and Martine, your smiles, jokes and recipes helped me through tough days.

A special thought to the fellow doctoral students who defended or will defend alongside me. You are the ones who lived through the thesis journey most similarly to mine, and

our discussions were invaluable. Damien, thank you for organizing those collections, honey waffles, and, most importantly, for your encouragement when imposter syndrome was too strong. Charles, we may have known each other for a shorter time, but we shared many laughs in between, and thank you for being the perfect rubber duck. Helen Mair (I did it like Romain!), thanks for improving my English (Oh, I do not mean my vocabulary!), thanks for laughing at my more or less successful jokes, and thanks for your encouragement and open-hearted discussions that helped me through the toughest times. Finally, Romain, thank you for believing in me, for your unwavering support even when things were not easy for you. Thank you for your valuable mathematical approach and contagious research passion. Thanks for our exciting discussions, from theology to sushi and everything in between. A very special thanks to Helen and you; of all the things I looked for in recent years, you are the ones I am most glad to have found.

To all my high school friends, especially Clément, thank you for reminding me that life goes on, and not everything revolves around the thesis. Those meals, pool games, and moments of laughter were refreshing breaks in this long adventure.

To my family, different thanks depending on the case, but equally important. Firstly, to my grandparents, I don't think I will ever stop being your first grandson trying to make you proud. I hope that's enough because I won't do it again, and thank you for your discreet support.

To Dad, thank you for refraining from asking too often, "How is it going?" and if "I will finish on time." Thank you for your presence and your humor as lame as mine (and no, sorry, we won't launch a start-up together after my thesis). Thanks to Belle-Maman for your knowledgeable encouragement and attentive listening; and especially for preventing Dad from asking "How is it going?" and if "I will finish on time" even more often. Finally, Mom, thank you for pretending that all of this was normal, for your unwavering support, and especially for your invitation to a nice restaurant if I manage to finish this thesis (my second-biggest motivation).

Bleurette, Anémone, Térébentine, and Robinson, thank you for remaining true to yourselves, it helped keep me grounded.

To Lio, Carole, Philippe, and all the others I don't have space to mention here, thank you for your interest in my work and your support in all culinary forms possible (harissa, dried fruits, and tomato sauce meatballs (there were enough, I promise)).

Finally, Marie, I think it's no secret that if this thesis eventually saw the light of day, it is in large part thanks to you. Firstly because my biggest motivation turned out to be becoming a doctor before you, as I jokingly said. More seriously, your presence in difficult times was decisive; I would have had little chance of finishing without you. Thank you for calming my anxieties, reminding me that things are progressing, keeping me grounded even when I felt disconnected. Thank you for enduring my neuroses, my procrastination, and my lack of organization. Thank you for reminding me often why I was doing this thesis and that I was capable of success. It would probably take more words than the eight chapters that follow to make an exhaustive list of all the reasons I have to thank you, but I'll stop here; I think you already know the rest. Thank you.

Table of Contents

Table of Contents	13
List of Figures	17
List of Tables	19
1 Introduction	21
1.1 Motivations	22
1.1.1 Various potential applications	22
1.1.2 A hindering lack of structure but several possible solutions	23
1.2 Objectives and contributions	24
1.2.1 Context	27
1.3 Publications	27
2 Iconographic Heritage: a focus on Paris	29
2.1 Introduction	29
2.2 Iconographic heritage: a challenging object of study	30
2.2.1 Heterogeneous data	31
2.2.2 General organization and structure	33
2.3 A test dataset focusing on Paris	34
2.3.1 City of Paris' image data	35
2.3.2 Stereopolis dataset	36
2.3.3 Other providers	37
2.3.4 Summary of the final dataset	40
2.4 Conclusion	43
I Automatic Retrieval and Re-ranking	45
3 Related Work and its Evaluation	47
3.1 Introduction	47
3.2 Image retrieval	49
3.2.1 Hand-crafted descriptors	49
3.2.2 Learned descriptors	50
3.2.3 Similarity search	56
3.2.4 Spotlight on How and ASMK	58
3.2.5 CBIR and iconographic heritage	60
3.3 Re-ranking	64
3.3.1 Late fusion	65
3.3.2 Geometric verification	65
3.3.3 Transformers-based re-ranking	67
3.3.4 Query expansion	68

3.4	Image descriptors evaluation	72
3.4.1	Descriptors evaluated and evaluation choices	72
3.4.2	Choice between How-A, R101-GeM and DELG	74
3.4.3	Choice between How-A and CV-Net	77
3.5	Re-ranking methods evaluation	78
3.5.1	Re-ranking choices	78
3.5.2	Aggregation	79
3.5.3	Pseudo-relevance feedback	80
3.5.4	Transformers-based	80
3.5.5	Geometric verification	80
3.5.6	Diffusion methods	81
3.6	Conclusion	82
4	Our Contributions to Re-ranking	85
4.1	Introduction	85
4.2	Geometric query expansion	86
4.2.1	A 3D based proof of concept	86
4.2.2	A 2D geometric query expansion proposition (R2D)	90
4.3	Metadata exploitation	96
4.3.1	Metadata structure weighting scheme	96
4.3.2	Structuring with location information	98
4.4	Conclusion	99
5	Re-ranking Strategies Evaluation	101
5.1	Introduction	101
5.2	Contributions' evaluation	102
5.2.1	3D geometric query expansion	102
5.2.2	2D geometric query expansion	103
5.2.3	Location proximity weighting	104
5.3	Combination of re-ranking methods	105
5.3.1	Multiple re-ranking combinations' evaluation	105
5.3.2	Insight on provider entropy impact for re-ranking	109
5.4	Runtime of methods and combinations	113
5.5	Key takeaways	114
5.6	Conclusion	115
II	Graph-based Semi-automatic Structuring	117
6	Structuring, Spatializing and Visualizing Iconographic Heritage	119
6.1	Introduction	119
6.2	Image spatialization	120
6.2.1	Spatialization paradigms overview	120
6.2.2	Manual spatialization	124
6.2.3	Semi-automatic spatialization	125
6.2.4	Automatic spatialization	128
6.3	Iconographic heritage structuring and visualization	131
6.3.1	Single modality platforms	131
6.3.2	Multiple modalities combinations platforms	134
6.4	Conclusion	140

7	Graph-based Semi-automatic Re-ranking	143
7.1	Introduction	143
7.2	Graph representation of the structured dataset	144
7.2.1	Graph links considered	145
7.2.2	Presentation of the graph	147
7.3	Structuring process overview	150
7.3.1	Semi-automatic iterative process	150
7.3.2	Location propagation	152
7.4	Graph-based visualization platform	152
7.4.1	Visualization needs and technical solutions	153
7.4.2	The visualization platform	154
7.4.3	Visual clues as analysis support	159
7.5	Semi-automatic structuring process evaluation	166
7.5.1	Automatic quantitative evaluation	166
7.5.2	Qualitative visual evaluation	168
7.6	Conclusion	172
8	Conclusion and Perspectives	175
8.1	Contributions	175
8.2	Perspectives	177
8.2.1	Retrieval and reranking	177
8.2.2	Semi-automatic structuring platform improvements	179
8.2.3	Generalization to other types of data	179
	Bibliography	181
	Résumé détaillé de la thèse en français	203

List of Figures

2.1	Sample of images of iconographic heritage	30
2.2	Sacré-Cœur Basilica, at different times, from different perspectives using various media	32
2.3	Sample of images in our dataset	35
2.4	Examples of images from the Parisienne de Photographie	36
2.5	Example of images from the Stereopolis dataset	37
2.6	Example of images from the Paris 6K public benchmark	37
2.7	Example of images from the Médiathèque du Patrimoine et de la Photographie	38
2.8	Example of images from the Musée Albert Kahn’s collection	39
2.9	Example of images from the Cité de l’Architecture et du Patrimoine	39
2.10	Example of images from the Commission du Vieux Paris	40
2.11	Example of images from the Conservation des Œuvres d’Art Religieuses et Civiles of the city of Paris	40
2.12	Dataset statistical representation based on classes	42
2.13	Dataset statistical representation based on providers	42
2.14	Dataset global statistical representation	43
3.1	Global Content-Based Image retrieval pipeline	48
3.2	Overview of approaches for approximate nearest neighbors search	58
3.3	Overview of the architecture for How local features	59
3.4	Example of pattern spotting in art collections	61
3.5	Example of long term image retrieval	62
3.6	Re-ranking methods paradigms	64
3.7	Example of late fusion using multiple distances	65
3.8	Examples of RANSAC-based matches selection	66
3.9	Transformer-based pipeline for re-ranking using global and local features	68
3.10	Pseudo-relevance feedback pipeline	69
3.11	Graph-based approach to re-ranking	70
3.12	GNN-Reranking process	71
4.1	Illustration of the 3D reconstruction	87
4.2	Illustration of the relocalization of a correct image	89
4.3	Illustration of the relocalization of an incorrect image	90
4.4	Illustration of the R2D point set creation steps	92
4.5	Visual example of exploiting the level of detail in the R2D re-ranking process	95
4.6	Example of contradiction between the metadata information (location) and the automatically estimated visual similarity	97
4.7	Example of structuring based on visual similarities compared to a structuring metadata	97
4.8	Plot of distance weighting function $w_{i,j}$	99

5.1	Visual example of re-ranking methods combination	106
5.2	Workflow of re-ranking steps combinations	108
5.3	Example of artificial increase of provider entropy while keeping the same mAP	111
5.4	Comparison of provider entropy between How-A, How-A + RANSAC-SG, How-A + R3D-SG and How-A + R2D-SG	112
5.5	Comparison of provider entropy between How-A, How-A + RANSAC-SG and How-A + RANSAC-LG	112
5.6	Comparison of provider entropy between How-A + RANSAC-LG, How-A + RANSAC-SG + R3D-SG and How-A + Location weighting (Sp) + R3D-SG	113
6.1	Examples of scalable georeferenced 3D models (data from IGN). 1st row: CityGML LoD1 buildings (French "Ref3DNat" reference), available on the whole territory; 2nd row: Superposition with terrestrial LiDAR point cloud acquired by Stereopolis at the scale of the city.	124
6.2	Manual 2D spatialization of contents in the French national library's collaborative platform	125
6.3	Interactive selection of 2D-3D pairs of points (colored bullets) in the photograph and in the 3D scene modeled with LiDAR points, as input of a 6-DoF pose estimation tool (iTowns web application (Blettery et al., 2020)). In this example, we observe differences between the old photograph and the recent version of the scene (disappearance of the bridge, new buildings, roadway modification), which highlights the challenge of the points selection for the pose estimation (images from Musée Nicéphore Niepce and IGN).	126
6.4	Illustration of the influence of the intrinsic parameters on the pose estimation	127
6.5	Pose estimation by fusing relative poses as proposed by (Song et al., 2016) (figure from (Song et al., 2016))	129
6.6	Classes of methods for pose estimation (figure from (Humenberger et al., 2023))	130
6.7	Graph of a collection organized in the Oronce-Fine platform	132
6.8	Preview of a collection visualized via PixPlot	133
6.9	Preview of the WhatWasThere platform	133
6.10	Preview of the Remonter le temps platform, comparison of a 1950 ortho-photograph and a 1850s map of Paris' center	134
6.11	Preview of the Navilium platform, with a located image selected	135
6.12	Preview of the HistoryPin platform, with an image visualized in pseudo-3D in the scene	135
6.13	Preview of the Hist4D platform, where images are located in 6D within a tailored 3D model	136
6.14	Preview of the SmapShot platform, displaying an image in its 3D context .	137
6.15	ALEGORIA Project Search Engine, images from (Geniet et al., 2022) . . .	138
6.16	ALEGORIA project visualization platform, allowing for the visualization of multiple images at the same time (all "pyramids" correspond to an image), image from (Blettery et al., 2020)	139
6.17	The heritage content coarsely spatialized and visualized in the 3D scene of the <i>UD-Viz</i> platform from (Jaillot et al., 2021)	140
7.1	The 3D graph-based visualization platform	144
7.2	Minimal representation of the graph-based representation of the dataset . .	147
7.3	Detailed representation of the graph-based representation of the dataset . .	149

7.4	Visualization of the subgraphs obtained using different similarities	150
7.5	Overview of the semi-automatic structuring process	151
7.6	Overview of the platform's interface	155
7.7	Overview of the platform's menu	155
7.8	Overview of the nodes and links displayed	156
7.9	Overview of the platform's tools	156
7.10	Node information display	157
7.11	Grove extension macro buttons examples	158
7.12	Aggregated similarity links visualization	160
7.13	Example of cross-community link validation	161
7.14	Highly central node edge clearing	162
7.15	Spatial similarity links creation process aided by the tree representation . .	164
7.16	Unconnected nodes' reconnection process	165
7.17	Example of the location propagation process	169
7.18	Visual evaluation of the dataset's structure on the 5 first global links . . .	170
7.19	Visual evaluation of the dataset structure, using the 5 to 10 best global links	171

List of Tables

2.1	Summary of the providers of the dataset, their acronyms and how much they account for in the dataset	41
2.2	Summary of the 31 classes in the dataset and their size	41
3.1	Main network backbones for image retrieval	55
3.2	Image retrieval datasets usable for landmark retrieval applications	63
3.3	Image retrieval datasets' specific heterogeneity	63
3.4	mAP score of tested image descriptors	74
3.5	mAP provider vs provider with How-A descriptor	75
3.6	mAP provider vs provider with R101-GeM descriptor	75
3.7	mAP comparison for classes with specific small details	76
3.8	mAP comparison for classes with great visual similarity	76
3.9	mAP provider vs provider with How-A descriptor on Full dataset without distractors	78
3.10	mAP provider vs provider with CV-Net global descriptor on Full dataset without distractors	78
3.11	Modification of mAP score with re-ranking	79
3.12	Geometric verification on the 135 first similar images	81
3.13	Summary of the evaluated descriptors and re-ranking methods	82
5.1	Evaluation of the 3D reconstruction-based approach	102
5.2	Evaluation of the re-ranking method using a 2D pseudo-reconstruction	103
5.3	Evaluation of the location weighting re-ranking approach	104
5.4	mAP scores for multiple combinations of re-ranking steps	108
5.5	Evaluation of the impact of diffusion depending on provider's entropy	111
5.6	Mean computation time for each re-ranking strategy, including combinations, for $k = 135$ images	114
6.1	Overview on public image datasets dedicated to landmarks, exploited for training purposes or as spatialization reference (MMS stands for Mobile Mapping System)	123
7.1	mAP scores evolution through iterations combining automatic and manual linking	167
7.2	mAP improvements comparison against no re-ranking at all, depending on the number of re-ranked images and the computation time	168

Chapter 1

Introduction

1.1 Motivations	22
1.1.1 Various potential applications	22
1.1.2 A hindering lack of structure but several possible solutions . . .	23
1.2 Objectives and contributions	24
1.2.1 Context	27
1.3 Publications	27

From archiving purposes to immersive visualization via new research purposes, exploiting digital and digitized iconographic heritage contents could have numerous impactful applications. This thesis presents approaches furthering the structuring of such contents, promoting their use. This first chapter outlines the goals, challenges, motivations and contributions of our work.

The starting point of this thesis is the increased availability of digital or digitized iconographic heritage contents, opening new areas of research and offering new potential applications. This increase in digitization stems from two paradigms. On the one hand, GLAMs (Galleries, Libraries, Archives and Museums) exploit digitization for the purpose of conservation and exploitation of their contents. On the other hand, the French administration is leading a large and accelerating campaign to promote open data for disseminating and promoting public data, which further lead french GLAMs down the path of digitization for open access distribution.

However, this growing availability revealed that the specific characteristics of these contents prove to be challenging for a complete use of their potential. Indeed, due to their specific organization, structuring issues prevent a large scale usage. To alleviate those challenges, owners of these contents developed structuring standards and methods. However, those remain very collection-based, thus barely solving the large scale use issues.

From another perspective, large-scale indexing of image collections is a growing field for the past decade, that benefited greatly from technological advancements in terms of storage and computational capabilities, as well as the advancements in the field of computer vision, proposing efficient large-scale solutions for image indexing. However, those solutions are developed with rather recent contents and barely tested against the more difficult cases that are brought on by digitized iconographic heritage.

This thesis aims to bridge the gap between large-scale automatic image indexing meth-

ods and the specificities of iconographic heritage contents, to propose new structuring paradigms for iconographic heritage contents. More structure and interlinking of contents could further their large scale use for multiple goals. Those are what motivate our thesis and we detail them in the following section.

1.1 Motivations

The digitization process started by many GLAMs has two main purposes. On the one hand, this process ensures an obvious objective of digital archiving of all contents. On the other hand, many new applications could make use of those newly available contents. This section introduces some potential applications as well as the challenges hindering their deployment.

1.1.1 Various potential applications

The uses of iconographic heritage contents are multiple. The applications examples presented here are not exhaustive but rather highlight the multiplicity and diversity of potential uses, especially when the contents are digitized and linked together.

First, digital collections are much more accessible via web-based platforms. This allows GLAMs to reach more and more people, promoting their collections, their work, but also serving potential educative purposes. Indeed, as George Santayana said, "Those who cannot remember the past are condemned to repeat it". Thus, ensuring access to information on the past is paramount for history or geography teachers, students, but also the general public. Especially in our time in which visualization is key, allowing such contents to be readily available to students or to the media for instance becomes paramount. The French National Audiovisual Institute for instance has started to do such comparisons between past reactions and current reactions to a similar problematic.

From another perspective, further from the general public, the accessibility of digitized iconographic contents proves to be a new and rich source of data for multiple fields of research, especially in social sciences and humanities. Even though those contents were already present, their digitization furthered somewhat their accessibility, making it a more readily usable source of information. These new contents prove usable as illustrations, source or support of new theories within research projects (examples are the ALEGORIA project (ALEGORIA project, 2018) or the Archival City project (Archival City Project, 2019; Blettery et al., 2020)).

Furthermore, exploiting such iconographic heritage contents could also serve more current and pressing issues. Indeed, satellite data for instance is now widely use for change detection and land use monitoring. It allows to follow the territory's evolution throughout decades. Iconographic heritage in our thesis focuses more on street-level depictions of landmarks. They could be used for similar purposes of change detection and analysis. Indeed, whether for ecological studies or urban planning studies, knowing how an area was before and how it was affected by different types of human interventions

is immensely useful to plan new urban developments or to protect sensitive areas. "A picture is worth a thousand words" to convey an idea and can help understand what multiple written reports would tell in a less intuitive way. A similar field that could benefit from these contents is architectural refurbishments projects. Indeed, visualizing a building at its prime, in an immersive fashion, and comparing it to the actual scene should help promote new ideas to both do repairs and respect the original spirit of the building, especially in this time when renovation plans are more and more numerous.

Apart from the specific studies, visualization of the collections can also be performed in multiple, more or less intuitive and immersive ways. From simple browsing of images on the web to advanced visualization platforms using Augmented/Virtual Reality paradigms, the showcasing of contents can be pushed very far. Going further with virtual reality applications could lead to full virtual guided tour to "visit the past", making the past a more tangible object to visualize, explore, and analyze. The augmented reality paradigm could lead to actual guided tour with the possibility to visualize a place as it was fifty or a hundred years ago for instance. Those visualization paradigms could as well be applied to all types of modern studies like urban planning. Indeed, standing in front of a future urban development and seeing immersively how it was fifty years ago could foster new thinking and prevent planners from repeating past mistakes they may have overlooked "on paper".

Finally, even within the collections' own organization, the digitization process introduced a new way of organizing contents, leading to new structuring and added information to the contents, as links are inherently created as a new, digitized structuring is created. Furthermore, as new structuring and new content (digitally available ones) are available, more fields of study can exploit this data. Thus, combining multiple organizations, approaches and methods from different communities focusing on different aspects of the data can be beneficial for all communities, as explained by (Meinecke, 2022).

1.1.2 A hindering lack of structure but several possible solutions

As the potential applications are numerous and attractive, their implementation is often hindered by the lack of structure both intra-collection and inter-collections. Indeed, due to the silo-based organization of the collections, few links between collections are available. Furthermore, as each collection has its own specific organization and structuring, links are hard to create. Structuring within a collection is furthermore based solely on the choices made at creation time and can be hard to modify afterwards.

Having links between contents either within collections or between collections can serve multiple purposes. First of all, to easily query relevant contents, which allows for a more global use of all -rather sparse- data sources (that image collections often are). Linking contents also allows checking for incoherences between metadata associated to contents. Indeed, two images linked together as similar but not sharing the same address for instance may reveal an issue with the address of one of them. Furthermore, links between images also allow to propagate information between contents. Thus, if links

are created between a collection where all images depicting buildings have an address and another where all images depicting buildings have a construction date and architect name, then similar images can be provided with all three information at once.

Though structuring may be hard, linking contents within each collection or between collections could be based on several information. For instance, based on a localization information (two images at the same address probably depict the same object). It can also be based on semantic tags depicting what is shown on the images. It can also be based on automatically computed visual similarity scores. In our context of iconographic heritage, the metadata are often sparse and organized differently for each collection, making their use difficult for automatic linking. However, visual similarity is agnostic to all this, thus making it a potential solution for automatic linking.

Aware of the huge potential of exploiting iconographic heritage and the potential solutions to structure collections, we detail in the next section the contributions we propose to help interlinking and structuring collections of iconographic heritage contents.

1.2 Objectives and contributions

This section details our starting objectives and paradigms and the contributions they led us to.

Starting hypotheses

The main objective of our work is to improve the structuring between various heritage iconographic contents by leveraging existing automatic methods while remaining aware of the specificities of the data considered. This main objective can thus be subdivided into the following ones:

- Exploring the suitability of state-of-the-art large scale automatic image indexing approaches;
- Proposing new approaches more suited to the specificities of the data considered;
- Exploiting structures specific to the considered collections to enrich the interlinking of contents.

Due to the specificity of the data considered and the potential applications, the structuring must be performed with the greatest understanding of the methods possible. Indeed, interpretability of the approaches is paramount to understand why a method is suited for some specific contents and not others. That allows to adapt methods accordingly to the observed behavior. We will focus in our study on interpretable methods for whose the impact of the data on their performance is explainable.

Furthermore, as shown previously, iconographic heritage contents depictions are multiple. We focus our work on visual representations of buildings depicting Parisian heritage throughout the 20th century as presented in Chapter 2. Working with such objects allows

to exploit several types of link between contents, either visual or spatial, as each building is an immovable object. The linking of contents in our work may thus use information from either the depicted object (the building) and the depiction of the object (the iconographic content) but both should not be confused.

Outline and contributions

To address the objectives listed before, the main contributions of this thesis are detailed next. They all aim at improving structuring within and between image collections. The various proposed approaches do not attack the problem from the same angle, either at query level or at collection-level, but all take into account the specificity and challenges of the data considered. They are detailed in the outline of our thesis which consists of eight chapters with two main parts with our contributions:

Chapter 1: Introduction. This first chapter introduces the general context of our work. More specifically, it describes what motivates our research and outlines our contributions.

Chapter 2: Iconographic Heritage: a focus on Paris. Out of the multiple potential aspects of iconographic heritage, we choose to work on representations of buildings, both well-known and regular ones. This chapter further details the specificities of iconographic heritage contents and presents the dataset we gathered for the remainder of our experiments.

Part 1: Automatic Retrieval and Re-ranking. This part focuses on our work on automatic, content-based image retrieval methods, applied to our heritage dataset.

Chapter 3: Related Work and its Evaluation. This chapter details our **first contribution**. It is the review and evaluation of state-of-the-art approaches for content-based image retrieval faced with iconographic heritage. Both image descriptors and re-ranking approaches are evaluated with the dual objective of correctly linking similar images but also ensuring a strong inter-collection retrieval, the goal of our thesis being the automatic linking of contents between different collections. This led us to define which methods are suited for this data and which are not but also to explain why. Furthermore, it also provided interpretable insights that led to our second contribution.

Chapter 4: Our Contributions to Re-ranking. The **second contribution** consisting of three re-ranking methods alleviating the challenges brought by iconographic heritage on existing approaches are introduced in this chapter. For all of them, the idea is to exploit information at a more global scale than just the query itself. In our case, we exploit spatial information in two different contexts. Furthermore, one main focus of these methods is for them to remain interpretable as to how and why they work better with our data than existing approaches.

First, in a geometric query expansion setting, extending the classical geometric verification paradigm. The idea of our first two propositions is to exploit the geometric information gathered within the first retrieved images to extend the encoding of the scene’s geometry, thus evaluating the geometric coherence of the retrieved images not only against

the query’s geometry but against a larger geometry. Our two methods exploit for one a 3D reconstruction and for one an approximate 2D reconstruction of the scene.

Second, our third approach exploits the location information that is partially available with the images. The idea is to evaluate the coherence of the visual similarity automatically computed by weighing it against the spatial proximity between images. The idea being that far away images are less likely to represent the same scene. If this method exploits spatial information in our case, it could use any structuring information associated to the images.

Chapter 5: Re-ranking Strategies Evaluation. The proposed approaches are evaluated in this chapter. The most-suited paradigms of re-ranking are then defined and evaluated in depth to better understand which parameters are most important for automatic retrieval and re-ranking on iconographic heritage contents.

Part 2: Graph-based Semi-automatic Retrieval. The second part of our thesis focuses on a semi-automatic framework designed for structuring iconographic collections, leveraging the large-scale impact of automatic methods and the focused impact of expert knowledge.

Chapter 6: Structuring, Spatializing and Visualizing Iconographic Heritage. A review of methods designed for spatializing, visualizing and more globally structuring iconographic heritage is presented in this chapter. This chapter aims to ground our proposed framework within an ecosystem of platforms and approaches which all leverage specific aspects of iconographic heritage for structuring collections.

Chapter 7: Graph-based Semi-automatic Re-ranking. The **third and final contribution** of the thesis is developed in this chapter. It is a framework for semi-automatic visual validation and manual correction of a collection’s structuring at a global level. This framework first exploits the most suited automatic approaches evaluated in the first part of the thesis to automatically estimate the best possible structuring. The structured collection is then visualized in a graph-based fashion on a web-based visualization platform. Several visual clues are then introduced to guide the expert’s manual interventions on impacting areas. We show that, though semi-automatic approaches are costly in expert time, they can be optimized for the expert to intervene on the most important structuring issues that automatic methods can not solve. This impacting corrections then see their effects multiplied after diffusion throughout the entire structure. This framework feeds on the insights from the first part in terms of what is most-suited for re-ranking and structuring collections. It also provides a visual platform ideal for evaluating automatic approaches, furthering our objective of interpretability of the structuring.

Chapter 8: Conclusion. This final chapter summarizes our work and offers insights as to what the next step should be to further improve collection structuring in a more automated and large-scale fashion.

1.2.1 Context

The Ph.D. work presented in this thesis was part of a Cifre Ph.D. between the city of Paris and the LaSTIG laboratory of the French Mapping Agency. 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 and 2023-AD011013510R1).

Furthermore, collections from several GLAMs were exploited during this work, namely the city of Paris, the Médiathèque du Patrimoine et de la Photographie, the Musée Albert Kahn and the Cité de l'Architecture.

1.3 Publications

The work in this thesis led to the following publications:

International Conferences

1. Emile Blettery, Nelson Fernandes and Valérie Gouet-Brunet, "How to Spatialize Geographical Iconographic Heritage", Proceedings of the 3rd Workshop on Structuring and Understanding of Multimedia heritAge Contents, ACM Multimedia, Chengdu, China, 2021, 31-40.
2. Emile Blettery and Valérie Gouet-Brunet, "Re-ranking Image Retrieval in Challenging Geographical Iconographic Heritage Collections", 20th International Conference on Content-Based Multimedia Indexing, Orléans, France, 2023.

Under review Emile Blettery and Valérie Gouet-Brunet, "Heritage Iconographic Content Structuring: from Automatic Linking to Visual Validation", Journal on Computing and Cultural Heritage.

Chapter 2

Iconographic Heritage: a focus on Paris

2.1	Introduction	29
2.2	Iconographic heritage: a challenging object of study	30
2.2.1	Heterogeneous data	31
2.2.2	General organization and structure	33
2.3	A test dataset focusing on Paris	34
2.3.1	City of Paris' image data	35
2.3.2	Stereopolis dataset	36
2.3.3	Other providers	37
2.3.4	Summary of the final dataset	40
2.4	Conclusion	43

2.1 Introduction

The previous chapter introduces the motivations of this thesis and the challenges they face. Indeed, many of the characteristics of heritage contents are challenging for automatic approaches relying on visual content. This chapter first presents how iconographic heritage contents are challenging due to their visual aspect but also their organization in Section 2.2. To further study and evaluate this in the rest of the thesis, we design a specific dataset depicting Paris throughout the last century, multiplying variations in terms of collection of origin, color, viewpoint, level of detail, etc. This chapter extensively details this dataset and its specificities in Section 2.3.



Figure 2.1: Sample of images of iconographic heritage¹

2.2 Iconographic heritage: a challenging object of study

According to the Collins dictionary, a country's heritage is all the qualities, traditions, or features of life that have continued over many years and have been passed on from one generation to another. This heritage may be transmitted through the ages in various forms as listed by the UNESCO². Iconography is "the art of representing or illustrating by pictures, figures, images, etc.". Hence, iconographic heritage consists of all past visual representations of a society's way of life, culture, buildings, technological innovations and so on. It can represent cultural aspects (monuments of course as well as mundane places and scenes of life), but also natural or geographical landscapes, depicting scenes at different times in the past. Iconographic heritage describes a specific state at a specific time, based on the principle that iconography is a snapshot. Furthermore, even if a common view of heritage implies that it comes from a somewhat distant past, we assume that past photographs from even a year ago rapidly become part of iconographic heritage as they represent a dated cultural aspect of society in a quickly evolving world. As all iconography can be exploited for sociological and historical analysis, even recent pictures are a testimony of a less distant past and become integral part of the iconographic heritage.

Because it would be difficult to perform an extensive study for all types of contents related to iconographic heritage, in this thesis we choose to focus on one category of

¹From top to bottom and left to right: Internet, CC-BY-NC 2.0 ; IGN, Stereopolis ; Internet, CC-BY-NC-SA 2.0 ; Internet, CC-BY-NC 2.0 ; Internet, CC-BY-NC 2.0 ; Internet, CC-BY-NC 2.0 ; Internet, CC-BY-NC 2.0 ; IGN, Photothèque ; Archives Nationales, LAPIE ; Internet, CC-BY-SA 2.0 ; Internet, CC-BY-NC 2.0

²<http://www.unesco.org/new/en/culture/themes/illicit-trafficking-of-cultural-property/unesco-database-of-national-cultural-heritage-laws/frequently-asked-questions/definition-of-the-cultural-heritage/>

iconographic heritage, related to all the visual representations of geographical landmarks, as illustrated in Figure 2.1 with various aerial and terrestrial iconographic representations of landmarks.

From one side, this category of contents is extremely widespread because accounting for a large portion of what people have liked to capture through drawings or amateur or professional photographs. This leads to a huge visual testimony of our environment that can benefit use cases and applications, ranging from historical and sociological studies up to mobile mapping scenarios, through digital tourism, landscape ecology or remote sensing. And from the other side, the visual representations associated with these objects of interest are extremely diverse given the various acquisition conditions and the evolution of landmarks over time.

With the development of powerful scanning tools and the availability of storage and sharing infrastructures, increasing digitized or digital data are made available. In this thesis, we focus on iconographic cultural heritage data which proves in its heterogeneous aspect (Section 2.2.1) and its specific organization and structuring (Section 2.2.2) to be of specific scientific interest as their analysis, *i.e.* their description, comparison, learning, indexing and retrieval are complex and challenging.

2.2.1 Heterogeneous data

Various data types. When it comes to iconographic cultural heritage data, even so the digitized format may be similar between images, the original medium used for the depiction transfers its specificity to the digital image. Indeed, whether it is a painting, an historical map, a postcard, or even printed-then-digitized photographs, the final image content will have features specific to its original medium. Hence, iconographic cultural heritage data cannot be apprehended as a single type of object but as a constellation of objects, with both specificities and similarities when compared to one another. In particular, this multiplicity of representations and data types leads to a large variety of visual characteristics (color, texture, shapes, local patterns, etc.) that may be hard to link together using off-the-shelf automatic indexing methods.

Increasing number of objects and representations. For the last decade, GLAMs (Galleries, Libraries, Archives and Museums) have increased their digitization processes, mostly as a new mean of conservation but also to render their collections available on a dedicated platform.

Examples of those platforms are the Terra³ media library of the French Ministry of Ecological Transition and Territorial Cohesion but also the Europeana Collections⁴ regrouping millions of digitized heritage items. Platforms like these allow the user to browse through the collection using more or less precise search conditions (location, date, theme, etc.). This "dive into the past" has made available many more representations of the same objects or places (see Figure 2.2) and revealed depictions of old places that may

³<https://terra.developpement-durable.gouv.fr/>

⁴<https://www.europeana.eu/>

now have disappeared. This multiplicity of depictions, on various media, at several dates, from different points of view intensifies the heterogeneity of the images available.

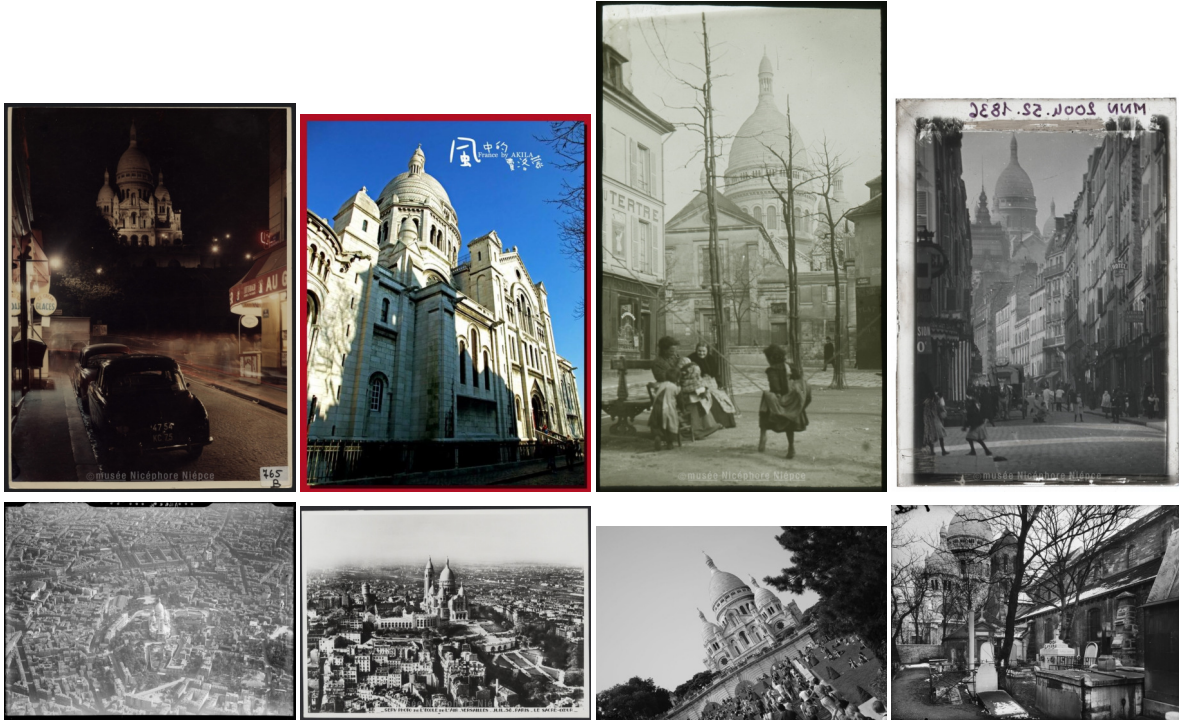


Figure 2.2: Sacré-Cœur Basilica, at different times, from different perspectives using various media⁵

Heterogeneous metadata. The nebulous and heterogeneous aspect of iconographic heritage data is also heightened due to the variability between associated metadata. Indeed, metadata are linked to the image depending on many factors: the considered collection, the original medium, the curator of the collection, the information provided, the application and audience targeted, etc. Hence, for instance, if the geographical location depicted is considered, the level of metadata may be very different. From a simple description of the main object of the depiction to a precise address or even a precise 2D, 3D or 6D location (usually provided by mapping agencies), the level of detail is very disparate. Furthermore, the formatting of the information is very collection-dependent. For an address, you can have a single sentence containing the full address or three separate entries for the street number, the street type, and the street name. Furthermore, as we are on the topic of toponyms, as time passes, place names are modified to suit political needs of because of city evolution (for instance Paris during the XIXth). Saint-Petersburg for example has been renamed three times along the XXth. Hence, addresses associated with images at a certain time may not correspond to current addresses if correspondences between passed and current addresses are not known. The variety of combinations and representations deepens even further the heterogeneity of the available data, which com-

⁵From top to bottom and left to right: Musée Nicéphore Niepce, Internet ; Internet, CC-BY-NC 2.0 ; Musée Nicéphore Niepce, Internet ; Musée Nicéphore Niepce, Internet ; IGN, Photothèque ; Musée Nicéphore Niepce, Internet ; Internet, CC-BY-NC-SA 2.0 ; Ville de Paris, Edouard Desprez / DHAAP / Roger-Viollet

binéd with its specific organization makes requesting these contents harder than modern datasets.

Metadata can also be structured in various models according to the needs and practices of the different institutions in charge of the funds, which requires a prior harmonization step to be fully exploitable together. Following the recommendations of the Web of Data best practices for representing metadata, it is generally conventional to adopt a graphical data model (RDF), accompanied by common and linked ontologies and repositories to describe cultural objects (e.g. the ICA RiC-O international standard⁶). Alongside this model, it is considered good practice to extend the publication principles in accordance with the FAIR data principles⁷.

2.2.2 General organization and structure

In silo structuring. Cultural heritage data is traditionally preserved, organized, and displayed by GLAMs. This physical distribution of the data has limited interactions between collections, inducing this "in silo" organization. Generally, this structuring model sees every collection as an independent, self-sufficient entity. The connections are very sparse between them, the similarity and complementarity across collections being more or less identified. Furthermore, as described in the previous section, metadata associated to the iconography may be very collection-dependent. Indeed, metadata selection and organization reflects the representation each institution has of its collections. This can be influenced by institutional tradition, the date of creation of the collection, etc. This collection-specific organization explains the difficulty to link collections between each other but interlinking within a single collection may also prove difficult.

Traditional manual indexing. Structuring collections using indexes has mostly been done manually, even with the help of IT tools. These manual methods require laborious work when it comes to updating indexes, for instance, to add new information or to add a new object to the collection. Manual indexing requires expert knowledge (hence the necessary employment of an archivist), but this knowledge is difficult to pass on from one generation of archivists to the next. The consistency of the indexing process and the indexes becomes one of the main obstacle when it comes to linking objects within a collection or with other collections. To alleviate this obstacle, the folksonomy (a spontaneous decentralized collaborative annotation system, based on indexing by non-specialists) allows for a quick annotation of numerous objects by multiple users using tags which in the end create a classification and interlink contents. However, this bottom-up strategy offers a quality that may be inferior to a taxonomy (top-down strategy, classification made by the owners of the contents) and this may cause issues. Thus, this solution is often incompatible for professional applications and remains for general public applications (YouTube for instance).

Weak intra-linking and interlinking. The "Linked Data" initiative (Berners-Lee,

⁶RiC-O standard: <https://www.ica.org/standards/RiC/ontology>

⁷FAIR principles: <https://www.go-fair.org/fair-principles/>

2006) encourages the publication of structured data on the Web, not as isolated silos of data, but by linking them together to form a global network of information, helping sharing and querying information between machines, regardless of where it is stored. In order to gain visibility and to make their data exploitable in other contexts, GLAMs have started participating in this organization, mainly through the assignment of URI (Uniform Resource Identifier) to collection items, relying on metadata. To go deeper, this kind of initiative should also involve techniques relying on the content itself, to overcome the potential problems associated with metadata (annotation cost, lack of metadata, metadata not standardized, too specialized, etc.), and CBIR has a great role to play in such initiative. Several research projects aim to alleviate those limitations: an example is the Indian Ocean iconographic heritage network (Indian Ocean iconographic heritage network). Its objective is first to improve digitization of the iconographic heritage of countries of the Indian Ocean. The second objective is the sharing of resources and expertise and the development of common tools to "create cultural content in a digital form throughout the Indian Ocean area". From a more automated point of view, the ALEGORIA project (ALEGORIA project, 2018) for instance works on automatically interlinking and exploiting French institutional funds. Its consortium regroups ICT and social sciences research laboratories, archives and museums. The idea being for the iconographic heritage content to be a material for research, the product of the research in turn serving the better linking and exploitation of the iconographic heritage. The existence of those projects and the collaboration of the different actors reflect the realization by all actors of the pitfall that this weak linking represents. The iconographic heritage and all its challenges appear as a scientific object both challenging and full of promises.

To evaluate how these specificities impact the newly developed automatic approaches dedicated to interlinking of image contents, we design and present in the following section a specific dataset illustrating those specificities.

2.3 A test dataset focusing on Paris

As presented in the previous section, iconographic heritage is extremely diverse and for this thesis, we focus on a specific iconography. We study the city of Paris, from 1900 onwards. We more specifically exploit images depicting Parisian architecture mostly taken at street-level, depicting both known landmarks and classical facades, a sample is shown in Figure 2.3. This dataset combines multiple collections from various GLAMs, all with specific characteristics. We first present our two starting collections, the city of Paris' one in Section 2.3.1 and the Stereopolis one in Section 2.3.2. We then present the other providers in Section 2.3.3 and summarize the dataset in Section 2.3.4. This dataset depicting a certain type of iconographic heritage but reflecting all the diversity and heterogeneity of the contents will then be used as a support for evaluating existing automatic approaches and guide the proposition of new, more suited approaches.



Figure 2.3: Sample of images in our dataset⁸

2.3.1 City of Paris’ image data

One of the two starting datasets was created by *La Parisienne de Photographie*, for digitization and promotion purposes of the Roger-Viollet photographic archives agency’s collections, owned by the city of Paris and now archived by the Department of Architectural History and Archeology (DHAAP). The whole dataset comprises more than 13 691 photographs, depicting Paris throughout the 20th century between 1910 and 1979. The dataset depicts various aspects of Paris, from photographs of paintings or city plans to photographs of life scenes, via shots of archeological digs and findings. Most importantly, it depicts the parisian architecture in ample details, showing the variability of Parisian buildings and the evolution of the city through the century. In total, the architectural part of the dataset consists of 8390 images. Examples of the dataset are shown in Figure 2.4.

The main specificity of this dataset is its sparsity, both spatially and temporally. Indeed, a specific building can be shot from the same angle every two years, whereas another can be shot from ten different angles twice thirty years apart. This sparsity increases the variability in terms of visual aspect, as both photography techniques and Paris evolved throughout the years.

Alongside those images, an image caption is almost always available. It may be a textual description of the shot, an address (more or less detailed), a date of acquisition (more or less precise). Indeed, the information linked to the image were manually added by the photograph right after the acquisition or much later by archivists in an effort to add information to the shots, which have resulted in errors and imprecisions.

Exploiting automatic, content-based indexation appears like a solution to detect obvious errors in the metadata-linking process but also to enrich the existing metadata between images depicting the same buildings. Indeed, if one has an address while another has not, the address could be propagated. On the contrary, if they depict the exact same building while not having the same address, it reveals a mistake that could be manually

⁸From 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

or automatically corrected.



Figure 2.4: Examples of images from the Parisienne de Photographie

2.3.2 Stereopolis dataset

The second starting dataset for our study is called Stereopolis (Paparoditis et al., 2014). It is a complete and systematic acquisition of all streets of Paris in 2015 using a mobile mapping system. Similar to other mobile mapping datasets such as RobotCar (Maddern et al., 2017) or Kitti (Geiger et al., 2013), it consists of both photographs taken every three meters and a complete 3D point cloud of the whole city. Images are taken all around and above the car, as shown in Figure 2.5.

Exploiting this data has multiple benefits. First, contrary to the "Parisienne de Photographie" data, there is no sparsity in the data. Indeed, the whole city is mapped, which provides us with a full ground truth at a specific date. Second, the metadata associated are certain, especially the very useful location information, thanks to the mobile mapping approach that provides us a 6-degrees of freedom pose for each image.

Thus, our hope of exploiting sparse heritage data in conjunction with this complete "picture" of Paris at a certain date is to bring the structure and certainty of the recent and systematic dataset into the heritage one.



(a) Fontaine Saint-Michel



(b) Musée d'Orsay



(c) Rue des Quatre-Fils

Figure 2.5: Example of images from the Stereopolis dataset. © IGN

2.3.3 Other providers

To further increase the multi-provider aspect of the dataset, which is essential to evaluate the efficiency of the whole process for interconnecting multiple collections, we have decided to use contents from six other providers described here.

Paris 6K dataset

The first dataset exploited is a well known public benchmark in the field of content-based image retrieval called Paris 6K (Philbin et al., 2008). We more specifically exploit the revisited version (Radenovic et al., 2018). It consists of images of well-known monuments and buildings in Paris taken from Flickr. The quality, resolution, color and viewpoint of the images are very diverse, even though the images are quite recent (after 2000). Furthermore, some images are from an aerial perspective, which we have kept as they could display multiple known buildings, presenting a certain challenge for the evaluated and proposed methods. Figure 2.6 represents some images kept in the proposed dataset.



(a) Dôme des Invalides



(b) Le Louvre



(c) Panthéon

Figure 2.6: Example of images from the Paris 6K public benchmark. © Flickr

Médiathèque du Patrimoine et de la Photographie

A second database we obtained images from is the Memoire database with images mainly from the Médiathèque du Patrimoine et de la Photographie⁹. The photographs come from either the Historical Monuments collections or the State collections, displaying more or less known monuments or regular buildings, as illustrated in Figure 2.7. More than 20 000 photographs of Paris can be found, however, not all can be used in the thesis as they depict interiors, tiny details or archeological digs.

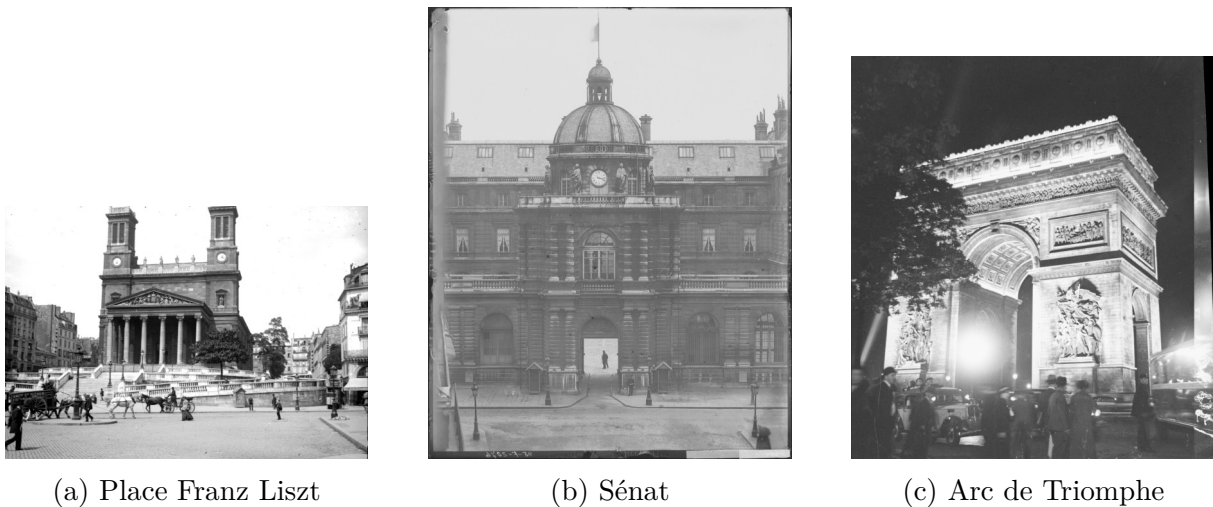


Figure 2.7: Example of images from the Médiathèque du Patrimoine et de la Photographie.
© Médiathèque du patrimoine et de la photographie

Musée Albert Kahn¹⁰

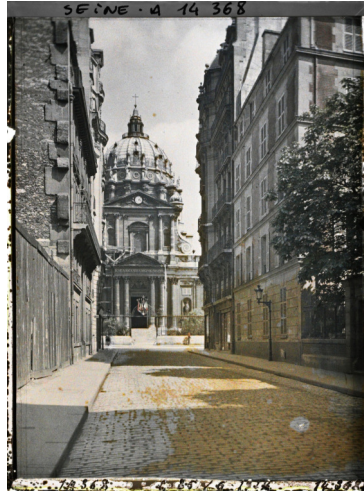
A third collection where we gathered images are the Planet’s Archives, a series of photographs taken by Albert Kahn in order to document the world and visually inventory the transformations of his time. A certain part of this collection (examples in Figure 2.8) concerns Paris with more than 4 000 photographs. However, all the images of this collection are not available at full resolution, hence part of the data we gathered is used at full resolution while some images are used at a lower resolution.

⁹<https://mediatheque-patrimoine.culture.gouv.fr/>

¹⁰<https://albert-kahn.hauts-de-seine.fr/les-collections/presentation/photographies-et-films/les-archives-de-la-planete>



(a) Tour Eiffel



(b) Eglise du Val de Grâce



(c) Place de la Bastille

Figure 2.8: Example of images from the Musée Albert Kahn's collection. © Musée Albert-Kahn, Archives de la Planète

Cité de l'Architecture et du Patrimoine

The fourth GLAM we asked data from is the Cité de l'Architecture et du Patrimoine¹¹. Its focus is more on the parisian monuments as its objective is to promote french architecture throughout the world. Thus, most images there concern known monuments with the specificity of finding photographs of mock-ups and work drawings for those monuments. Examples of images from this collection are represented in Figure 2.9.



(a) Place de la Concorde



(b) Opéra Garnier



(c) Gare du Nord

Figure 2.9: Example of images from the Cité de l'Architecture et du Patrimoine. © Cité de l'architecture et du patrimoine

Commission du Vieux Paris

The "Commission du Vieux Paris" data consists of multiple data sources including photographs (old or recent) gathered for evaluating the possibility of construction work (more specifically demolition works) with regard to its objectives of heritage preservation, thus using various viewpoints and detail levels, as presented in Figure 2.10. Also gathered by the Department of Architectural History and Archeology (DHAAP) of the city of Paris, this data is organized on an address basis, which makes it a quite certain collection (metadata-wise) but also an easily queried resource.

¹¹<https://www.citedelarchitecture.fr/fr>



(a) Avenue Rapp



(b) Panthéon



(c) Cathédrale Notre-Dame

Figure 2.10: Example of images from the Commission du Vieux Paris. © DHAAP / CVP

Conservation des Œuvres d’Art Religieuses et Civiles

Partner of the DHAAP, the Conservation des Œuvres d’Art Religieuses et Civiles of the city of Paris aims at listing and protecting more specifically the monumental heritage of Paris, from fountains to churches via statues. We exploited their collections more specifically for their photographs of churches which are otherwise quite lacking in more general collections of Paris and which we show in Figure 2.11.



(a) Basilique du Sacré-Coeur
© Ville de Paris, COARC/
Emmanuel Michot



(b) Eglise Saint-Sulpice
© Ville de Paris, COARC/
Jean-Marc Moser



(c) Eglise de la Madeleine
© Ville de Paris, COARC

Figure 2.11: Example of images from the Conservation des Œuvres d’Art Religieuses et Civiles of the city of Paris

2.3.4 Summary of the final dataset

Exploiting data from all the collections, which are summarized in Table 2.1, we assembled in total a dataset of 1,637 images of which an example was previously shown in Figure 2.3, 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). These classes are described in Table 2.2. Classes in **red** depict highly known monuments. Those in **orange** depict lesser known monuments of Paris. **Green** classes depict Parisian churches and **gray** classes represent classic Parisian buildings. 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.

Acronym	Full Name	Nb of images	% of dataset
Sp	Stereopolis	537	32.8
PdP	Parisienne de Photographie	193	11.8
MAP	Médiathèque du Patrimoine et de la Photographie	276	16.9
CVP	Commission du Vieux Paris	153	9.3
AK_HD	Albert Kahn - Haute Définition	214	13.1
AK_BD	Albert Kahn - Basse Définition	47	2.9
CA	Cité de l'Architecture	25	1.5
COARC	Conservation des Œuvres d'Art Religieuses et Civiles	30	1.8
P6K	Paris 6K	162	9.9
Total		1637	100

Table 2.1: Summary of the providers of the dataset, their acronyms and how much they account for in the dataset

Name and count			
Assemblée Nationale	26	Avenue Jean Jaurès	28
Avenue Rapp	35	Place de la Bastille	50
Place de la Concorde	71	Tour Eiffel	135
Fontaine Saint-Michel	27	Gare de l'Est	39
Gare de Lyon	37	Gare du Nord	38
Gare Saint-Lazare	31	Hôtel Hérouet	32
Dôme des Invalides	74	Le Louvre (place du Carrousel)	88
Eglise de la Madeleine	81	Musée d'Orsay	67
Place de la Nation	39	Cathédrale Notre-Dame	120
Opéra Garnier	63	Panthéon	89
Pharmacie (rue Saint-Honoré)	27	Place Franz Liszt	32
Rue de Lille	18	Rue de l'Université	18
Rue des Quatre-Fils	17	Rue Linne	15
Basilique du Sacré-Coeur	75	Eglise Saint-Sulpice	64
Sénat	24	Arc de Triomphe	119
Eglise du Val de Grâce	58		

Table 2.2: Summary of the classes in the dataset and their size. Legend: **highly known monuments**, **lesser known monuments**, **Parisian churches** and classic Parisian buildings.

Due to the variety of the classes, the number of images and providers per class can vary greatly. However, in every class both starting datasets must be present. Statistics on the class sizes and providers present are displayed in Figures 2.12, 2.13 and 2.14.

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, some 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 the 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).

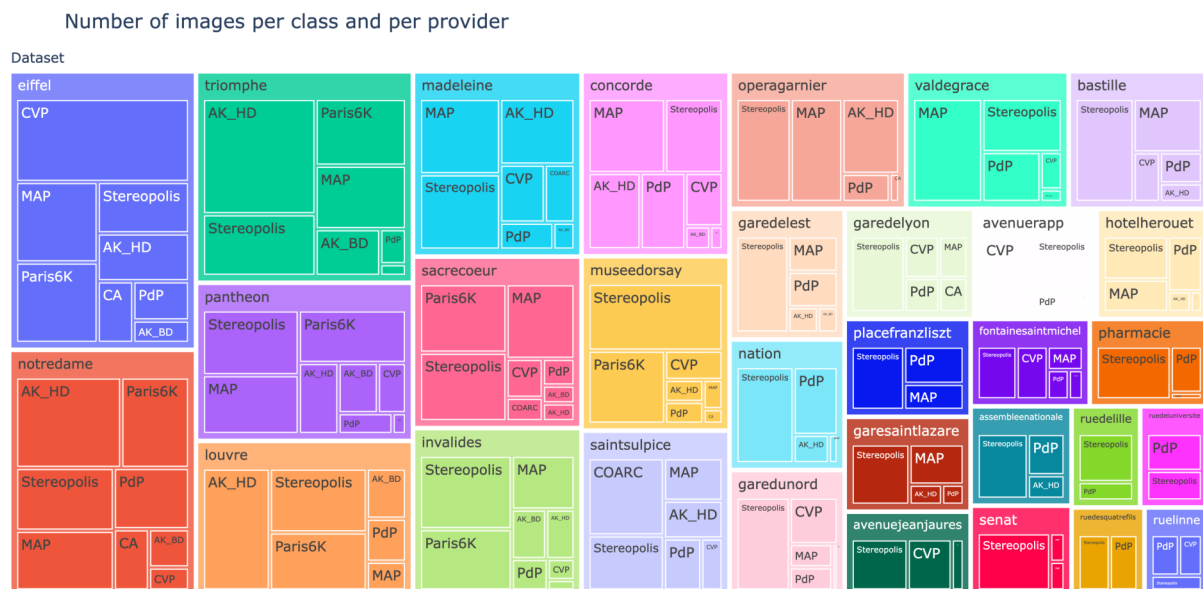


Figure 2.12: Dataset statistical representation based on classes

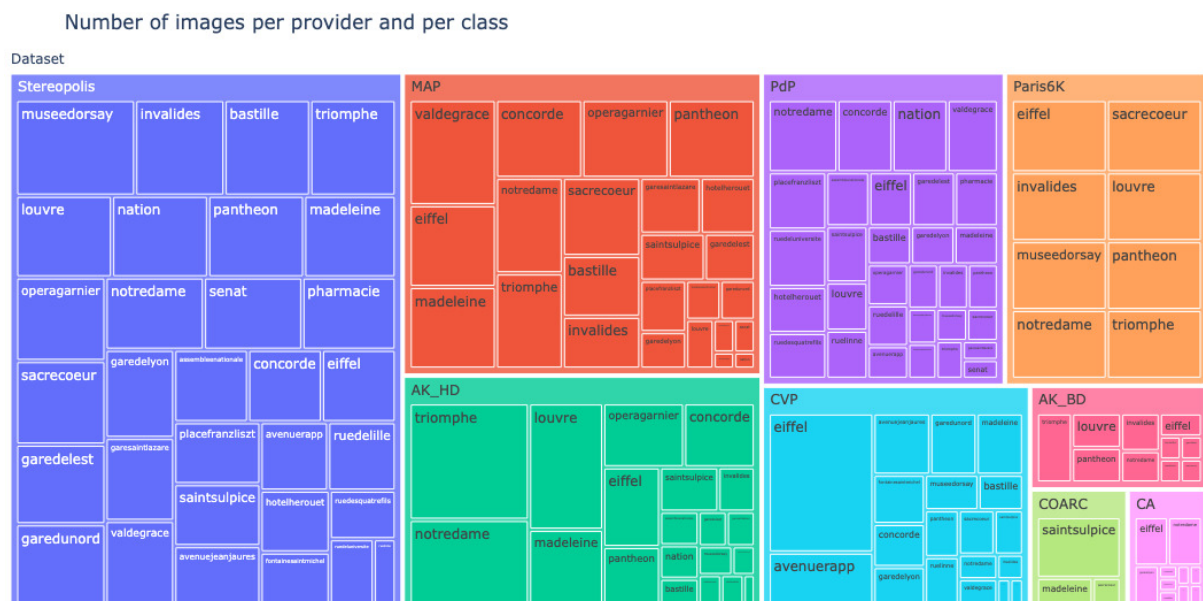


Figure 2.13: Dataset statistical representation based on providers

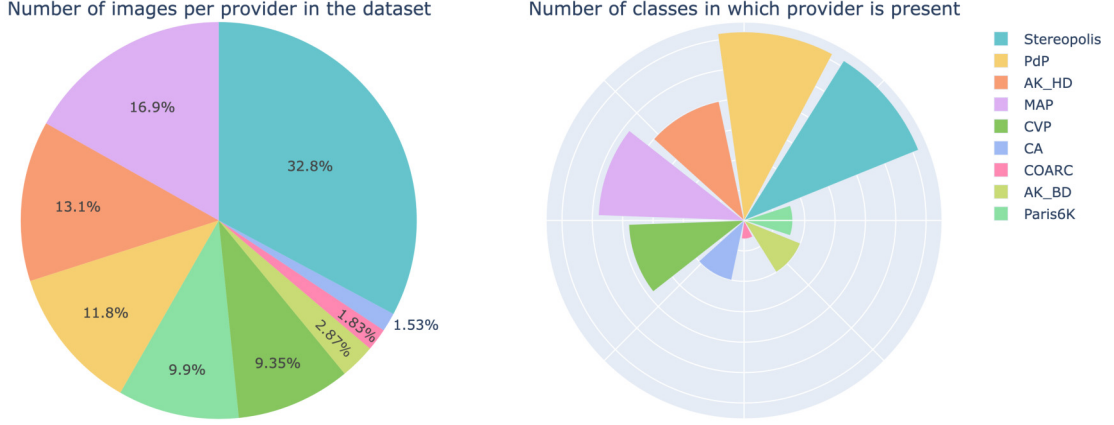


Figure 2.14: Dataset global statistical representation

2.4 Conclusion

This chapter first presents the specificities of iconographic heritage contents, both in terms of visual representations and in terms of organization. These specificities are prone to create challenges for existing state-of-the-art automatic image retrieval approaches. To evaluate such approaches and their suitability for large-scale use with iconographic heritage contents, but also to help formulate and assess our proposed methods, we design a dataset aggregating contents from multiple heritage collections.

This chapter thus introduces the proposed heritage iconographic content dataset depicting Paris throughout the last century, from 1910 to 2015. Due to the multiplicity of collections, the visual heterogeneity and variability is high, making this dataset one prone to challenge state-of-the-art methods of content-based image retrieval, as presented in Part I and more specifically in Chapter 3. Furthermore, this dataset will be used throughout the thesis to evaluate all newly proposed more suited approaches.

Part I

Automatic Retrieval and Re-ranking

Chapter 3

Related Work and its Evaluation

3.1	Introduction	47
3.2	Image retrieval	49
3.2.1	Hand-crafted descriptors	49
3.2.2	Learned descriptors	50
3.2.3	Similarity search	56
3.2.4	Spotlight on How and ASMK	58
3.2.5	CBIR and iconographic heritage	60
3.3	Re-ranking	64
3.3.1	Late fusion	65
3.3.2	Geometric verification	65
3.3.3	Transformers-based re-ranking	67
3.3.4	Query expansion	68
3.4	Image descriptors evaluation	72
3.4.1	Descriptors evaluated and evaluation choices	72
3.4.2	Choice between How-A, R101-GeM and DELG	74
3.4.3	Choice between How-A and CV-Net	77
3.5	Re-ranking methods evaluation	78
3.5.1	Re-ranking choices	78
3.5.2	Aggregation	79
3.5.3	Pseudo-relevance feedback	80
3.5.4	Transformers-based	80
3.5.5	Geometric verification	80
3.5.6	Diffusion methods	81
3.6	Conclusion	82

3.1 Introduction

Since the 90s, there exist a large panel of approaches for the description, matching and indexing of visual contents (Veltkamp and Tanase, 1999; Dharani and Aroquiaraj, 2013; Zhou et al., 2017; Chen et al., 2021). They can be classified according to the contents con-

sidered, even if today’s machine learning techniques, associated with dedicated training datasets, tend to reduce this difference. Specific datasets (*e.g.* fingerprints, faces, etc.) usually exploit dedicated techniques of description, while generic visual contents, including landmarks, are based on more generic approaches such as global descriptors (color, texture and shape) or local ones (points of interest, blobs, regions, etc.). Because of the variability of the content encountered in iconographic heritage, we address here such generic approaches, which gathers a very large panorama of techniques of description. Note that some of the references considered in the following concern the domain of RSIR, *i.e.* Remote Sensing Image Retrieval: this is a fast-growing research field where contents at large scale (large image datasets and/or high-resolution images) highly benefit from CBIR for retrieval as well as classification tasks. It mainly concerns satellite or vertical aerial imagery, sometimes with dedicated modalities (*e.g.* multispectral, hyperspectral, SAR imagery) but some approaches share characteristics with those of iconographic heritage dedicated to landmarks, especially when considering multi-temporal imagery (Li et al., 2021).

To easily present the global process of Content-based Image retrieval, Figure 3.1 proposes an overview of the whole process. A first step of image retrieval is performed by comparing image descriptors between that of a query and those of the images in the database. Images are then ranked in a decreasing similarity order. A second step of re-ranking can then be performed, using a variety of methods and based on the first retrieved images, to re-order the images according to another similarity criterion, improving the retrieval results.

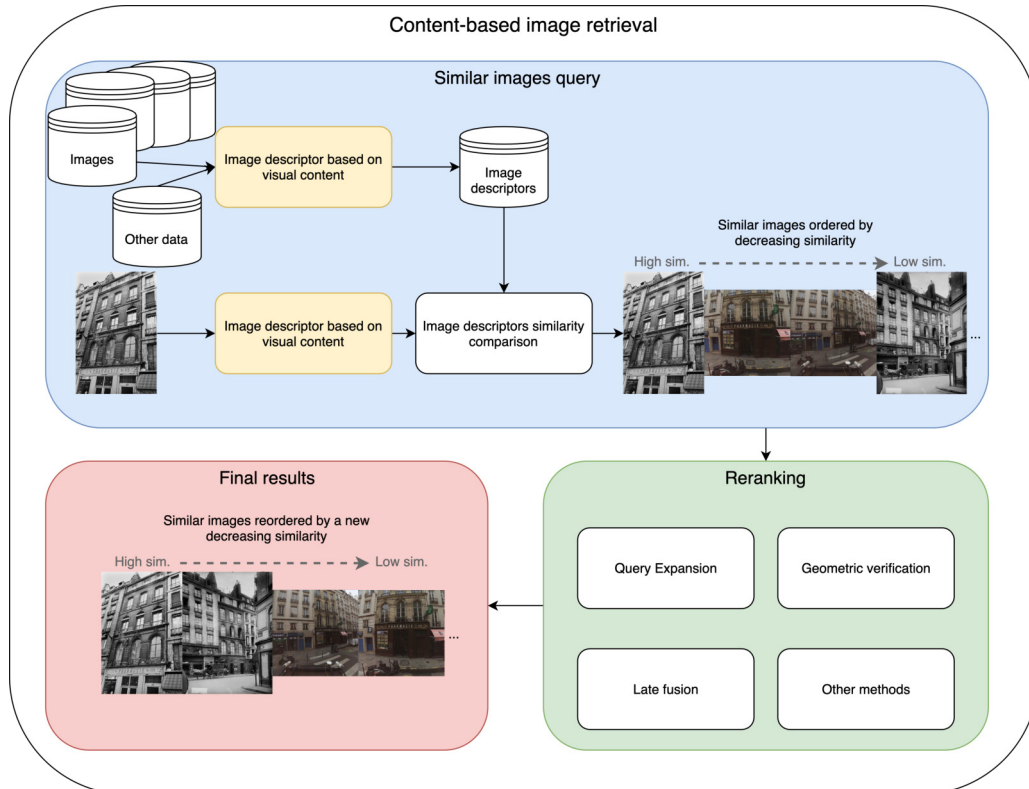


Figure 3.1: Global Content-Based Image retrieval pipeline

Due to its specificities, and mostly its large visual heterogeneity, iconographic heritage makes it very challenging for image descriptors to be efficient. Furthermore, their low accessibility at large-scale has precluded them from being used in benchmarks or training datasets, thus limiting the new methods from being suited to this specific type of data. Existing methods can thus be applied to heritage contents but with poorer performance.

This chapter presents methods for CBIR and their specificities, first for simple image retrieval and second for re-ranking, respectively in Sections 3.2 and 3.3. Some of those methods thought most-suited for our problem are then tested to estimate the ones most adapted to the dataset we gathered and its specificities, in Section 3.4 and Section 3.5.

3.2 Image retrieval

As presented in Figure 3.1, the first part of any image retrieval process is the creation of image descriptors that can be compared to evaluate the similarity between two images in order to be able to order images based on this similarity measure. In this section we will thus develop a state of the art on descriptor extraction and similarity measurement between images based on their visual content. First, Section 3.2.1 presents handcrafted methods for image description. Second, Section 3.2.2 introduces new learned methods for descriptor extraction with their specific paradigms, networks and training datasets. Then, Section 3.2.3 explains the specificities of the similarity search between descriptors. 3.2.4 details further the descriptor that will prove essential for our work. And finally, Section 3.2.5 presents specific image retrieval methods applied to some types of iconographic heritage.

3.2.1 Hand-crafted descriptors

Content-based image retrieval methods leaned on various handcrafted image descriptors based on different paradigms, the main one being the scale at which they are computed : either at a global scale or by aggregating local features.

Global features. A first category of handcrafted image descriptors uses the image in its entirety to estimate a single image signature. It can use color, texture or shape information or even combine them. A simple color-based descriptor is the histogram, however lacking in efficiency as two images can have identical histograms and yet represent two completely different scenes. Texture-based methods are for instance the GIST feature (Oliva and Torralba, 2001) of the Fourier transform, which can also be used to describe the shape (outlines) of the images' objects. Those global descriptions may be efficient when it comes to memory use for instance, however, they lack in invariance and robustness to transformations like rotation or illumination changes. Hence, using more robust local features and aggregating them has proven to be a successful alternative to global features.

Aggregated local features. An alternative to global features are local features, which prove to be more robust to various variations appearing on the images. Using local features to describe the image can be decomposed in three steps : first, detecting interest

points (or regions), second, describing those interest points, third, computing the image signature. Multiple methods exist to detect and describe (and sometimes both at the same time) the interest points. As detectors for instance, one can mention the famous SIFT (Lowe, 2004) and its adaptation SURF (Bay et al., 2008), but also ORB (Rublee et al., 2011), BRISK (Leutenegger et al., 2011) or Hessian-affine (Mikolajczyk and Schmid, 2004). As for descriptors, one can find once again SIFT and its adaptation Root-SIFT (Arandjelovic and Zisserman, 2012), and SURF, ORB but also BRIEF (Calonder et al., 2010). Detectors and descriptors can then be combined to achieve the most accurate, robust and invariant description. To establish a global signature for the image using those local features, two ways can be followed to aggregate the local information. First, using a sparse representation aggregating all interest points. However, as those representations can rapidly increase the memory use and complexity of the description, for large volumes of images, other means of aggregating the features using "bag-of-features" have been developed. Indeed, since (Sivic and Zisserman, 2003), quantization methods have been widely adopted in image retrieval. Local features in an image are seen as words in a text. Hence, using a "visual dictionary" (built using clustered local features extracted from a large visual database), local features are assigned to the nearest visual word and then the frequency of each visual word in the image becomes the image's signature. Several methods have then improved the aggregation or the assignment process like the Hamming embedding (Jegou et al., 2008), Fisher Vectors (FV) (Perronnin et al., 2010) or the Vectors of Locally Aggregated Descriptors (VLAD) (Jégou et al., 2011). New learning-based methods of aggregation have also been developed like (Passalis and Tefas, 2017).

3.2.2 Learned descriptors

Since AlexNet (Krizhevsky et al., 2012) at the ImageNet challenge in 2012, Deep Neural Networks have been a source of interest and successes when applied to the field of CBIR (Content-Based Image Retrieval) because they can compute powerful representations of images, especially with Convolutional Neural Networks first and later with Vision Transformers models. Multiple surveys have inventoried existing models based on different perspectives. (Piasco et al., 2018) describes the methods used in Visual-Based Localization, (Masone and Caputo, 2021) focuses on its applications regarding place recognition, whereas (Chen et al., 2021) and (Dubey, 2022) focus in more detail on the models' implementation. However, all surveys agree to distinguish three steps: the encoding of the representation, the similarity search between the query image and the database images, and finally the post-processing refinement of the results (which will be addressed later in Section 3.3).

The first step towards CBIR is the computation of the representation of the images (both the database images and later the query one) by the network. To this end, two strategies are used in the literature: using off-the-shelf models and applying them directly to the desired dataset or fine-tuning models to adapt them to the specificity of the dataset.

3.2.2.1 Off-the-shelf models and feature enhancement

Using off-the-shelf models means exploiting models trained for a specific task (mostly classification) and applying them without retraining to solve a different task (in our case an image retrieval task). In this case of off-the-shelf use, weights and parameters are used as is and not recomputed. This choice implies a wider domain shift, hence, results potentially disappointing as the network’s representations are not initially suited for image retrieval purposes. Multiple networks used as backbones in image retrieval tasks are presented further in Table 3.1.

To reduce the gap, the main contributions that can be applied to the results of these models are feature enhancement. The idea is to transform or combine extracted representations to obtain a new representation more suited to the problem of image retrieval. Hence, using this method, two steps are essential : extracting the features and enhancing them.

Feature extraction. Feature extraction process is impacted by several aspects of the model’s architecture.

- *First*, the data passed through the network. Indeed, it can be the whole image at once in a single pass forward or multiple passes forward using each time different patches of the image before aggregating the features computed. The strategy to choose those patches also impacts the features extracted. It can use a rigid grid, pyramid modeling, dense patch sampling or even use region proposal networks (Ren et al., 2017; Kong et al., 2016) to select patches.
- *Second*, the extraction layer. It can be either a fully connected layer or a convolutional layer. The fully connected layer is an intuitive solution that has been pursued (Jun et al., 2019; Song et al., 2017), but it has obvious limits: as each neuron is connected to all previous neurons, spatial information is largely lost, as well as local geometric invariance. Using a convolutional layer preserves the local information, as the receptive fields are smaller, propagating more structural information. Local detectors extracted from convolutional layers are used in image retrieval. One can mention SPoC (Babenko and Lempitsky, 2015) or (Ng et al., 2015) that uses VLAD instead of BoW and thus paves the way for NetVLAD (Arandjelovic et al., 2018).
- *Third*, the level of fusion of the extracted features if need be. Indeed, fusing features within a feature extractor aims at combining the various characteristics to enhance the feature specificity. This fusion process can happen between layers within a single model. Fusing multiple fully connected layers concatenates global features to enrich the combined global feature (Yu et al., 2018). As RSIR applications mostly require better features than low-level ones, (Zhou et al., 2015) use an auto-encoder to fuse low-level features into a sparse middle-level feature better suited for image retrieval purposes. The fusion can also concatenate features from both fully connected and convolutional layers (Yu et al., 2017). Fusing different types of layers allows for concatenating both global and local features. However, methods

like Multi-layer Orderless Fusion (Li et al., 2016) which are pooling-based, tend to remove the subtleties brought by local features as they are considered equivalent to global ones. To reduce these limitations, (Yu et al., 2017) and (Cao et al., 2020) tend to use local features in a second time for the re-ranking of the results obtained using local features. (Tolias et al., 2020) proposes a deep local descriptor How designed to learn and aggregate (using ASMK) local features specifically to obtain a global descriptor for instance-level retrieval. How and ASMK will be further detailed in Section 3.2.4. (Teichmann et al., 2019) proposes a regional ASMK to aggregate features extracted at a regional level with VLAD. Instead of layer fusion, the fusion can be performed at model level, either intra-model or inter-model. Intra-model fusion combines features obtained with similar or highly compatible models. For instance, (Ding et al., 2019b) combines ResNet-26 and ResNet-50, while (Kim et al., 2018) trains three attention modules to extract different features. On the contrary, inter-model fusion combines models whose structures are much more different and tries to combine very different features. Hence, (Ozaki and Yokoo, 2019) combines six different descriptors out of six models (using various combinations of backbone models, loss and data augmentation strategies) to improve retrieval performance. This fusion strategies can be applied at two times : "early fusion" concatenates features and then learns a metric on the concatenated feature while "late fusion" learns metrics for each feature and concatenates optimal features.

Feature enhancement. To enhance the discriminativeness of the features computed, several enhancement strategies can be applied.

- *Feature aggregation.* The idea behind is to improve the discrimination of features using pooling methods on feature maps (Zheng et al., 2016). Hence, sum or average pooling will limit the effect of highly activated features, while max pooling will enhance the power of sparse features that are rarely activated. R-MAC (Tolias et al., 2016b) allows for a more targeted approach during the pooling. GeM (Radenovic et al., 2019) generalizes both max and average pooling to achieve current state-of-the-art regarding pooling layers.
- *Feature embedding.* The principle behind feature embedding is to obtain compact features by embedding convolutional maps into a high dimensional space. The embedding methods are widely similar to those used in hand-crafted methods (cf. 3.2.1), namely, BoW (high dimensionality and sparsity, not ideal for large datasets), VLAD (affected by the number of centroids but more effective than BoW), or FV (extends BoW, captures more statistics but costs more memory-wise). As those embedding methods have to be added to the networks as a new layer, it led to the NetVLAD (Arandjelovic et al., 2018) approach, where the network is fine-tuned using VLAD; and Patch-NetVLAD (Hausler et al., 2021), where the network extends NetVLAD by deriving patch-level features from NetVLAD residuals to increase invariance to many visual changes. Hash embeddings can also be further used to

limit the high dimensionality of the feature representations. Hash functions can be plugged into deep networks (Wang et al., 2018)(Deng et al., 2020) to embed features into compact codes, allowing for higher efficiency both storage and computation wise, especially for RSIR applications in regard to the large amount of data considered (Li et al., 2020).

- *Attention mechanisms.* The idea behind attention is to focus on relevant features and limit "distractions" caused by irrelevant ones, using attention maps. Those maps can be computed with or without using deep networks. Without deep networks, no parameters are learned, the basic methods are either channel wise or spatial pooling and are applied to convolutional layers. Spatial-based methods weigh activations on feature maps to determine the most relevant ones but in (Ng et al., 2020) authors not only weigh activations at different spatial locations, but they also explore the correlations between the different activations. Separately, channel-wise methods like (Xu et al., 2018) rank weighted feature maps to select regional features. Using deep networks to learn attention maps is now widely used in literature, especially thanks to the greater adaptability of those methods. Indeed, full feature maps or patches can be fed to the network to predict relevant features (Noh et al., 2017). Full images can also be used as input (Hu and G.Bors, 2020), separately training the feature description network and the attention computation network. Specifically for our problem of image retrieval, those mechanisms can be used in combination with metric learning (cf. section 3.2.3) to improve retrieval (Ng et al., 2020).

Transformers models. A specific focus must be made on transformers for CBIR. Indeed, unlike all approaches described previously that build on CNN architectures, new networks for vision tasks now rely on the transformer architecture. The advent of transformers models introduced with (Vaswani et al., 2017) for Natural Language Processing tasks, allowed for new approaches to image retrieval with vision transformers (Dosovitskiy et al., 2021). Relying on self-attention layers, transformers process the image as a series of patches (like words in a sentence) and embed information globally across the whole image. Vision transformers variants have reached state-of-the-art performances on multiple vision tasks, like (Dosovitskiy et al., 2021) first when proposing vision transformers (ViT) for image classification. (Carion et al., 2020) combines a traditional CNN with a transformer for object detection, removing the need for post-processing steps for clearing the results of the first feature extraction and object detection steps. (Touvron et al., 2021a) combines transformers models with knowledge distillation to improve performance and limit training overhead for image classification tasks. (Touvron et al., 2021b) proposes a class-attention layer to improve image classification performance of transformer models by feeding the model the class token later in the training. (Bao et al., 2022) proposes BEiT, a BERT like pre-training approach for vision transformers that allows for pre-training a network that can be used for image classification or image segmentation. Building on this, (Touvron et al., 2022) revisits the pre-training approaches for vision transformers, inspired by pre-training approaches used for CNNs like ResNet-50. (Wang

et al., 2021) combines the transformer architecture with the pyramid approach of CNNs to create a versatile backbone for many vision tasks. (Zhang et al., 2021) further improves this multiscale approach for high-resolution image encoding, resulting in a backbone for several vision tasks. (Han et al., 2021) embeds a transformer architecture within a global transformer architecture to further extract information from image patches for both image classification and object detection. (Liu et al., 2021) proposes a shifted window approach to vision transformer to perform self-attention on changing parts of the image layer after layer, thus encoding more information. (Dong et al., 2022) builds on this with CSWin, its self-attention using a cross-shaped window divided between heads (self-attention on vertical and horizontal stripes is computed simultaneously). It also introduces a new positional encoding scheme supporting arbitrary input resolutions, ideal for the different downstream tasks.

Several of these approaches are designed as backbones suited for multiple downstream vision tasks as shown in Table 3.1 and could be exploited for image retrieval. The main drawbacks of transformer-based networks however is the fact that they require a large amount of training data, as well for training from scratch as for fine-tuning.

3.2.2.2 Fine-tuning models

Instead of using off-the-shelf models, fine-tuning existing models allows for a thinner domain gap as the model’s parameters and weights are slightly adapted to the specific dataset, which is often essential when working with challenging data. As presented in Table 3.1, multiple networks have been developed and can be used as backbones to be fine-tuned on the specific data used. However, it requires to have enough training data for the fine-tuning to be efficient. Depending on the dataset used, the fine-tuning step can be either supervised or unsupervised.

Supervised fine-tuning. These methods of fine-tuning can be considered when enough information can be gathered regarding the dataset. It can be a classification or a similarity evaluation between images. Using a cross-entropy loss on a classified dataset can improve the features computed (either global or local), but the focus remains on inter-class variability and fails to distinguish intra-class specificities. Hence, using information describing similarity or dissimilarity between images offers more opportunities for robustness to both inter-class and intra-class variability. The principle is to fine-tune the network to learn a metric that preserves the similarity (or dissimilarity) between the features computed. Three approaches are possible to preserve feature similarity close to image similarity.

- *First*, using a transformation matrix (Garcia and Vogiatzis, 2019), that is concatenating features from two images and maximizing or minimizing the similarity score estimated using this concatenated feature to conform to the binary label similar/dissimilar assigned to the image pair.
- *Second*, using Siamese networks and only two images, either a similar pair or a

Table 3.1: Main network backbones for image retrieval

Backbone	Date	Training dataset	Testing dataset	Goals	Improvements
VGG (Simonyan and Zisserman, 2015)	2015	ImageNet	ImageNet	Image classification	Uses a deeper model Up to 19 layers
ResNet (He et al., 2016)	2016	ImageNet Cifar-10	ImageNet Cifar-10	Image classification and localization, object detection	Residual learning Increase representation depth
Inception ResNet (Szegedy et al., 2017)	2017	ImageNet	ImageNet	Image classification	Network with Inception module and residual learning
Xception (Chollet, 2017)	2017	JFT	FastEval14k	Image classification	Extreme Inception module
DenseNet (Huang et al., 2017)	2017	CIFAR-10, CIFAR-100, SVHN, ImageNet	CIFAR-10, CIFAR-100, SVHN, ImageNet	Image classification	Dense connection between all layers
ResNext (Xie et al., 2017)	2017	ImageNet, CIFAR, COCO	ImageNet, CIFAR, COCO	Image classification, object detection	Aggregates a set of smaller transformations, increasing cardinality
NASNet (Zoph et al., 2018)	2018	CIFAR-10	CIFAR-10, ImageNet	Image classification, object detection	Learn the model's architecture using the dataset and transfer it to other datasets
Fishnet (Sun et al., 2018)	2018	ImageNet, COCO	ImageNet, COCO	Image classification, object detection	Combines pixel-level, region-level and image-level information
SENet (Hu et al., 2018)	2018	ImageNet, COCO	ImageNet, COCO	Image classification, object detection	Add the Squeeze and Excitation block
EfficientNet (Tan and Le, 2019)	2019	ImageNet	ImageNet, other datasets for transfer learning	Image classification	New architecture, new method to scale width, depth and resolution simultaneously
ResNeSt (Zhang et al., 2020a)	2020	ImageNet, COCO	ImageNet, COCO	Image classification, object detection	Add a Split-Attention block
ViT (Dosovitskiy et al., 2021)	2021	ImageNet, ImageNet-21k, JFT-300M	ImageNet, CIFAR-10, CIFAR-100, VTAB	Image classification	Introduces vision transformers with self-attention layers
Swin (Liu et al., 2021)	2021	ImageNet, COCO, ADE20K	ImageNet, COCO, ADE20K	Image classification, object detection, semantic segmentation	Uses a shifting window for self-attention
Pyramid Vision Transformer (Wang et al., 2021)	2021	ImageNet, COCO, ADE20K	ImageNet, COCO, ADE20K	Image classification, object detection, semantic segmentation	Uses a pyramid-like structure like CNN's but with transformers
Multi-Scale Vision Longformer (Zhang et al., 2021)	2021	ImageNet	ImageNet, COCO	Image classification, object detection	Uses a multiscale approach and adapts Longformer for images
CSWin (Dong et al., 2022)	2022	ImageNet-21K, COCO, ADE20K	ImageNet	Image classification, object detection, semantic segmentation	Uses a cross-shaped self-attention and locally enhanced positional encoding

dissimilar one. The network's weights are shared between layers. (Ong et al., 2017) for instance, uses both Fisher vectors and a Siamese network to compute features.

- *Finally*, following the Siamese network idea, using a Triplet Network allows for optimizing the metric using at the same time a similar and a dissimilar pair, each sharing the same "anchor". Using a triplet loss trains the model to learn representations minimizing the dissimilarity with the positive example and maximizing it with the negative one. Improving those two distances improves the final relevance and discriminative power of the computed features.
- *Furthermore*, those methods of fine-tuning can also be combined with various modules to further improve the quality of the features. Region Proposal networks allow the network to adopt a more local approach to the feature computation, focusing on more relevant parts of the image (Gordo et al., 2017). Attention modules can also be plugged into deep networks to focus on specific regions, improving inter-class but also intra-class feature discrimination. Finally, a combination of losses may be of interest to exploit, for instance the inter-class discriminative power of the classification loss and the intra-class discriminative power of the triplet loss (Jun et al., 2019).

Unsupervised fine-tuning. The issue regarding supervised fine-tuning is the cost of annotating the dataset to have ground truth data regarding the class or the similarity between images. Hence, unsupervised fine-tuning methods are a way to explore when dealing with poorly annotated datasets as they do not require a ground truth. A first idea behind this type of fine-tuning is to exploit/mine the data to estimate which images are similar to each other to estimate the relevance of an image compared to another in order to use this estimation for the latter step using positive examples and negative examples (e.g. a triplet network). This method is called manifold learning. Using the first output of the network, an affinity matrix is computed (similar to a weighted kNN-graph), then the pairwise similarities are reevaluated in light of all other pairwise affinities. After several iterations, the deep representations are spatially organized in the manifold space and using a distance in this space the positive and negative examples are mined to fine-tune the initial network. A second idea to exploit the new unannotated dataset is to use AutoEncoders within image retrieval frameworks. An AutoEncoder is a neural network that wishes to reconstruct its output as similar as possible as its input. Hence, it will encode the input into a deep feature then decode it to obtain in our case an image very similar to the one fed as input. It can be used within a network aiming at hash-encoding images, as (Shen et al., 2020c) aims at limiting prior computation of a graph of image similarity (with manifold learning, for instance) by using an AutoEncoder framework.

3.2.3 Similarity search

After the computation of the most optimal features for both the database images and the query image, the second step in image retrieval is the search of neighbors in the

feature representation space, i.e. the images similar to the query. The similarity search is composed of two steps: first, computing a distance between features, whether local or global, and second, querying efficiently the database to find the most similar images.

Similarity measures. Various distances can be used as a similarity measure to compare two vectors of features, *e.g.* the l^1 norm, the most-used l^2 norm, the inner product. These measures can be used for global descriptors or multiple local descriptors (*e.g.* (Kim et al., 2015) sums similarity of each feature to obtain the global similarity). For local features, other metrics can be used like SMK or its aggregated version ASMK (Tolias et al., 2016b) (more details in Section 3.2.4).

Similar images search. The search of similar images can then be seen as a straightforward step of finding the k nearest neighbors in the feature space. However, this task can be quite expensive when dealing with a large database or with high-dimensionality features. Several improvements have been proposed to speed up this process.

Feature dimension reduction. First, approaches aiming to reduce the dimension of the features have been proposed, reducing the computation cost of the matching. For instance, (Arandjelovic et al., 2018; Gordo et al., 2017) use Principal Component Analysis to reduce the dimension of CNN-based features, it is the mainly used approach. This can also be used jointly with a whitening step like in (Tolias et al., 2020) that helps ensure data consistency and limiting the impact of co-occurrences during similarity computation. Using quantization methods like binarization (Cao et al., 2020) reduces the storage requirement and speeds up the computation at search time.

Approximate nearest neighbors. Second, exact nearest neighbors are costly to identify as it supposes an exhaustive similarity computation between the query and all images in the database. Thus, using various indexing structure stopping criteria, approximate nearest neighbor methods speed up the process immensely, performing a non-exhaustive search (Johnson et al., 2019; Magliani et al., 2019). The nearest neighbor search can also be improved by using several features per image and then using Dominant Set Clustering (Zemene et al., 2019) to find the most similar images.

Other search methods. Several other approaches have been devised to perform the search in the database. (Kim et al., 2015) for instance exploits SVM classifier to estimate the robustness of the descriptors and select those to compare in the database. (Stumm et al., 2015) matches graphs of visual words present in the images using a graph kernel. Grouping images of the database and then matching the query to the clusters also proves to improve the search efficiency. Graph-based approaches like Hierarchical Navigable Small World (Malkov and Yashunin, 2020) have proved efficient for approximate nearest neighbors search. Indeed, it uses a proximity graph between images in the database. Starting anywhere, it compares the query with all its connected nodes and moves to the one closest to the query. Iteratively, the nearest neighbor is found. In the case of local features, text-based approaches of inverted files have been adapted to visual words, allowing to select potentially similar images based on their shared visual words, this is

used in (Tolias et al., 2020) for instance. Figure 3.2 summarizes approaches developed for searching nearest neighbors efficiently in a database. For a more in-depth study, readers may refer to (Wang et al., 2018).

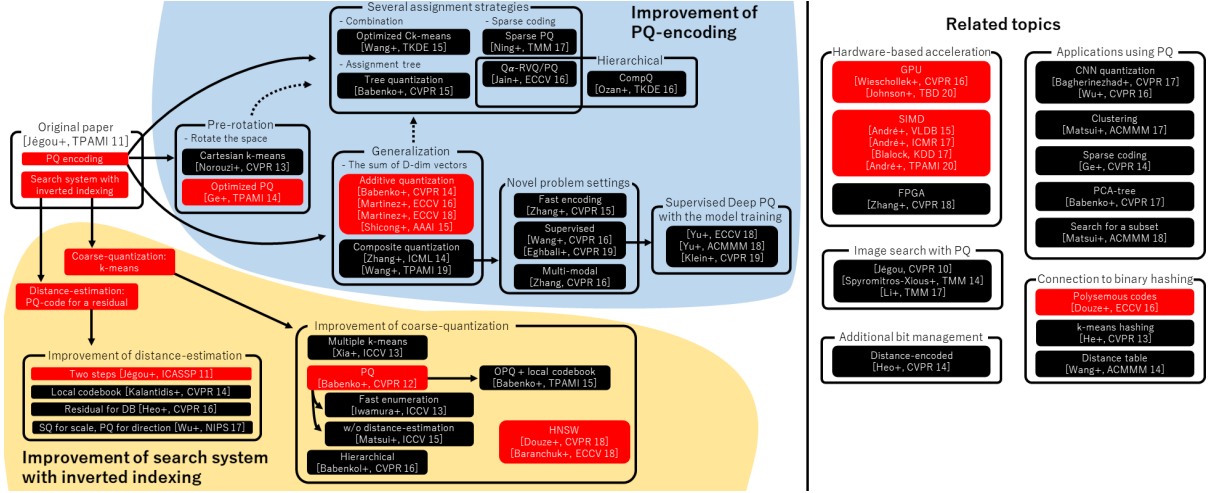


Figure 3.2: Overview of approaches for approximate nearest neighbors search (illustration from (Matsui et al., 2018))

3.2.4 Spotlight on How and ASMK

Later in this thesis, evaluations will show that How combined with ASMK is an efficient descriptor for image retrieval in our specific dataset. We thus present here the approaches in more details.

How

We present here the proposed descriptor How as defined in (Tolias et al., 2016a, 2020), used later in our experiments for image retrieval. We largely use the authors descriptions, equations and illustrations.

The idea behind How is to extract deep local features that are aggregated using a global sum-pooling at training time to obtain a global descriptor that is optimized at image level using contrastive loss. Intuition of the authors is that optimization at image-level with a contrastive loss implicitly optimizes local features in various ways. First, local background features’ impact is lowered, while local foreground features’ importance is heightened. Second, local descriptors of similar images are pushed closer in the feature space and the opposite for dissimilar images.

Furthermore, an attention metric to evaluate the strength of each local feature in the descriptor is used first for weighting the contribution of each local feature in the global descriptor at training time and then to select the strongest features for aggregation with ASMK at testing time.

An overview of the process is available in Figure 3.3.

This descriptor is optimized to be used with ASMK which proved an efficient way to aggregate local features and is described next.

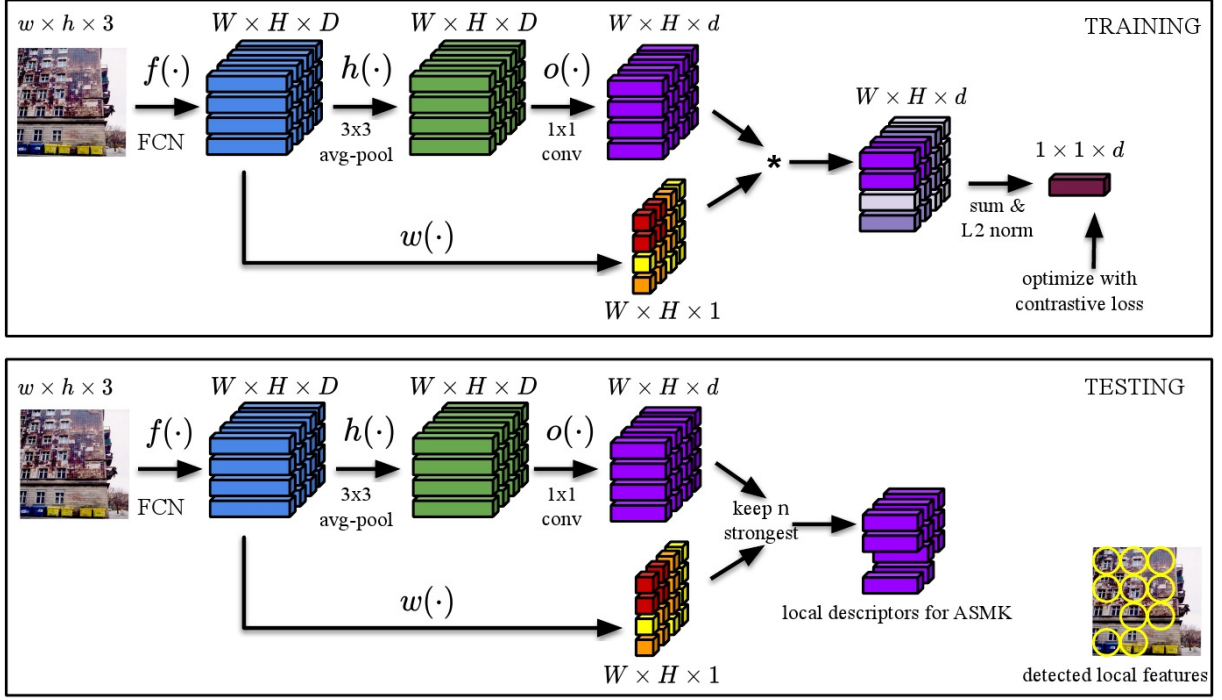


Figure 3.3: Overview of the architecture for How local features (illustration from (Tolias et al., 2020))

ASMK

We present here the binarized version of Selective Match Kernel (SMK) and its extension, the Aggregated Selective Match Kernel (ASMK) as defined in (Tolias et al., 2016a, 2020), used with How later in our experiments for indexing and retrieval. We largely use the authors equations and descriptions of the process.

An image is represented by a set $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d\}$ of $n = |\mathcal{X}|$ d -dimensional local descriptors. The descriptors are quantized by $q : \mathbb{R}^d \rightarrow \mathcal{C} \subset \mathbb{R}^d$. $\mathcal{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_k\}$ is a codebook comprising $|\mathcal{C}|$ vectors (visual words), extracted both in the original publication and our experiments on the *SfM120k* dataset (Radenovic et al., 2019).

Using the quantization process, every descriptor \mathbf{x} is assigned to its nearest visual word $q(\mathbf{x})$. It creates two types of subsets: $\mathcal{X}_c = \{x \in \mathcal{X} : q(x) = \mathbf{c}\}$ the subset of descriptors in \mathcal{X} assigned to visual word \mathbf{c} , and $\mathcal{C}_{\mathcal{X}}$ the set of all visual words that appear in \mathcal{X} .

Especially for its use with How, a binarization step is used: \mathbf{x} is mapped to a binary vector through function $b : \mathbb{R}^d \rightarrow \{-1, 1\}^d$ given by $b(\mathbf{x}) = \text{sign}(r(\mathbf{x}))$, where $r(\mathbf{x}) = \mathbf{x} - q(\mathbf{x})$ is the residual vector w.r.t. the nearest visual word and sign is the element-wise sign function.

The SMK similarity of two images, represented by \mathcal{X} and \mathcal{Y} respectively, is estimated by cross-matching all pairs of local descriptors with match kernel

$$S_{\text{SMK}}(\mathcal{X}, \mathcal{Y}) = \gamma(\mathcal{X})\gamma(\mathcal{Y}) \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{y} \in \mathcal{Y}} [q(\mathbf{x}) = q(\mathbf{y})] k(b(\mathbf{x}), b(\mathbf{y})), \quad (3.1)$$

where $[\cdot]$ is the Iverson bracket and $\gamma(\mathcal{X})$ is a scalar normalization that ensures unit

self-similarity. Function $k : \{-1, 1\}^d \times \{-1, 1\}^d \rightarrow [0, 1]$ is given by

$$k(b(\mathbf{x}), b(\mathbf{y})) = \begin{cases} \left(\frac{b(\mathbf{x})^\top b(\mathbf{y})}{d} \right)^\alpha, & \frac{b(\mathbf{x})^\top b(\mathbf{y})}{d} \geq \tau \\ 0, & \text{otherwise} \end{cases} \quad (3.2)$$

where $\tau \in [0, 1]$ is a threshold parameter. Only descriptor pairs that are assigned to the same visual word contribute to the image similarity in Equation 3.1. In practice, not all pairs need to be enumerated and image similarity is equivalently given by

$$S_{\text{SMK}}(\mathcal{X}, \mathcal{Y}) = \gamma(\mathcal{X})\gamma(\mathcal{Y}) \sum_{\mathbf{c} \in \mathcal{C}_{\mathcal{X}} \cap \mathcal{C}_{\mathcal{Y}}} \sum_{\mathbf{x} \in \mathcal{X}_{\mathbf{c}}} \sum_{\mathbf{y} \in \mathcal{Y}_{\mathbf{c}}} k(b(\mathbf{x}), b(\mathbf{y})) \quad (3.3)$$

where cross-matching is only performed within common visual words.

In the case of ASMK, the local descriptors assigned to the same visual word are first aggregated into a single binary vector. This is performed by $B(\mathcal{X}_{\mathbf{c}}) = \text{sign}(\sum_{\mathbf{x} \in \mathcal{X}_{\mathbf{c}}} r(\mathbf{x}))$, with $B(\mathcal{X}_{\mathbf{c}}) \in \{-1, 1\}^d$. Image similarity in ASMK is then given by

$$S_{\text{ASMK}}(\mathcal{X}, \mathcal{Y}) = \gamma(\mathcal{X})\gamma(\mathcal{Y}) \sum_{\mathbf{c} \in \mathcal{C}_{\mathcal{X}} \cap \mathcal{C}_{\mathcal{Y}}} k(B(\mathcal{X}_{\mathbf{c}}), B(\mathcal{Y}_{\mathbf{c}})). \quad (3.4)$$

Aggregating the descriptors proves more efficient computationnally and memory-wise. Furthermore, it deals better with the burstiness problem, detrimental to image retrieval.

For retrieval, an inverted-file indexing structure is used to perform efficient search.

3.2.5 CBIR and iconographic heritage

Due to the diversity of the iconographic heritage, the methods presented above apply to various types of contents. Here we first present applications dedicated to digitized paintings or manuscripts and then applications for long-term landmark retrieval.

3.2.5.1 Pattern spotting and artwork recognition

While image retrieval aims at finding similar images, pattern spotting goes further and consists of retrieving in a collection of historical document images occurrences of a graphical object and estimating its location on the image. Methods first applied traditional handcrafted image retrieval descriptors and methods like (En et al., 2016a) and (En et al., 2016b). But deep learning approaches have been designed specifically for this problem as new datasets have become available, for instance (En et al., 2016c). (Wiggers et al., 2019b) presents two previous methods (Wiggers et al., 2018, 2019a) in a comparative way. Indeed, one method (Wiggers et al., 2018) is based on a AlexNet and pre-processes the data using a Selective Search algorithm to extract multiple candidates out of a single image, all those candidates being later described and compared to the query. On the other hand (Wiggers et al., 2019a) exploits a Siamese network instead of a conventional CNN and outperforms the previous method using feature maps of higher dimension. (Úbeda et al., 2019) uses RetinaNet and its pyramidal architecture to extract features at various

levels. Combining this with a NonText classifier to search only on relevant image parts, it improves the state-of-the-art of (En et al., 2016c) when it comes to pattern spotting, however lowering the results of "simple" image retrieval. (Úbeda et al., 2020) further improves the work of (Úbeda et al., 2019), enhancing their results both on pattern spotting and image retrieval. When it comes to heritage content, artworks collections have become a new object of interest as their digitization increases and their heterogeneity make them a very interesting object of study. Hence (Yang and Min, 2020) uses a CNN (DenseNet) to classify oilpaint brush, pastel, pencil and watercolor artworks. Similarly, (Ufer et al., 2020) proposes a multi-style feature fusion approach to reduce the domain gap. (Hu et al., 2023) proposes an approach for cross-domain retrieval (paintings, sketches, photographs and so on) by learning domain-agnostic features using the more generic frequency domain, which proves less variable than the actual representations. They however apply it only on specific objects and not actual scenes. (Shen et al., 2019) aims at finding similar instances of details amongst different artworks but simultaneously proves that its methods brings improvement on localization for historical photographs datasets, as illustrated in Figure 3.4. (Shen et al., 2020b) further extend their work to find watermark in historical documents, a challenging pattern recognition problem.



Figure 3.4: Example of pattern spotting in art collections (illustration from (Shen et al., 2019))

3.2.5.2 "Very" long-term image retrieval

Image retrieval on heritage content can be viewed as "very" long-term image retrieval. Indeed, heritage content often depicts known visual landmarks that a human can compare to a more recent depiction. However, as (Fernando et al., 2015) explains, old and new images can be considered as belonging to two different domains, making it difficult for



Figure 3.5: Example of long term image retrieval (illustration from (Shen et al., 2019))

automated methods to link the two depictions.

Linked to pattern spotting, a first specific lead is followed, aiming at finding similar details on artworks or historic documents and it is applied on visual landmarks by (Shen et al., 2019) as illustrated in Figure 3.5 where the areas of interest to the model (non changing) are identified both on the old and the recent image.

With a very applied aspect to their research (building a 4D visualization platform for heritage contents), (Maiwald et al., 2021) applies deep features to heritage content. However, to further ensure the quality of the retrieval, they manually select three queries of a same object with different viewpoints and select the intersection of the retrieved image lists as the final list. Further along, (Maiwald et al., 2023) uses a first metadata-based searching step to subsample the collection in which they apply the CBIR process.

Another lead is to extend the research on long-term visual localization to even more expand the temporal gap and thus the domain gap. Several image retrieval datasets focus on landmark retrieval as presented in Table 3.2. However, the time gap or heterogeneity of the data may not be representative when compared to cultural heritage content.

Table 3.2: Image retrieval datasets usable for landmark retrieval applications. MMS stands for Mobile Mapping System.

Dataset	Number of images	Viewpoint	Time gap
Large TimeLags Locations (Fernando et al., 2015)	500	Street-level	150 years
Google Landmarks Dataset v2 (Weyand et al., 2020)	Over 5M	Street-level and aerial	Unspecified
\mathcal{R} Oxford (Radenovic et al., 2018)	Over 5k	Mostly street-level and some aerial	Unspecified
Aachen Day-Night (Sattler et al., 2018)	7712	Street-Level	2 years
Extended CMU-Seasons (Sattler et al., 2018)	Over 110k	Street-level MMS camera	1 year
RobotCar Seasons (Sattler et al., 2018; Maddern et al., 2017)	Over 35k	Street-level MMS camera	1 year
SILDa Weather and Time of Day (Balntas, 2019)	Over 14k	Street-level and aerial	1 year
HistAerial (Ratajczak et al., 2019)	4.9M	Vertical aerial	1970-1990
ALEGORIA (Gominski et al., 2019)	13175	Street-level and aerial	1920's-today

Table 3.3: Image retrieval datasets' specific heterogeneity

Dataset	Color	Domain	Illumination	Occlusion	Scale	Orientation
Large Time Lags Locations (Fernando et al., 2015)	✓		✓		✓	✓
Google Landmarks Dataset v2 (Weyand et al., 2020)			✓	✓	✓	✓
\mathcal{R} Oxford (Radenovic et al., 2018)	✓		✓	✓	✓	✓
Aachen Day-Night (Sattler et al., 2018)			✓	✓	✓	✓
Extended CMU-Seasons (Sattler et al., 2018)			✓	✓		✓
RobotCar Seasons (Sattler et al., 2018; Maddern et al., 2017)			✓	✓		✓
SILDa Weather and Time of Day (Balntas, 2019)			✓	✓		✓
HistAerial (Ratajczak et al., 2019)				✓	✓	
ALEGORIA (Gominski et al., 2019)	✓	✓	✓	✓	✓	✓

Most datasets do not reflect correctly the heterogeneity representative of cultural her-

image contents as summarized in the Table 3.3. Indeed, the heterogeneity may come from multiple factors. The main differences are the **color** of the image (is it a regular RGB image, a sepia one, a black and white one ?), the **domain** of the image (a photograph, a drawing, an engraving, etc.), the **illumination** variation (between seasons or day and night for instance), the **occlusion** that may be present (preventing from correctly discerning the landmark), the **scale** (meaning the size the main object takes in the whole image, influencing the level of detail to which one can see it but also the number of distracting elements in the image) and finally the **orientation** of the image (i.e. the viewpoint used to depict the landmark). To address this issue, (Gominski et al., 2019) proposes the ALEGORIA dataset consisting of multi-source, multi-date and multi-view images with wide intra-class variations (viewpoint, illumination, date, media, etc). This new dataset allows (Gominski et al., 2019) to challenge seven state-of-the-art image descriptors (including six deep features). It concludes that there are still many difficult cases to be handled by image retrieval methods and methods need to become more robust to bigger changes in viewpoint or illumination for instance. Extending their work and dataset in (Gominski et al., 2021), they explain the difficulties encountered by state-of the art deep features in the context of cultural heritage content, detailing the most problematic variations. Especially, they affirm that the existing datasets do not cover wide enough variations in content to properly train feature descriptors to be used on heritage content. In order to improve the results, work on precision and recall is done. However, place for improvement remains for deep features to be generalised to such datasets.

3.3 Re-ranking

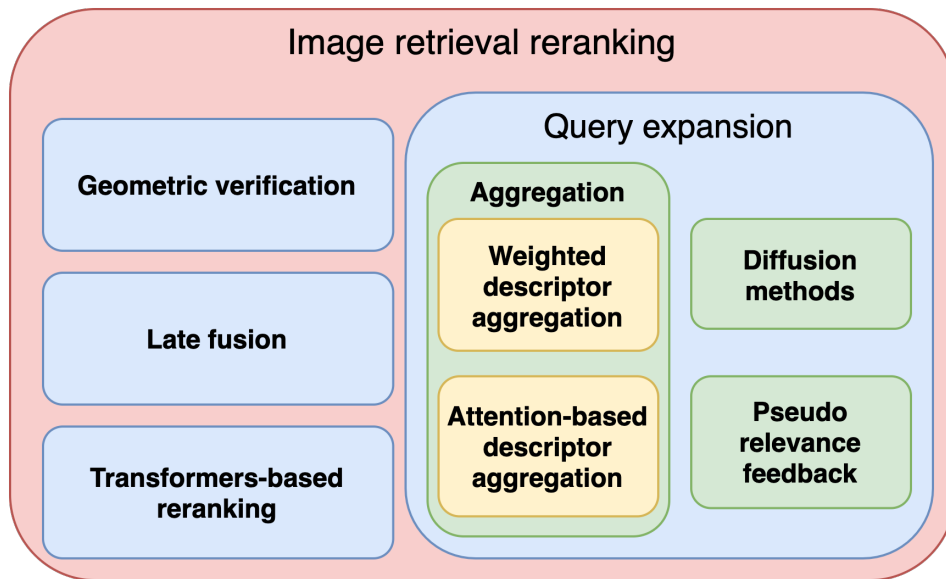


Figure 3.6: Re-ranking methods paradigms

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, as presented at the beginning of the chapter in Figure 3.1, 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. Multiple paradigms of re-ranking exist as described in Figure 3.6. In this section, we will further present these paradigms and the methods associated.

3.3.1 Late fusion

The principle behind late fusion is to exploit two (or more) separate retrieval processes and fuse their similarity lists or scores afterwards to obtain a new list of similar results combining optimally the previous results. The idea is to exploit the efficiency of different approaches (for instance global description and local description) while alleviating their drawbacks, this on a dataset where both approaches perform very differently depending on each query. (Zhang et al., 2012) for instance exploits both global and local features similarities as two local graphs that they fuse to obtain the best similarity result overall. (Ye et al., 2012) exploits similarity results as several ranking matrices and finds the most common ranking matrix between them. Another approach in (Zheng et al., 2015) exploits the similarity score curve of each descriptor to estimate its efficiency with regard to each specific query and uses it to weigh the impact of this descriptor in the fusion scheme. Even in the paradigm of learned features late fusion can be exploited as in (Wang et al., 2020b) where the image features are compared using multiple distances to be the most discriminative possible, as illustrated in Figure 3.7.

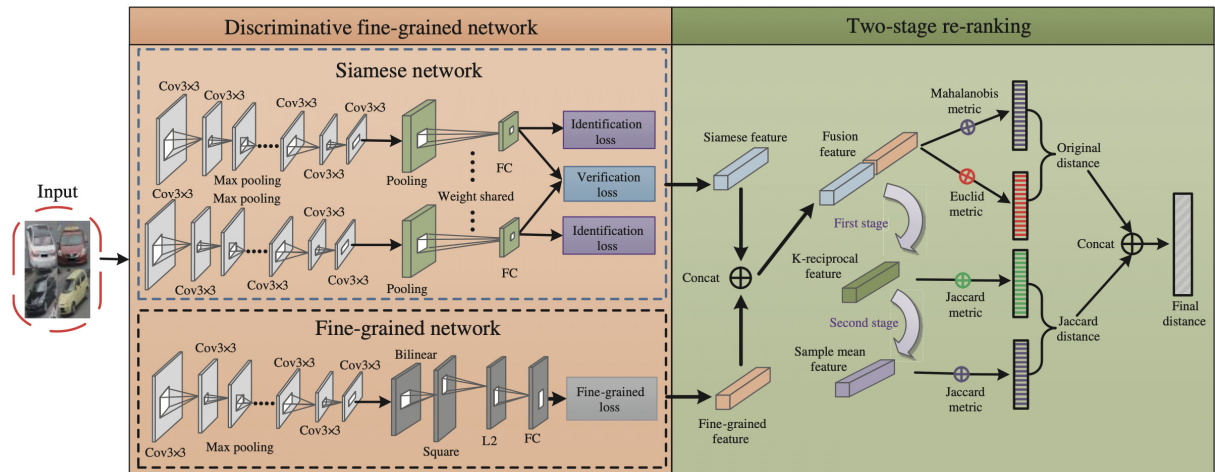


Figure 3.7: Example of late fusion using multiple distances (illustration from (Wang et al., 2020b))

3.3.2 Geometric verification

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 geomet-

ric transformation parameters (affine transformation) using a robust approach such as RANSAC (Fischler and Bolles, 1981; Cao et al., 2020) which fits data to a model while being robust to outliers (in our case wrong matches). It consists of four steps:

1. local features are extracted for the query image and its k most similar images,
2. the features are matched between the query image on one side and each of the k images,
3. out of this k matched sets, an affine transformation is estimated via RANSAC,
4. the k images are re-ranked based on the number of inliers kept by the RANSAC process.

An example of matches considered as inliers is shown in green in Figure 3.8 while outliers (incorrect matches) are shown in red. The idea behind this geometric transformation is to check the consistency of the matching between specific points in both images. Thus, two images displaying the same place will be geometrically speaking more consistent than two images which local features match but without any coherence, indicating that they are less likely to display the same object.

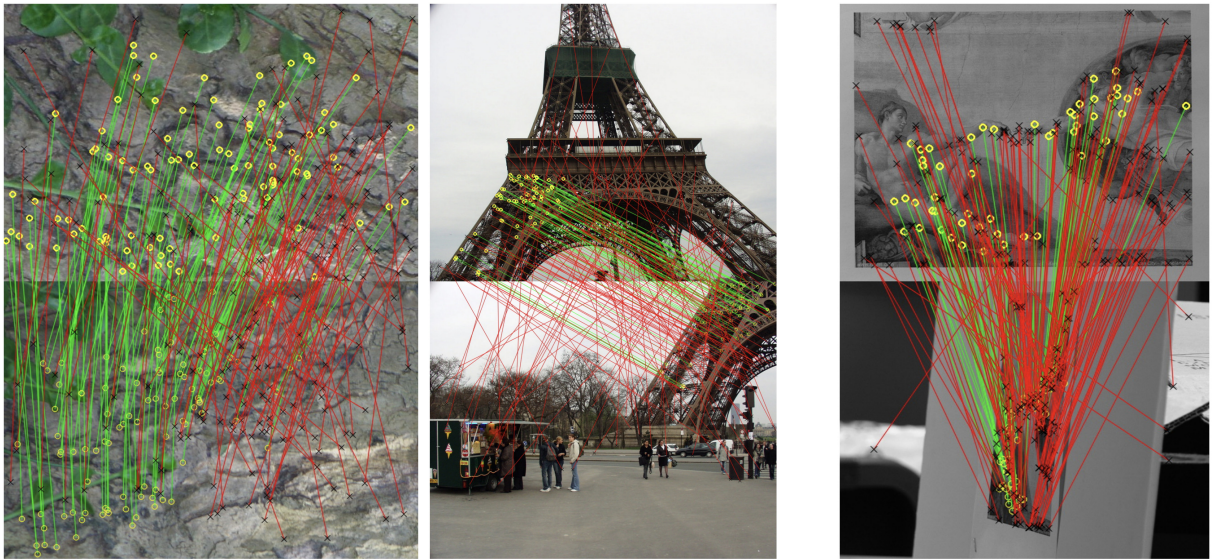


Figure 3.8: Examples of RANSAC-based matches selection (illustration from (Wang et al., 2020a))

The local features used can directly be the ones used during the retrieval step as in (Noh et al., 2017) with DELF. However, more precise local features have been developed and used in other computer vision tasks, such as Structure from Motion (SfM) which estimates a 3D structure from 2D images of a same object but taken from different viewpoints. These more precise features are often more efficient for the geometric verification process, even though they require another feature extraction step. First of all, most handcrafted features described in Section 3.2.1 can be used for this task. However, newly developed learned feature extractors prove to be the new state-of-the-art, especially when dealing

with a high variability in the contents. Examples are D2Net (Dusmanu et al., 2019), SuperPoint (DeTone et al., 2018) or DISK (Tyszkiewicz et al., 2020).

Once local features are extracted, the matching between them can be computed in various ways. The most recent specific approaches for matching features are learned one. Multiple approaches exist such as SuperGlue (Sarlin et al., 2020) and its improvement LightGlue (Lindenberger et al., 2023) which exploits a self attention mechanism to base its matching on priors specific to the underlying 3D scene. With OANet (Zhang et al., 2019), the matching exploits both the local and the global context extracted from the existing sparse correspondences between two sets of local features.

(An et al., 2023) proposes an adaptation of this classical RANSAC using topological relations instead of spatial ones to improve re-ranking without requiring fine-tuning, which could prove useful in our setting of iconographic content heritage retrieval.

The detection and matching can also be concomitant as in (Sun et al., 2021) which uses a transformers-based approach to mitigate the cost of first detecting and describing features and then matching them.

This geometric verification step can also be included directly in the descriptor extraction process as in DOLG (Yang et al., 2021), DELG (Cao et al., 2020) or CV-Net (Lee et al., 2022) which does not apply RANSAC with local features, but a dense cross-scale feature correlation to assess the coherence between images. The efficiency of spatial verification is such that this geometric verification step is now further embedded in the descriptor extraction process. (Zhang et al., 2023b) proposes to extract global features embedding directly their spatial context. The subsequent matching exploits both information (visual and spatial) at once rather than in two separate steps.

(Cai et al., 2023) proposes a dataset and a method to disambiguate visually similar images (in their case similar facades of monuments, front and back for instance). They see this problem as a classification task, binarily deciding whether or not two images display the same side of the building. Their geometric verification is thus a trained network exploiting geometric matches for a classification task.

3.3.3 Transformers-based re-ranking

The advent of transformers model introduced with (Vaswani et al., 2017) proposed new approaches to image retrieval with vision transformers (Dosovitskiy et al., 2021). The use of transformers for re-ranking became the next logical step. Exploiting self-attention mechanisms, these networks learn new scores based on different input informations.

On one side, (Tan et al., 2021) proposes RRT and exploit the re-ranking lists and both global and local descriptors to estimate a new similarity score between each pair of images, said score being then used for re-ranking. Similar to RRT, (Zhang et al., 2023a) proposes ETR, a new transformer block jointly exploiting self-attention and cross-attention when estimating the similarity within an image pair. (Zhu et al., 2023) proposes a unified pipeline (shown in Figure 3.9) for both retrieval and re-ranking based on transformers. A self-attention mechanism selects the informative local features extracted by the trans-

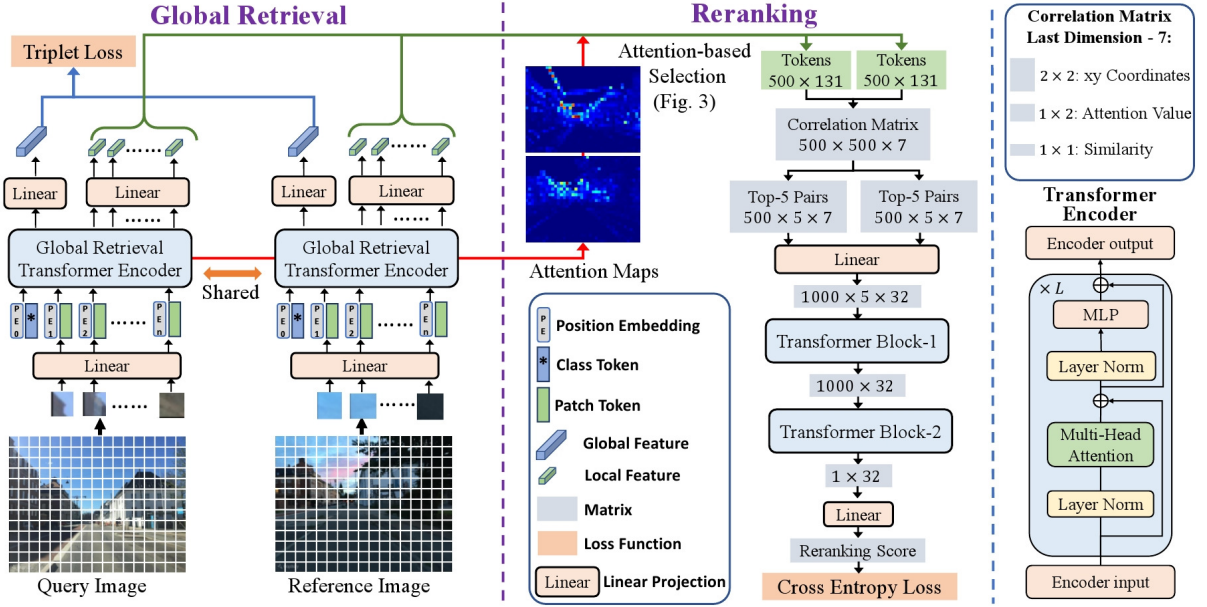


Figure 3.9: Transformer-based pipeline for re-ranking using global and local features (illustration from (Zhu et al., 2023))

former used afterwards during the re-ranking process using jointly correlation, attention and positional information between feature pairs.

On the other side, (Ouyang et al., 2021) exploits the lists of results to estimate a certain cross-similarity between the lists of similar images of two similar images. Indeed, two similar images should logically be similar to the same images. Using both lists of results exploits this logic to reestimate the similarity between two images.

3.3.4 Query expansion

A large 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. We detail further all types of approaches.

3.3.4.1 Descriptor aggregation

A first approach to query expansion is the aggregation of the image descriptors most similar to the query image. The principle is to use true positives returned by the initial process and create a new, richer query by "combining" the initial query and those positive results, in order to "attract" more positive results during a second querying step. The number of descriptors aggregated and the weighting schemes can greatly vary. Multiple adaptations have been proposed, such as changing the aggregation weighting scheme. Average-QE (AQE) (Chum et al., 2007) considers all retrieved results as equal when averaging the descriptors. An adaptation is AQEwD, meaning AQE with decay, where further an image is in the result list, lesser the impact of its descriptor is in the aggregation. HQE (Tolias

and Jégou, 2014) exploits Hamming embeddings to enrich the descriptors. (Arandjelovic and Zisserman, 2012) proposes a discriminative query expansion (DQE) which rather determines which descriptors not to aggregate rather than those to aggregate like all AQE-based methods. α -QE (Radenovic et al., 2019) in turn aggregates descriptors based on similarities weighted by an alpha coefficient. (Klein and Wolf, 2021) exploits the graph of nearest neighbors to aggregates descriptors using both the descriptors of most similar images and the descriptors of their most similar images. Finally, new learned approaches like (Gordo et al., 2020; Zhang et al., 2022) have been developed to automatically select meaningful parts of the images that should be aggregated into a new descriptor.

3.3.4.2 Pseudo relevance feedback

Similar to descriptor aggregation, the idea is also to create a new, richer descriptor, however, this approach is inspired by semi-automatic relevance feedback approaches, used for instance in commercial platforms to propose new products to a client based on its first selection. Instead of simply using the first results as correct results, other results are assumed to be incorrect to compute the new descriptor as both closer to the query and correct retrieved images and further from incorrect retrieved images. As this is in a fully automatic setting, this is only a "pseudo" relevance feedback process.

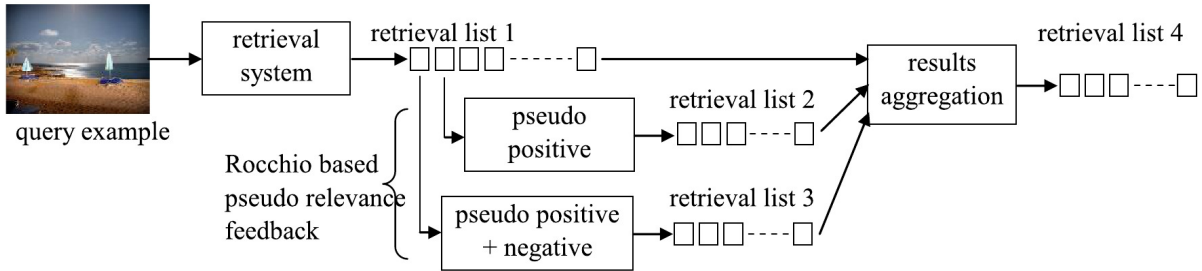


Figure 3.10: Pseudo-relevance feedback pipeline (illustration from (Lin, 2019))

(Lin, 2019) propose the pipeline from Figure 3.10 and exploit the Rocchio algorithm (Rocchio Jr, 1971) from Equation 3.5 for descriptor aggregation and also proposes a combination of result lists using the Borda count to fuse the retrieved images list, aggregating either only pseudo positive results or both pseudo positive and pseudo negative results.

The Rocchio algorithm modifies the query vector q_0 into the modified vector q_m following this formula:

$$q_m = \alpha q_0 + \beta \frac{1}{|D_r|} \sum_{d_j \in D_r} d_j - \gamma \frac{1}{|D_{nr}|} \sum_{d_j \in D_{nr}} d_j, \quad (3.5)$$

with D_r the set of pseudo-positive (relevant) images and D_{nr} the set of pseudo-negative (non-relevant) images and α , β and γ the weights of the three components of the modified query vector.

(Lin, 2022) builds on the Rocchio algorithm using a block based approach. Indeed, instead of using the same weight for all positive images and the same weight for all

negative ones, the sets of positive and negative results are divided into blocks to which a different weight is assigned, in order to preserve the order of similarity. That way, a block of positive images closer to the query will be weighed higher than the following block in a decreasing order of similarity. This aims at limiting the impact of incorrect images wrongly selected in the set of positive images.

3.3.4.3 Diffusion-based approaches

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 in order to re-rank the list of results without a new step of querying like descriptor aggregation methods.

Such solutions have achieved state-of-the-art performance on many benchmarks. (Delvin-
ioti et al., 2014) proposes three extended similarity measures for comparing the neighborhoods of candidates in the list of results and then re-rank them. (Iscen et al., 2017) transforms a similarity matrix into an affinity matrix in order to assign gradually each image to a cluster of its similar images, exploiting the manifold of the dataset. (Zhong et al., 2017) proposes to encode the set of k -reciprocal nearest neighbors of an image to reestimate a similarity between images using an adapted Jaccard metric. Using heat diffusion properties, (Pang et al., 2019) re-rank the images by estimating how much an image is heated by the query (heat source) and then how much this image contributes to heating the cluster of heated images. (Bai et al., 2019) alternates a diffusion step and a fusion step to further increase affinity between similar images while gradually removing the noise created by dissimilar images. (An et al., 2021) exploits hypergraphs intra- and inter-images to diffuse the spatial similarity information alongside the visual similarities. (Shen et al., 2021) finally leverages the structure of the similarity graph to re-rank images, refining the similarity of multiple subgraphs and aggregating them into a new global graph of similarity, as simply shown in Figure 3.11.

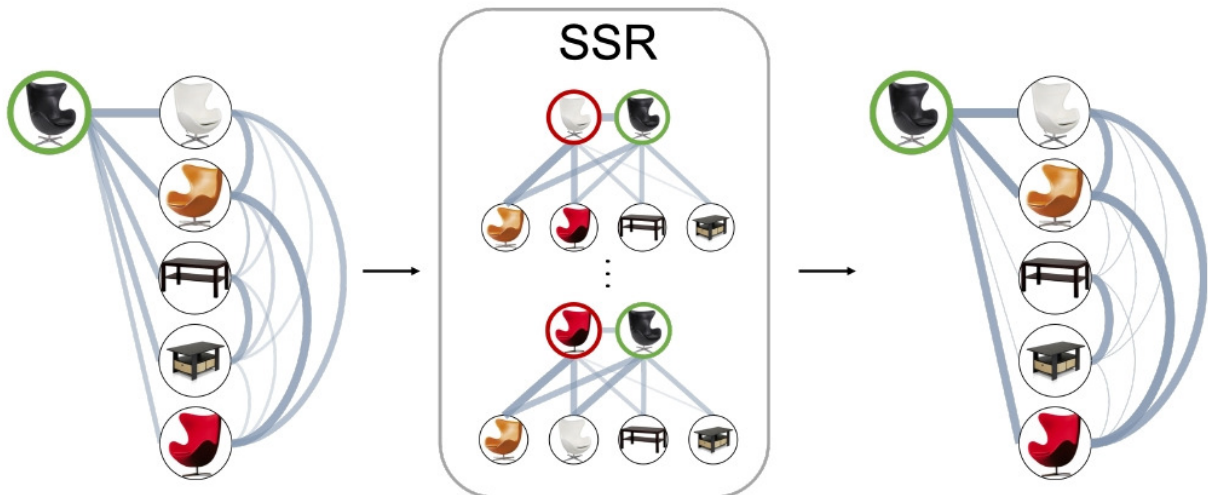


Figure 3.11: Graph-based approach to re-ranking (illustration from (Shen et al., 2021))

GNN-Reranking

During our experiments, one diffusion-based approach similar to (Shen et al., 2021) proved to be highly efficient with our dataset. It is called GNN-Reranking (GNN-R after) and was proposed in (Zhang et al., 2020b). We describe it in more details, largely using the authors equations.

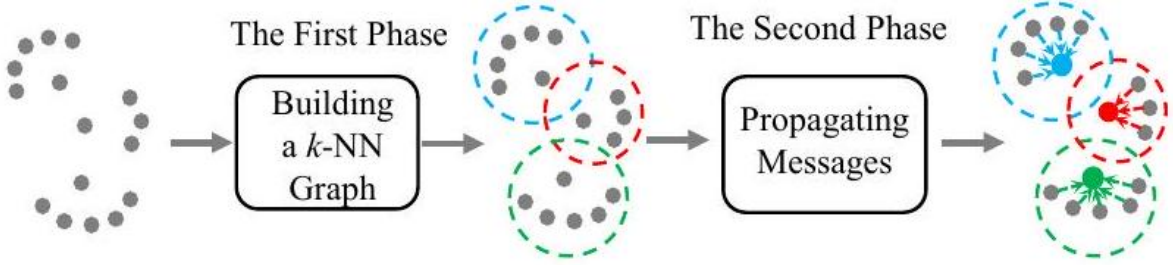


Figure 3.12: GNN-Reranking process (illustration from (Zhang et al., 2020b))

Their key idea is that image similarities can be seen as relations in a graph. That way, discrimination between sets of similar images can be performed in the graph and updating the image features can be seen as a message propagation in GNN approaches. The approach is done in two steps as shown in Figure 3.12. First, it creates a k -NN graph using all images and similarity information. The images are nodes and the similarity score between two images (two nodes) is encoded in the edge. Second, the features are updated by aggregating the features of the closest nodes weighted by the edges. The final re-ranking list is computed by comparing the updated image features.

Graph creation. The process 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 based on the similarity matrix S :

$$A_{i,j}^* = \begin{cases} 1 & \text{if } j \in \mathcal{N}(i, k_1) \wedge i \in \mathcal{N}(j, k_1) \\ 0 & \text{if } j \notin \mathcal{N}(i, k_1) \wedge i \notin \mathcal{N}(j, k_1) \\ 0.5 & \text{otherwise} \end{cases} \quad (3.6)$$

By encoding differently similar images and reciprocally similar images, this matrix encodes more finely the adjacent information of similar images. Using this matrix, the feature of the images (nodes) are defined as h_i for the node n_i , which can be extracted from the i -th row of the symmetric adjacent matrix A^* :

$$h_i = [A_{i,0}^*, \dots, A_{i,n}^*]. \quad (3.7)$$

The features used to represent the images are based on their set of neighbors rather than the original features, in order to more easily remove hard negatives which can be occasionally similar to one image but are rarely similar to the same set of images.

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 j that are used during the aggregation step to update the node (image) feature; k_2 is lower than k_1 (usually much more lower).

Features update. Once the graph is constructed, the message propagation is performed using the aggregation scheme defined here with $h_i^{(l)}$ the feature of image i at the l -th layer:

$$h_i^{(l+1)} = h_i^{(l)} + \sum e_{ij}^\alpha \cdot h_j^{(l)}, j \in \mathcal{N}(i, k_2) \quad (3.8)$$

with $h_i^{(l)}$ regularized with L_2 norm after every message propagation on the graph.

In our experiments, as evaluated by the authors, we use two consecutive layers and the last GNN layer outputs the transformed node features $h_i^{(l)}$. In the end, the final ranking list is computed with the cosine similarity of refined features.

This method exploits the manifold of the dataset and is very efficient because the message propagation is concurrent between all nodes. The high-parallelism GNN propagates the message on the sparse graph efficiently, the whole dataset is re-ranked in one passage.

3.4 Image descriptors evaluation

In this section, we evaluate some state-of-the-art descriptors presented earlier to define which descriptor is the most suited for the iconographic heritage contents we have selected in this thesis.

3.4.1 Descriptors evaluated and evaluation choices

3.4.1.1 Evaluation choices

Training strategy. First of all, we decided in our thesis not to retrain or fine-tune existing networks but rather use pretrained weights. Indeed, as presented in Section 2.2 of Chapter 2 the ever-changing aspect of heritage iconographic content collections, due to ongoing and increasing digitization presents difficulties for training networks.

A first aspect, inherent to iconographic heritage collections, is the sparsity both inside each collection and between collections. Each collection often has a specific object or area of study, or a specific acquisition protocol which makes linking between collections complex. Thus, automatically creating a large enough dataset for training, especially a correct ground truth is very complicated and not realistic.

A second aspect is the continuous change of the available data in digital humanities due to the digitization, which would require to fine-tune networks on each new data to be able to capture its new specificities.

For those reasons, the networks used are the implementations of the authors and the weights exploited are the ones provided by the authors.

Evaluation framework. The first experiments were run on a RTX 3060 GPU with 12 Go RAM and 4 CPU cores. Most of the experiments of this thesis are then run on a

Tesla-V100 GPU with 16 Go RAM and 10 CPU cores.

Evaluation metric. We evaluate the efficiency of the approaches mostly with the mean Average Precision score (mAP); the implementation used is from (Radenovic et al., 2019)¹.

3.4.1.2 Evaluation datasets

During the quite lengthy process of creating the complete dataset as presented in Chapter 2, we started our evaluations on smaller subsets of the dataset.

The first one, called DB_{small} afterwards, consisted of only 1306 images instead of 1637, coming from only five providers instead of eight and without distractors. The main differences with the final one are for one the even smaller part of heritage iconographic content compared to more recent Stereopolis and Paris6K data and second the fact that none of the Albert Kahn data was in high resolution, which was more detrimental to How-A due to the fact that it is a local descriptor.

The second one, called DB_{large} , is comprized of the 1637 selected images in the 31 classes described earlier, and includes DB_{small} .

The final one, called $DB_{large+dist}$, regroups the 9834 images, aggregating DB_{large} and the numerous distractors.

3.4.1.3 Evaluated descriptors

Four state-of-the-art image descriptors were evaluated, all methods are deep detectors and descriptors:

1. **DELG** (Cao et al., 2020), a global descriptor, trained on Google Landmarks Dataset v2. Out of a first set of image features, it produces a global descriptor using GeM pooling and simultaneously extract local features using an attention-based process. The local features are used in an integrated re-ranking process for the images retrieved using the global descriptor. The global descriptor was evaluated on our specific dataset.
2. **R101-GeM** (He et al., 2016; Radenovic et al., 2019) was also tested. Trained on Google Landmarks Dataset v2, it produces a global descriptor simply by combining a ResNet 101 backbone and using GeM as a pooling function.
3. **How** (Tolias et al., 2020) showed promises when used with heritage iconographic content as presented in (Gominski et al., 2021). Trained on the *SfM120k* dataset (Radenovic et al., 2019), it exploits attention to produce local features. The matching used is ASMK (Tolias et al., 2016a) which exploits a codebook of visual words created on *SfM120k* to evaluate the similarity between images. This descriptor will be called How-A for the remainder of the manuscript.

¹<https://github.com/filipradenovic/cnnimageretrieval-pytorch>

4. **CVNet-Global** (Lee et al., 2022) appeared as a new state-of-the-art for image retrieval, also trained on Google Landmarks Dataset v2. CV-Net makes in one pass image retrieval and re-ranking. For this evaluation, we extracted the global features and used them for image retrieval.

Table 3.4 resumes all mAP scores for all the tested descriptors. The two next sections will further justify our final choice of descriptor.

Table 3.4: mAP score of tested image descriptors

	DB_{small}	DB_{large}	$DB_{large+dist}$
DELG (Cao et al., 2020)	53.2	-	-
R101 - GeM (He et al., 2016; Radenovic et al., 2019)	57.9	53.3	38.5
How + ASMK (Tolias et al., 2016a, 2020)	53.8	55.1	41.0
CV-Net global (Lee et al., 2022)	-	67.3	37.1

3.4.2 Choice between How-A, R101-GeM and DELG

A first evaluation was performed on the DB_{small} with the first three image descriptors. As shown in Table 3.4, R101-GeM outperforms largely the other global descriptor DELG. It also outperforms How-A, by almost as much. Because of our very specific dataset, we compared more in depth How-A and R101-GeM in order to base the choice of descriptor not only on a global mAP score. This more in-depth evaluation was motivated by the fact that a local descriptor appeared more suited to retrieve images with a high variability in terms of levels of details in the contents. Furthermore, our objective is interlinking connections, thus the descriptor chosen must be efficient both intra-provider and inter-providers. This more in-depth evaluation was motivated by the fact that a local descriptor appeared more suited to retrieve images with a high variability in terms of levels of details in the contents.

To evaluate this, inspired by confusion matrices, inter-provider mAP scores were computed. An ideal descriptor for our problem should perform retrieval with the right trade-off between intra-provider and inter-provider retrieval (between the diagonal terms and the others). In the following Tables (3.5 and 3.6 but also all others of the kind), the mAP is each time computed using all the queries from the provider in the first column and for whom the positives are the images of the same class and from the provider in the first row. In all following tables, FD refers to the Full Dataset (being either DB_{small} , DB_{large} or $DB_{large+dist}$ depending on the case).

Comparing results of the Tables 3.5 and 3.6, we can identify several key aspects that helped decide on the best descriptor for our dataset. First of all, we see that for R101-GeM (Table 3.6), the diagonal terms are much more salient, indicating that the retrieval is efficient to find correct images from the same provider but in comparison less efficient to find correct images from another provider, thus probably with a different visual aspect. Secondly, we will focus on the two starting providers of the dataset (Parisienne de la

		Retrieved Image's Provider						
		FD	Sp	PdP	AK	CVP	P6k	Mean
Query Provider	FD	53.75	39.61	10.55	20.38	13.73	13.27	
	Sp	65.02	76.40	05.72	07.61	06.12	08.29	25.16
	PdP	55.21	20.86	43.14	20.09	09.30	12.61	25.27
	AK	29.54	04.43	02.02	45.58	02.96	03.60	13.05
	CVP	55.69	19.82	07.29	12.59	47.18	12.68	24.05
	P6k	51.85	15.33	04.55	15.52	09.66	29.82	19.71
	Mean		27.37	11.17	19.39	13.71	12.80	

Table 3.5: mAP provider vs provider with How-A descriptor

		Retrieved Image's Provider						
		FD	Sp	PdP	AK	CVP	P6k	Mean
Query Provider	FD	57.86	42.58	11.82	24.07	14.75	12.92	
	Sp	64.02	80.53	04.63	07.30	05.82	06.31	28.10
	PdP	49.34	17.10	50.13	18.20	06.90	07.53	24.87
	AK	55.62	12.54	06.04	61.80	05.19	06.54	24.62
	CVP	59.14	20.82	06.61	12.38	53.64	11.19	27.30
	P6k	50.02	16.11	04.21	15.54	09.41	29.45	20.79
	Mean		31.61	13.91	23.21	15.95	12.32	

Table 3.6: mAP provider vs provider with R101-GeM descriptor

Photographie and Stereopolis) to evaluate the efficiency of the inter-provider retrieval that is paramount for collection interlinking. Indeed, what we are looking for is a descriptor that ensures that all providers will retrieve the starting providers and vice-versa. The comparison can be done on the two columns and two lines for Stereopolis (Sp) and the Parisienne de la Photographie (PdP):

- with Sp as a query, How-A slightly outperforms R101-GeM,
- when PdP is a query, How-A outperforms R101-GeM even more,
- for the column with Sp as the retrieved images, How-A performs slightly worse, with the particular case of AK which we explain later,
- as for PdP as the retrieved images, How-A performs slightly better, with the same exception for AK.

Globally, it appears that How-A performs better than R101-GeM if we consider the intra-provider aspect of the retrieval, which is paramount in our study. Third, we want to reiterate the fact that all results based on Albert Kahn (AK) images are biased due to the low resolution of these contents at the time of the experiments, which is more detrimental for How-A as it is a local descriptor.

To further validate the choice of How-A, we looked further into specific classes of the dataset that prove to be difficult or ambiguous for image descriptors due to their similarity

to another class or to the very small detail that represents them. We used class-based mAP scores, meaning that queries and positive images can come from any provider but only from one specific class. The two following qualitative examples further validate the use of How-A in our case.

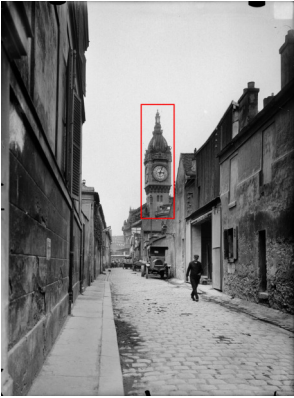
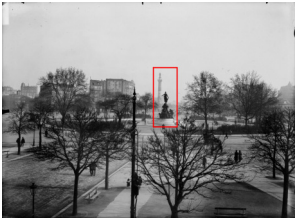

Class	Gare de Lyon	Nation	Sacré Coeur
Example			
Detail	Clock tower	Fountain and column	Dome in the background
mAP w/ How-A	92.9	46.3	67.6
mAP w/ R101-GeM	76.3	38.9	53.3

Table 3.7: mAP comparison for classes with specific small details

The examples in Table 3.7 show that for classes whose main element is often a small detail of the images, How-A performs better than R101-GeM.



Class	Invalides	Val de Grâce
Example		
mAP w/ How-A	44.5	48.2
mAP w/ R101-GeM	35.6	33.8

Table 3.8: mAP comparison for classes with great visual similarity

In Table 3.8 we compare the mAP for two classes which are quite similar. Indeed, the main element in both pictures is the dome. In both cases, it is in the background with

buildings in the forefront. Furthermore, both domes are quite similar, even to a human observer. Both classes could be confused with each other. However, it appears once more that How-A is more discriminative than R101-GeM, especially when confronted to small details.

Going even further than those experiments, as shown in Table 3.4, we evaluated both descriptors with DB_{large} and $DB_{large+dist}$. In both cases, How-A performs better than R101-GeM, showing its higher robustness to the added distractors.

This detailed analysis of performances between How-A and R101-Gem led us to finally select How-A for further experiments as it corresponds better to our needs in terms of discriminativeness and inter-provider retrieval efficiency. Both those characteristics are essential for our goal of interlinking widely diverse heritage image collections.

3.4.3 Choice between How-A and CV-Net

Even though How-A appeared most suited for our needs, CV-Net (Lee et al., 2022) appeared in 2022 as a new state-of-the-art with a substantial mAP gain. Therefore, we evaluated it against our dataset to keep using the best descriptor as a baseline for our retrieval process. At the time, CV-Net and How-A were evaluated on both DB_{large} and $DB_{large+dist}$.

When evaluated on DB_{large} , CV-Net largely outperforms How-A with a 12.2% global mAP gain as shown in Table 3.4. When comparing its inter- and intra-provider retrieval capacities against those of How-A in Tables 3.9 and 3.10, several conclusions can be made:

- CV-Net is largely better in terms of intra-provider retrieval,
- in terms of inter-provider retrieval, several points should be noted:
 - with PdP or AK (HD and BD) (the two main heritage collections) as the retrieved images, How-A outperforms CV-Net for 5 out of 9 providers,
 - when PdP images are the queries, How-A outperforms or is on par with CV-Net for 6 out of 9 providers.

This observations does not completely justify the fact to choose How-A over CV-Net, even though the performance is a little better with the two biggest "heritage-like" providers. However, when adding the distractors, the performance of CV-net drastically drops, scoring almost 4% below that of How-A. Indeed, being a global descriptor, CV-Net appears less efficient at dealing with the noise brought on by the distractors depicting similar scenes all over Paris.

Thus, despite CV-Net being a strong candidate, for the remainder of our experiments, we decided to keep working with How-A because of its adequate trade-off between intra-provider and inter-providers retrieval efficiency and its high degree of discriminativeness that helps dealing with distractors.

		Retrieved Image's Provider										
		FD	Sp	PdP	AK HD	MAP	CVP	CA	COARC	AK BD	P6k	Mean
Query Provider	FD	55.1	37.5	10.0	17.9	10.3	11.3	01.6	18.0	06.5	11.7	
	Sp	62.0	75.8	04.8	06.4	04.8	05.5	00.7	07.2	02.5	07.8	12.8
	PdP	56.1	17.4	35.8	16.5	15.7	07.5	02.1	17.8	06.9	10.3	14.4
	AK HD	52.7	09.5	06.4	51.4	08.2	04.4	01.1	10.8	05.8	07.5	11.7
	MAP	48.2	12.2	11.6	15.7	22.1	06.0	01.6	10.1	03.4	10.2	10.3
	CVP	53.9	17.2	07.0	09.3	06.9	44.7	02.7	11.6	05.7	10.2	12.8
	CA	48.7	11.9	08.3	16.6	10.7	11.2	15.7		07.2	06.3	11.0
	COARC	76.2	12.5	06.7	10.8	10.5	04.3		64.4	01.0	37.5	18.5
	AK BD	33.0	06.8	02.5	16.1	04.1	04.1	00.9	01.6	48.6	04.3	09.9
	P6k	50.8	14.0	04.0	13.5	07.7	07.9	01.2	16.9	03.2	23.6	10.2
	Mean		19.7	09.7	17.4	10.1	10.6	03.2	17.5	09.4	13.0	

Table 3.9: mAP provider vs provider with How-A descriptor on Full dataset without distractors

		Retrieved Image's Provider										
		FD	Sp	PdP	AK HD	MAP	CVP	CA	COARC	AK BD	P6k	Mean
Query Provider	FD	67.3	38.9	12.4	20.4	17.4	13.8	01.8	20.7	06.1	14.7	
	Sp	67.3	85.6	04.1	06.5	05.8	06.0	00.3	08.8	01.6	10.6	19.7
	PdP	63.6	16.3	51.3	09.9	28.1	07.7	01.9	17.8	04.6	08.6	21.0
	AK HD	67.7	12.6	05.5	71.9	10.2	05.7	00.9	09.7	07.2	08.7	20.0
	MAP	66.7	13.8	17.5	11.0	45.3	07.0	02.4	16.4	03.4	09.6	19.3
	CVP	74.4	22.5	06.7	09.2	11.2	58.5	03.0	14.9	07.1	12.3	22.0
	CA	53.5	09.4	09.4	14.6	23.3	11.6	23.7		03.5	04.5	17.1
	COARC	89.1	16.6	06.5	10.1	16.3	03.1		73.7	01.4	37.2	28.2
	AK BD	60.9	10.2	06.2	38.2	09.1	07.3	01.1	03.3	41.7	07.9	18.6
	P6k	65.5	20.5	03.0	13.9	09.3	08.6	00.7	11.9	03.5	32.8	17.0
	Mean		24.6	12.3	20.6	17.6	12.9	04.0	19.7	08.0	14.7	

Table 3.10: mAP provider vs provider with CV-Net global descriptor on Full dataset without distractors

3.5 Re-ranking methods evaluation

In this section, we evaluate all paradigms of re-ranking presented in 3.3 in order to find the most relevant ones for our problem of automatic content linking using image retrieval on a challenging dataset.

3.5.1 Re-ranking choices

As presented in Section 3.3, 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.

For all trained methods, the choice was made not to retrain or fine-tune the networks, for reasons similar to those explained in Section 3.3. Indeed, retraining networks is quite costly in both computation resources and in annotation time to create an adequate dataset and a ground truth. Furthermore, as the digitization process and the linking of contents is an ongoing process, fine-tuning would only help temporarily and for each new collection, the network would have to be retrained. Thus, our thesis focuses on off-the-shelf networks and how to exploit them to the maximum without retraining.

Table 3.11 shows the performance of these approaches, by providing an idea of the improvement in terms of mAP when exploited as a single re-ranking step. For example, geometric verification with RANSAC in average improves results by 0.5-1.5 point of mAP. Further details on experiments and results explanations are presented in the next sections.

Table 3.11: Modification of mAP score with re-ranking

Approach	Order of magnitude
Weighted descriptor aggregation (Chum et al., 2007; Radenovic et al., 2019)	+ 0.1
Pseudo relevance feedback (Lin, 2019)	< + 0.5
Transformers-based : CSA (Ouyang et al., 2021), RRT (Tan et al., 2021)	- 10
Geometric Verification : RANSAC (DeTone et al., 2018; Sarlin et al., 2020) CV-Net Rerank (Lee et al., 2022)	+ 0.5-1.5 - 2
Diffusion (Shen et al., 2021; Zhang et al., 2020b)	+ 16

3.5.2 Aggregation

In terms of descriptor aggregation approaches, AQE and α -QE have been tested on the DB_{small} . For each, the number of aggregated descriptors varied between 5 and 20, and the value of α between 1 and 5. The experiments were run using descriptors from R101-GeM, DELG and How.

Independently of the descriptors, the final mAP gains were minimal, inferior to 0.1%. In the case of the How descriptor, the fact that it is based on multiple local features explains why an aggregation approach is not best suited. However, in the case of global descriptors like R101-GeM or DELG, the explanation behind this very slim mAP gain appears to be the high degree of variability of the visual contents. Indeed, global descriptors are less suited for retrieval in a dataset with a high degree of changes in color, viewpoint, detail level, illumination,... Thus, the aggregation of the global descriptors

does not encapsulate enough subtil information required to bridge the gap induced by the visual disparities.

As for learned aggregation paradigms, no pre-trained networks or code releases were available and we did not explore this option further.

3.5.3 Pseudo-relevance feedback

The pseudo-relevance feedback method was tested on the DB_{small} , concurrently with the classical descriptor aggregation approaches. We chose to test (Lin, 2019) as its algorithm was detailed and allowed for freedom to test several choices in the weighting paradigm. Indeed, it exploited the Rocchio algorithm (Rocchio Jr, 1971) for descriptor aggregation in order to get "closer" to positive images' descriptors and further from negative ones.

Exploiting simply this aggregation approach proved slightly better than simple aggregation. Indeed, by tweaking the number of pseudo-correct and pseudo-incorrect images and the weights associated, we managed to obtain a mAP gain of about 0.5% which remains quite small after so much adaptation.

3.5.4 Transformers-based

For transformers-based approaches, we tested two available methods: CSA (Ouyang et al., 2021) and RRT (Tan et al., 2021) as pure re-ranking steps on DB_{large} . Both approaches required global features. We tested them with respectively R01-GeM and DELG descriptors, following the authors experiments. However, the re-ranking step in both cases was highly detrimental in terms of mAP (about a 10% drop). In both cases, it appears that the networks should be retrained with data more similar to our own. Indeed, trained with Google Landmark Dataset v2 which is a recent dataset in terms of image quality and visual aspect, the networks do not deal well with the high visual variability of our dataset.

3.5.5 Geometric verification

For geometric verification purposes, we exploited a classical RANSAC implementation for finding an affine transformation between two images. The process is similar regardless of the local features or the matching scheme selected.

Based on work previous to this thesis, we chose to exploit SuperPoint features (DeTone et al., 2018) and SuperGlue (Sarlin et al., 2020) as a matching scheme. Indeed, they proved to be the most efficient when applied to heritage content. Furthermore, they were integrated in the Hierarchical Localization process of (Sarlin et al., 2019) whose approach was very inspiring in terms of adapting the retrieval to the downstream task desired. Further along, as LightGlue was proposed in 2023 (Lindenberger et al., 2023), we decided to test it as an alternative for matching due to it being an improvement of SuperGlue.

Table 3.12 presents the mAP increases due to classical geometric verification on the $k = 135$ first similar images. The mAP gain is higher than previous re-ranking methods

tested. On DB_{large} , the mAP gain is of 1.7% when using SuperGlue. When the distractors are added ($DB_{large+dist}$), the mAP gain is less, only 0.5%. However, changing the matcher with the new and improved LightGlue brings the score up by 0.9%. The mAP gain is not extremely high, however as we will discuss later in Chapter 5, the real gain due to geometric verification will appear when combined with other re-ranking approaches exploiting extended information to globally improve the retrieval, justifying this more in-depth testing.

Table 3.12: Geometric verification on the 135 first similar images

mAP	DB_{large}	$DB_{large+dist}$
How-A	55.1	41.0
How-A + RANSAC-SuperPoint/SuperGlue (DeTone et al., 2018; Sarlin et al., 2020)	56.8	41.5
How-A + RANSAC-SuperPoint/LightGlue (DeTone et al., 2018; Lindenberger et al., 2023)	-	41.9

The final geometric verification approach we tested is the re-ranking part of CV-Net (Lee et al., 2022). We used it directly after the retrieval step from CV-Net as it exploits features extracted by CV-Net. Whereas CV-Net may perform better than other descriptors in terms of retrieval, its re-ranking step however was detrimental to the whole process with a mAP loss of about 2%, which can be explained by the high visual variability that is not suitable for its dense correlation approach.

3.5.6 Diffusion methods

To evaluate the benefits of using diffusion methods for re-ranking, we tested several approaches exploiting in different ways the set of nearest neighbors of each query to exploit the global retrieval context.

First, two different types of approaches were compared. The comparison of the k-reciprocal sets of nearest neighbors from (Zhong et al., 2017) on one side and the graph-oriented approach of GNN-Reranking from (Zhang et al., 2020b). This test exploits the same retrieval results as an input and is performed on the DB_{large} . While the mAP gain from both methods is high, we obtain +14.1% for k-reciprocal and +16.9% for GNN-R, GNN-R outperforms the other approach.

Second, we compare our leading method GNN-R with the similar approach SSR (Shen et al., 2021). Indeed both approaches exploit the retrieval results as a graph. In one case, GNN-R exploits the k-nearest neighbors graph as a whole while SSR exploits subgraphs which are optimized independently before creating a new whole similarity graph. Tested on the $DB_{large+dist}$, without any retraining, both methods perform similarly with an order of magnitude of 16% (15.9% for SSR and 16.2% for GNN-R). While GNN-R outperforms only with a very small margin, we decided to keep using it going forward for another

specific reason: as it will be shown in Chapter 5, multiple runs of GNN-R can be done successively, further improving the re-ranking.

3.6 Conclusion

This chapter has provided an overview and an evaluation of image retrieval descriptors and re-ranking methods.

Table 3.13: Summary of the evaluated descriptors and re-ranking methods

Descriptor evaluation		
Descriptor	Specificities	Evaluation
DELG	Global descriptor Trained on GLDv2	Lowest performance on the simplest dataset
R101-GeM	Global descriptor Trained on GLDv2	Better than DELG Less efficient inter-provider Less robust to distractors
CV-Net	Global descriptor Trained on GLDv2	Largely better without distractors Slightly less efficient inter-provider Not robust to distractors
How-A	Local descriptor Trained on SfM120k	Not the highest score to start More robust to distractors Better inter-provider retrieval
Re-ranking approaches evaluation		
Re-ranking method	Principle	Evaluation
Weighted descriptor aggregation	Aggregates most similar image's descriptors	Almost no mAP improvement Visual heterogeneity is too great
Pseudo relevance feedback	Aggregates image descriptors to be closer to pseudo-similar images and further from pseudo-dissimilar images	Low mAP improvement Visual heterogeneity is too great
Transformers-based approaches	Exploits transformers Uses global and local features or a global ranking list	Detrimental to the mAP Need to be retrained
Geometric verification: RANSAC	Estimates an affine transformation between images using matches between local features	Low mAP improvement Useful in the rest of our experiments Multiple potential features and matchers
Geometric verification: CV-Net	Trained approach using dense cross-scale feature correlation to assess geometric coherence between images	Detrimental to the mAP Needs to be retrained
Diffusion methods	Exploits the graph of nearest neighbors to propagate the similarities between images	Great mAP improvement Choice of GNN-R because it can be repeated multiple times.

First, it presented a state-of-the-art of existing descriptors for image retrieval. Second, a review of the multiple approaches to re-ranking is detailed. Both aspects of image retrieval are then evaluated on our dataset in order to evaluate their relevance in our research, as summarized in the following Table 3.13. Methods highlighted in bold are the one used in our next experiments.

After existing methods have been evaluated, several insights can be exploited to devise new approaches for re-ranking. The following chapter will focus on presenting our proposed re-ranking methods, exploiting both insights from existing methods and the knowledge of our specific dataset.

Chapter 4

Our Contributions to Re-ranking

4.1	Introduction	85
4.2	Geometric query expansion	86
4.2.1	A 3D based proof of concept	86
4.2.2	A 2D geometric query expansion proposition (R2D)	90
4.3	Metadata exploitation	96
4.3.1	Metadata structure weighting scheme	96
4.3.2	Structuring with location information	98
4.4	Conclusion	99

4.1 Introduction

The previous chapter introduced and evaluated a large state of the art of the whole content-based image retrieval pipeline. First, the most suited image descriptors were compared with regard to our specific dataset and the objectives of our study, that is interlinking content within and between diverse collections. Afterwards, the second part of the pipeline namely the re-ranking step was evaluated. Many re-ranking paradigms have been devised, and we tested methods for all of them to evaluate how they could suit our needs for improving a first retrieval step that proves to be difficult due to the specific data considered.

This first study provided several insights that we leveraged in order to propose new re-ranking methods suited to the data we work with. First, aggregation methods perform poorly mainly due to the high visual variability in the data. Second, classical pairwise geometric verification approaches does not perform as well as commonly observed in the literature. Finally, diffusion methods, exploiting a large retrieval context, in a way a similarity structure, prove to be the methods performing best with our data.

Those insights led us to explore the use of a more global structure either at query level or later at dataset level when performing re-ranking. The idea is to leverage a more global similarity and alleviate the issues introduced by the high visual variability in our dataset when performing simple pairwise approaches.

We will first introduce in Section 4.2 a geometric query expansion step to build on classical geometric verification, using the geometric information contained in the first retrieved images to enrich that of the query image before performing pairwise geometric verification. We will then in Section 4.3 introduce a dataset-wide weighting scheme based on location weighting which leverages a global structure of the dataset to validate or invalidate retrieval results. All those methods will be evaluated later in Chapter 5.

4.2 Geometric query expansion

This section introduces our proposed contributions for geometric query expansion. First, Section 4.2.1 details a first 3D reconstruction-based approach that leverages an exact geometry of the scene. Second, Section 4.2.2 proposes a version of the first approach based on a 2D approximation of the scene’s geometry, less accurate but computation-wise less costly.

4.2.1 A 3D based proof of concept

The objective of this approach is to exploit geometric information more global than simply the geometric information contained in the query alone. Indeed, when levels of detail greatly vary between images, the number of matched points can be very small and yet be a very good match, which is hard to evaluate without any priors on the images. A first idea to enrich the geometric information related to an image is to exploit the 3D scene. However, this information is not available directly with our dataset, especially with heritage content. We propose an approach to reconstruct this scene and evaluate the geometric coherence of the retrieved images in this scene, we call it R3D.

4.2.1.1 3D scene reconstruction (R3D)

A main aspect in our dataset is the very large variation in viewpoint and level of detail. If it impairs classical geometric verification, it also means that 3D reconstruction can be considered.

Our reconstruction process is widely inspired by the hierarchical localization toolbox (Sarlin et al., 2019). We first extract keypoints in images using SuperPoint (DeTone et al., 2018), then match them using SuperGlue (Sarlin et al., 2020) or LightGlue (Lindemberger et al., 2023). Those points and matches are then fed to the Colmap Library (Schönberger and Frahm, 2016; Schönberger et al., 2016) for 3D scene reconstruction using Structure-from-Motion algorithms.

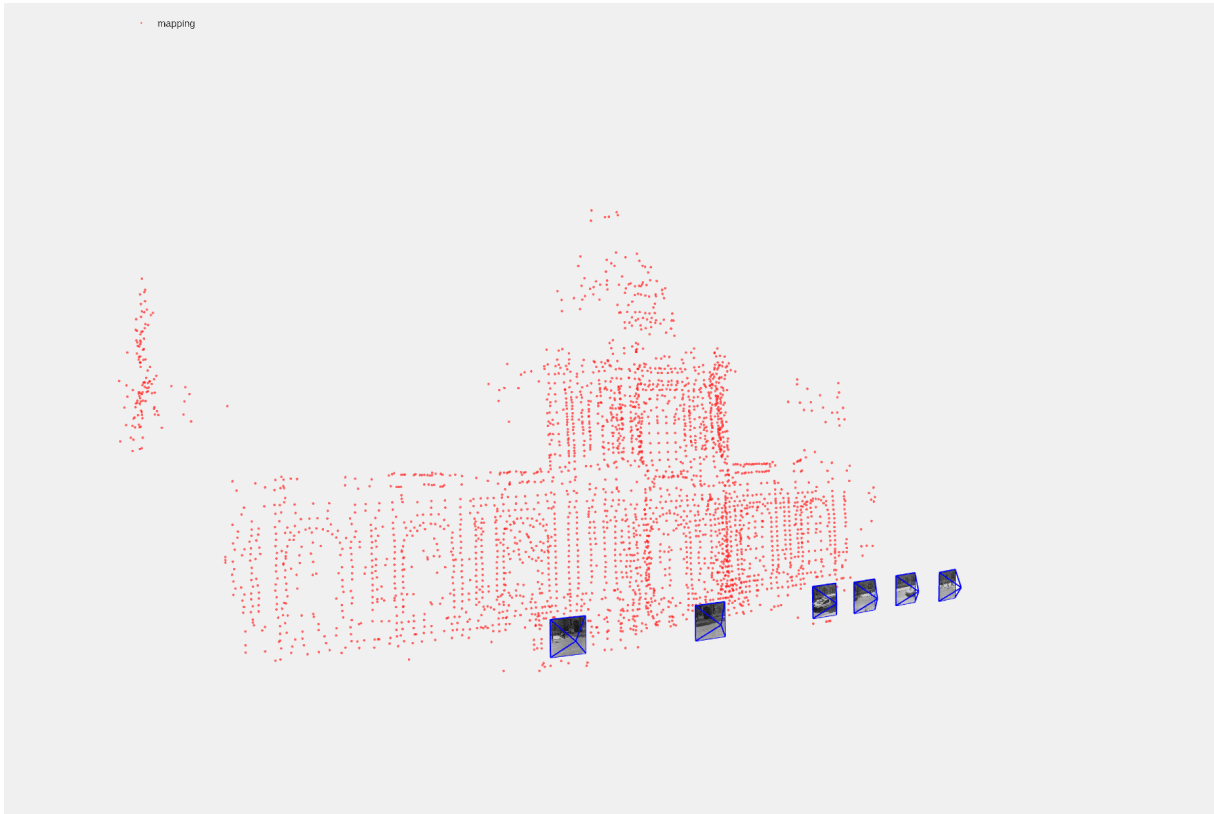
We do not particularly set specific parameters (maximum reprojection error for instance) to Colmap as we wish to obtain quality scenes. However, we had to use our keypoints detectors and matchers (SuperPoint and SuperGlue/LightGlue) as Sift proved to be completely unsuited for heritage content. Furthermore, this allowed us to remain coherent when comparing with classical geometric verification. Due to the difficulties at

the retrieval step, we allowed for a model to consist of a minimum of two images, in order to make sure that if a query manages to retrieve only one similar image in the set of images used for the reconstruction, a 3D scene, even a small one can be produced.

Our evaluation protocol is that for each query, itself and the ten first retrieved images are given to the 3D reconstruction. A model is then selected as correct if it consists of at least two images and includes the query. Otherwise, no model is selected and the images are not re-ranked. An example of a reconstructed model is shown in Figure 4.1 alongside the ten images given as input. One can notice that out of the ten images, only six are used and those are images from the mobile mapping Stereopolis dataset, well suited for this type of reconstruction.



(a) The ten images given as input to the SfM process



(b) The reconstructed 3D point cloud, using 6 out of 10 images (the highlighted ones)

Figure 4.1: Illustration of the 3D reconstruction

4.2.1.2 Image relocalization

Once a 3D reconstruction of the scene is reconstructed, the k first retrieved images are then evaluated against this model to check their geometric coherence, even images that potentially were used to create the model.

To evaluate the geometric coherence of an image I in the 3D model, its keypoints are extracted and matched against those used for the reconstruction, using the same extract-

ing and matching protocol as before. The matching is first performed in 2D between I and the images used to create the model. Then, keypoints of I matched to keypoints of the reconstruction's images that were projected in 3D are matched to those 3D points, conducting to 2D/3D matches. Using these matches, we try to estimate the pose of the image in the scene using Colmap and its 2D-3D registration algorithm. If an image is not re-positioned due to a lack of matches with the images used to create the model, its geometric coherence score will be null. However, if it is re-positioned, its coherence with the scene will be evaluated using a specific score, detailed below.

The geometric coherence score S_I is comprised of two parts. First a score evaluating the quality q of the reconstruction. This score (see Equation 4.1) uses the number of images used originally to perform the reconstruction n , the number of 3D points in the reconstruction p and the mean number of 2D/3D matches m per image used for the reconstruction:

$$q = np^{p/m}. \quad (4.1)$$

This quality score aims at promoting reconstructions with the largest number of images and 3D points but also reconstructions which extend the most the 3D scene. Indeed, if the mean number of points per image is the same as the number of points in the reconstruction, it means that not much geometric information was added to the scene whereas the opposite means that the scene extended further than what any single image "sees".

Once this quality score for the reconstruction is computed, the final score S_I used for the re-ranking is computed as:

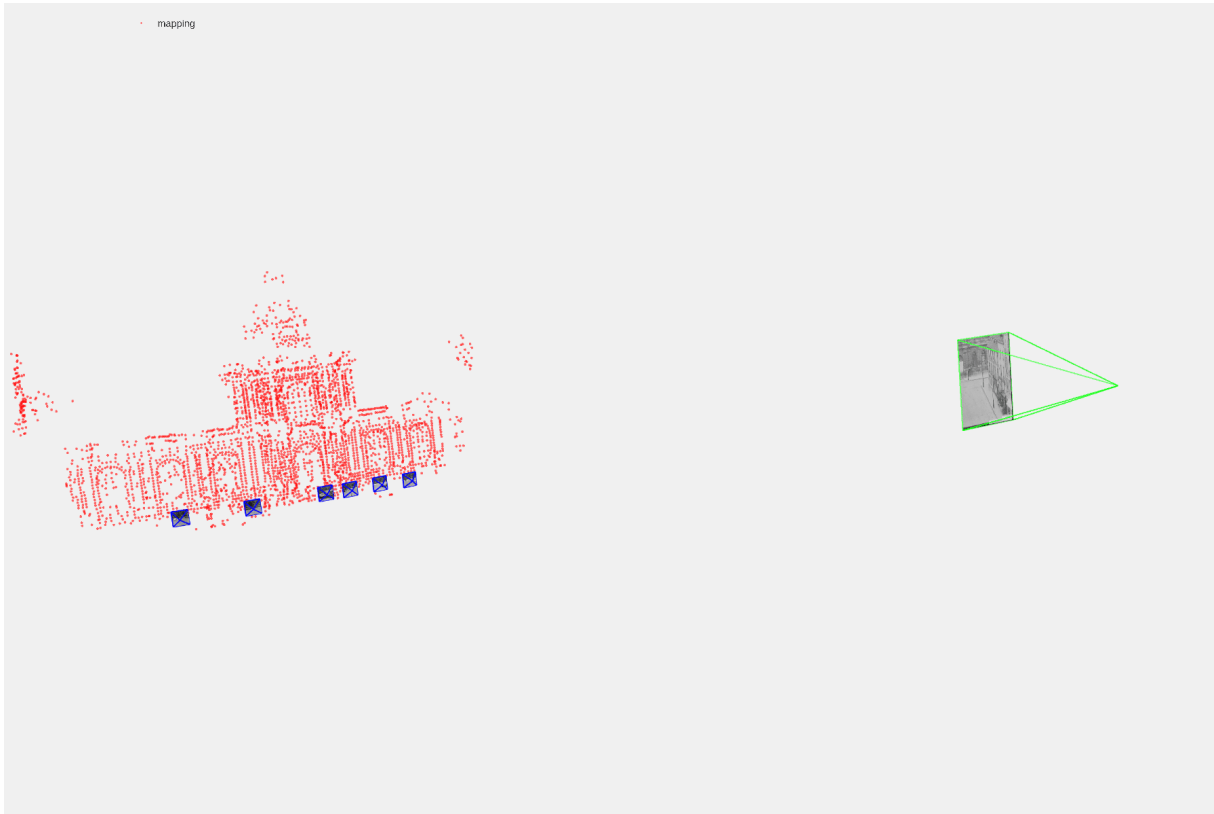
$$S_I = qp_I, \quad (4.2)$$

with p_I the number of points kept as inliers when estimating the pose of the re-ranked image I .

Two visual examples of image relocalization are shown in Figure 4.2 and 4.3. In the first example, the image represents the 3D scene, its relocalization produces a pose that is coherent, facing the scene. The second example however presents the relocalization of a totally different image. The pose obtained is thus degenerate, "inside" the 3D reconstruction, not coherent with the scene.



(a) The image (corresponding to the scene) to relocalize

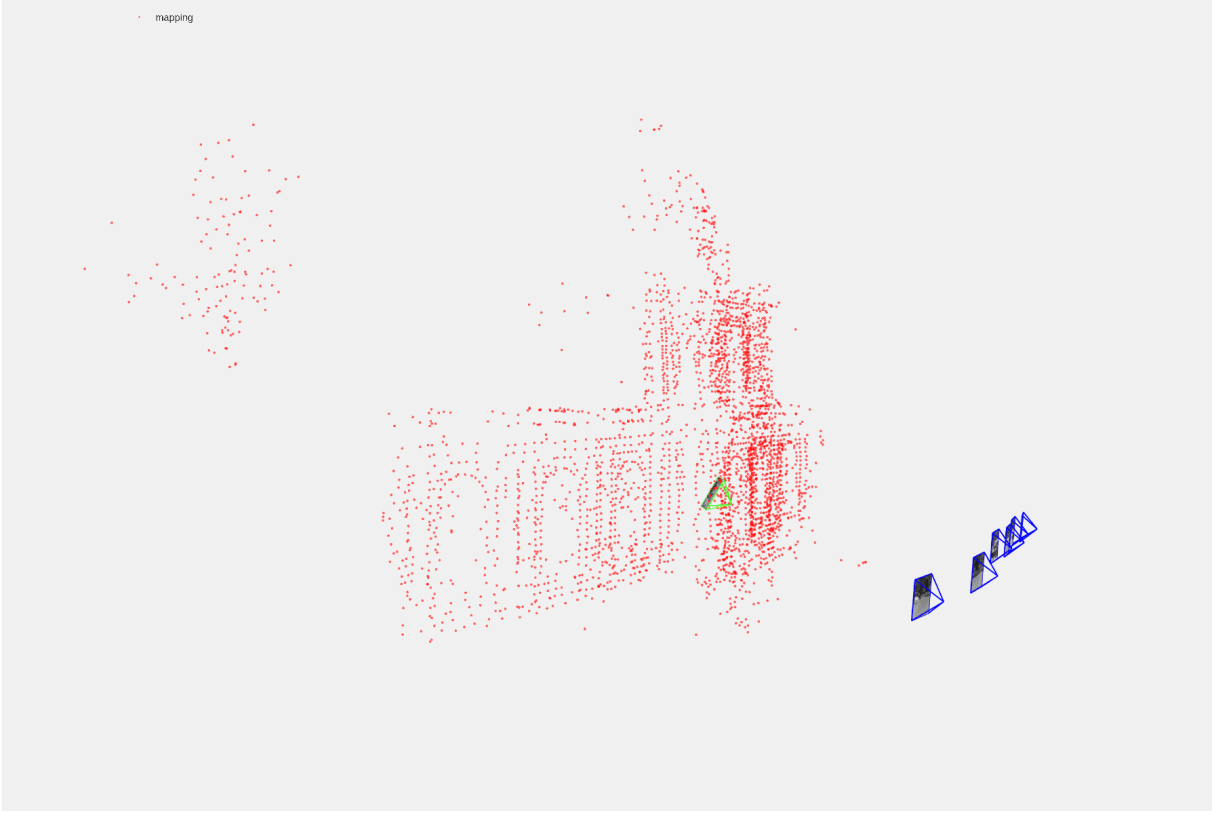


(b) The image localized (in green) in the 3D reconstruction, the pose estimation is coherent in the scene

Figure 4.2: Illustration of the relocalization of a correct image



(a) The image (not corresponding to the scene) to relocalize



(b) The image localized (in green) in the 3D reconstruction, the pose estimation is incoherent

Figure 4.3: Illustration of the relocalization of an incorrect image

4.2.2 A 2D geometric query expansion proposition (R2D)

As it will be evaluated in Section 5.2.1 of Chapter 5, using a 3D reconstruction proves to be a very efficient re-ranking method to improve retrieval. However, it is also very time-consuming (2 times more than classical geometric verification with RANSAC). To alleviate this drawback, we propose another approach for geometric verification that tries to get the best of both worlds. On one side, exploiting the global scene to expand the geometric information with which to compare retrieved images' geometry. On the other side, preventing the computational overhead due to the 3D reconstruction and the repositioning of all images. This approach can be apprehended as a 2D geometric query expansion as the idea is to aggregate the 2D geometric information from the first retrieved images to enrich that of the query image. Thus, the classical geometric verification is between an image and a geometrically extended query. We call this approach R2D.

4.2.2.1 2D point set creation

As previously explained, the objective is to enrich the set of 2D points of the query. The subsequent matching between the query and a retrieved image exploits the 2D geometry of a retrieved image and a 2D approximate representation of the scene of the query.

Building on the visual query expansion paradigm that fuses visual descriptors, our idea is to exploit the features extracted in similar images by reprojecting them in the query image to enrich its geometric significance and artificially enlarge the scene representation encoded. Thus, the query does not only encode the geometry of the scene it depicts, but also parts of the scene depicted by its most similar images. This way, the various viewpoints of the same scene add to the geometric of the scene in an approximate 2D reconstruction of the scene.

The first step of the proposed approach is the creation of a new set of keypoints for the query. To do this and remain comparable to the 3D reconstruction-based approach, we use the first ten retrieved images and the query to estimate the enriched set of keypoints of the query. We first extract keypoints in images using SuperPoint (DeTone et al., 2018), then match them using SuperGlue (Sarlin et al., 2020) or LightGlue (Lindemberger et al., 2023).

To ensure a certain quality in this geometric enrichment, we set some constraints that are described below and illustrated in Figure 4.4.

The first step consists in creating all triplets containing 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: $M_{Q,I_1}, M_{Q,I_2}, M_{I_1,I_2}$ illustrated in Figure 4.4 (a),
- define the query’s **solid matches** as:

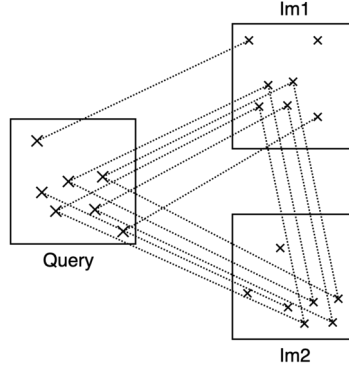
$$K_Q^s = \{k \text{ if } M_{I_1,I_2} \circ M_{Q,I_1}(k) = M_{Q,I_2}(k), \forall k \in K_Q\} ,$$
- define the query’s **uncertain matches**:

$$K_Q^u = \{k \text{ if } k \notin K_Q^s, \forall k \in K_Q\} ,$$
- 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:

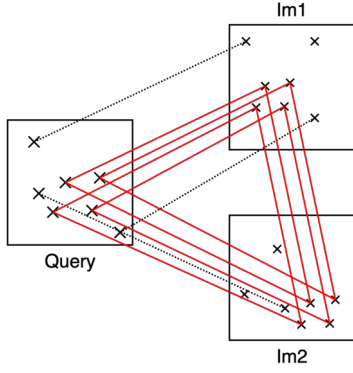
$$K_Q^h = \left\{ h_{I_1,Q}(k) \text{ if } k \notin M_{Q,I_1}[K_Q^s], \forall k \in K_{I_1} \right\} \\ \cup \left\{ h_{I_2,Q}(k) \text{ if } k \notin M_{Q,I_2}[K_Q^s], \forall k \in K_{I_2} \right\} .$$

The three types of points are shown in the example of Figure 4.4 (b), (c) and (d).

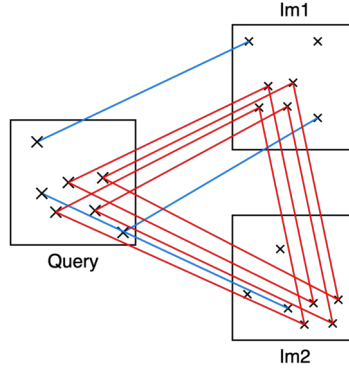
Once those steps are performed on all triplets, they are globally concatenated for each query: $K_Q^a = K_Q^s \cup K_Q^u \cup K_Q^h$, as illustrated in Figure 4.4 (e).



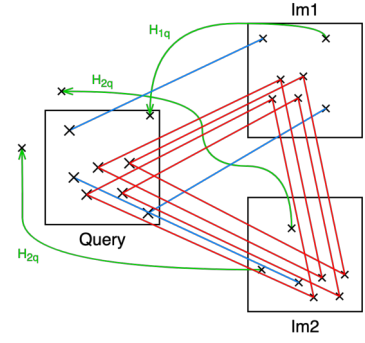
(a) Pairwise matching between images of the triplet



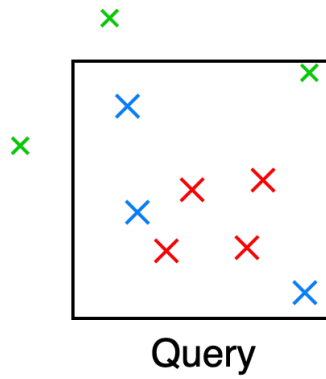
(b) Solid keypoints (red) matched in a loop pattern



(c) Uncertain keypoints (blue), simply matched



(d) Reprojection (green) of keypoints from other images to the query if enough solid matches are available



(e) Final set of points, colored with regard to their original category

Figure 4.4: Illustration of the R2D point set creation steps

4.2.2.2 Extended geometric verification

Once the new set of points is created, the re-ranking protocol is similar to the classical geometric one: the retrieved images are compared pairwise to the query image with the new set of points. Furthermore, the specificity of the points included in the query's set allows for a different, finer geometric verification than simply a classic RANSAC. The computation of the geometric similarity score $S_{Q,I}$ between the query Q and a retrieved image I is detailed below:

- match the keypoints K_Q^a and K_I (those of image I): $M_{Q,I}$,
- select the subsample of solid matches $M_{Q,I}^s$ (matches with a solid keypoint) or the subsample of solid and uncertain matches $M_{Q,I}^{s,u}$ 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,
- out of the kept matches, identify three categories of matches: $\hat{M}_{Q,I}^s$, matches with solid keypoints, $\hat{M}_{Q,I}^u$, matches with uncertain keypoints and $\hat{M}_{Q,I}^h$, matches with reprojected keypoints,
- the final score $S_{Q,I}$ is computed using the number of each type of match (kept after the RANSAC) and the subsample of points used for the RANSAC:

$$S_{Q,I} = \begin{cases} 2 \times \frac{|\hat{M}_{Q,I}^s|}{|K_Q^s|} + \frac{|\hat{M}_{Q,I}^u|}{|K_Q^u|} + \omega_s \times \frac{|\hat{M}_{Q,I}^h|}{|K_Q^h|} & \text{if } M_{Q,I}^s \text{ is used} \\ 2 \times \frac{|\hat{M}_{Q,I}^s|}{|K_Q^s|} + \frac{|\hat{M}_{Q,I}^u|}{|K_Q^u|} + \omega_{s,u} \times \frac{|\hat{M}_{Q,I}^h|}{|K_Q^h|} & \text{if } M_{Q,I}^{s,u} \text{ is used} \\ 2 \times \frac{|\hat{M}_{Q,I}^s|}{|K_Q^s|} + \frac{|\hat{M}_{Q,I}^u|}{|K_Q^u|} + \omega_{s,u,h} \times \frac{|\hat{M}_{Q,I}^h|}{|K_Q^h|} & \text{otherwise} \end{cases} \quad (4.3)$$

The re-ranking score has to distinguish between different cases in order to ensure the coherence of the approximate 2D reconstruction as detailed in Equation 4.3. Indeed, depending on the quality of the keypoints used for estimating the affine transformation between the query and the retrieved image, the validity of the transformation varies. Thus, matches with reprojected points that pass through the test using a highly certain transformation weigh more in the score than those passing through the test with an uncertain transformation.

The first weighting difference is between solid and uncertain matches as solid ones are weighed double what uncertain matches weigh.

The second weighting difference applies on matches with reprojected points. Indeed, the more they correspond to a transformation estimated without them, the more they are assumed to be correct, thus having $\omega_s \geq \omega_{s,u} \geq \omega_{s,u,h}$. Furthermore, as our objective is

to exploit those reprojected points to expand the query’s geometry, we give them weights that can be higher than solid matches weights.

Finally, as the matchers used (SuperGlue or LightGlue) have different performance to estimate correct matches, the confidence given to the reprojected points (and subsequent matches) is different depending on the matcher used. Thus, the optimal values found for our dataset are the following:

- **For SuperGlue:** $\omega_s = 4$, $\omega_{s,u} = 2$ and $\omega_{s,u,h} = 1$
- **For LightGlue:** $\omega_s = 8$, $\omega_{s,u} = 6$ and $\omega_{s,u,h} = 4$

As the ratios of matches over keypoints are all in $[0, 1]$, depending on the matcher and on the keypoints used for estimating the affine transformation, the scores of the images vary differently:

- **For SuperGlue:** $S_{Q,I} \in [0, 7]$ if $M_{Q,I}^s$ is used, $S_{Q,I} \in [0, 5]$ if $M_{Q,I}^{s,u}$ is used and $S_{Q,I} \in [0, 4]$ otherwise
- **For LightGlue:** $S_{Q,I} \in [0, 11]$ if $M_{Q,I}^s$ is used, $S_{Q,I} \in [0, 9]$ if $M_{Q,I}^{s,u}$ is used and $S_{Q,I} \in [0, 7]$ otherwise

This re-ranking score $S_{Q,I}$ is then used to re-rank the k first images in decreasing order of geometric similarity.

To summarize, this proposed approach exploits the global geometric information from the first retrieved images. It is less costly than a full 3D reconstruction while approximately encoding in 2D the geometric information of the scene. It is evaluated on our dataset in Section 5.2.2 of Chapter 5.

Level of detail and geometric verification

To go slightly further than the proposed extended geometric verification, we propose to exploit the difference in level of detail during the matching to impact the re-ranking. Indeed, with our proposed adaptation, the re-ranking can favor details of the query image or on the contrary, favor images of which the query image is a detail. This can be leveraged to focus image retrieval on a desired representation of a scene based on a query image. The previous score $S_{Q,I}$ is adapted into $S_{Q,I}^d$ such that:

$$S_{Q,I}^d = S_{Q,I} \times w_{Q,I}^d, \quad (4.4)$$

with $w_{Q,I}^d \in [0, 2]$ such as:

$$w_{Q,I}^d = \begin{cases} 1 + \tanh(\alpha\beta(r_{Q,I} - 1)) & \text{if } S(x_{I,J}) < 1 \\ 1 + \tanh(\alpha\beta(1 - \frac{1}{r_{Q,I}})) & \text{otherwise,} \end{cases} \quad (4.5)$$

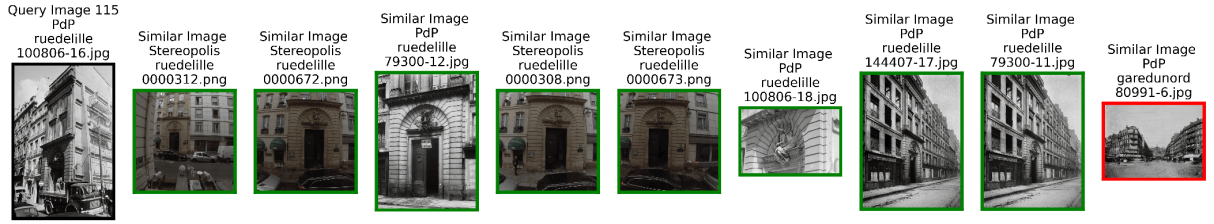
including:

$$r_{Q,I} = \frac{d_Q}{d_I}, \quad (4.6)$$

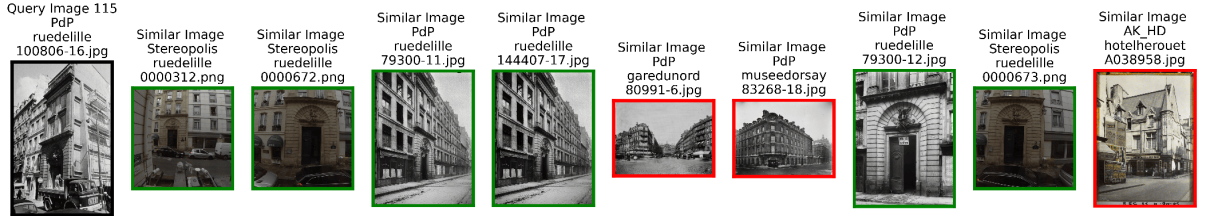
where d_Q and d_I are a measure of the level of detail of the matched keypoints in the query and the retrieved image respectively, α weighs the impact of this level of detail on the score and β indicates whether to favor detailing or englobing images.

First, the measure d of the level of detail of the matched keypoints in an image can be computed in multiple ways. Once the images are all normalized at the same size, it can be the area of the bounding box (or convex hull) of the matched keypoints. It could also for instance be a ratio between this area and the total area of the image. The ratio between the two level of details reflects whether the retrieved image is a detail of the query (ratio less than 1) or a more global view of the scene, englobing the query (ratio more than 1).

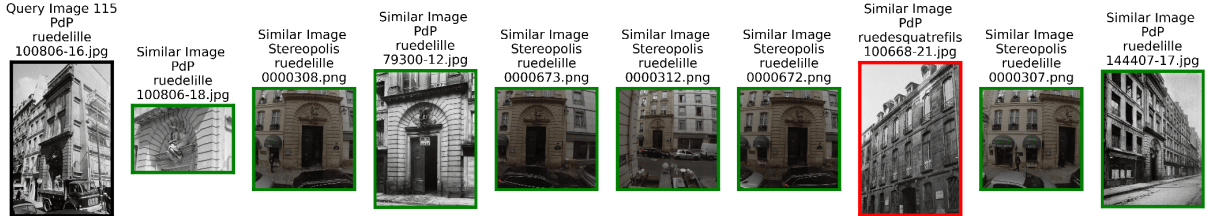
The global coefficient $\alpha \times \beta$ applied to the ratio is composed of two coefficients. First, $\alpha \in [0, \infty]$, that reflects the importance given to the level of detail in the geometric coherence score. If $\alpha = 0$, it means that the level of detail is not taken into account in the re-ranking score, $S_{Q,I}^d = S_{Q,I}$. Secondly, we have $\beta \in \{-1; 1\}$ that binarily indicates if image detailing the query should be favored (value is -1) or if we want to retrieve first images englobing the query (value is 1). Indeed, a positive global coefficient will favor a positive ratio (englobing images) and a negative one will promote negative ratios (detailing images).



(a) Re-ranking without taking into account the level of detail



(b) Re-ranking promoting englobing images



(c) Re-ranking promoting detailing images

Figure 4.5: Visual example of exploiting the level of detail in the R2D re-ranking process

An illustration of the impact on image retrieval is shown with the three retrieval results in Figure 4.5. On each retrieval result, the image framed in black is the query, the images to its right are the retrieved images in descending order of similarity from left to right. If framed in green, the image is correctly retrieved (same class as the query), whereas if it

is framed in red, it is an incorrect result (of a different class). The three examples clearly show the impact of weighting the geometric coherence by the level of detail. Indeed, in every case, the order of the retrieved images is different and the desired images are retrieved first.

4.3 Metadata exploitation

This section introduces another contribution, leveraging this time another type of structural information. Instead of extracting structure at the query level, in the first retrieved images, we aim at leveraging dataset-level structural indications.

4.3.1 Metadata structure weighting scheme

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 when considering multi-source collections, 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.

For geometric verification, a more global and structured representation (a 3D reconstruction) increases the efficiency of the re-ranking process. Building on this idea, we decided to use other structured linking representation of the data and combine it with the automatic visual content-based linking of CBIR. Structured linking can be extracted from the metadata associated with the images. A simple example of this is the position information associated with an image. We make the hypothesis that two images taken at the same location have a much better chance of depicting a similar scene than images distant of several kilometers. Our intuition is to compute a spatial coherence weight based on the metadata and use it to validate/invalidate the visual similarity automatically computed via CBIR. An example of this invalidation can be seen in Figure 4.6. In this illustration which belongs to the 3D representation platform introduced later in Chapter 7, each node represent an image, illustrated by the thumbnail. The blue arrows represent high visual similarity links automatically computed by the retrieval step. On the other hand, the grey links represent the link between an image and a location on the map. Whereas the images are considered highly similar by the automatic retrieval process (which a visual confirmation explains, they all depict the same type of scene coming from mobile mapping), the associated location data shows that they are quite far from each other, tending to disprove their supposed visual similarity.

If the most natural example to understand this concept is the location information, other metadata could be used to weight the visual similarity. For instance, a date information could be used if we assume that images taken closer in time have a bigger chance to be visually similar. Many other structural information could also be used as weight, even at a more global dataset level, in terms of dataset organization or creation, not only with an information purely associated to an image.

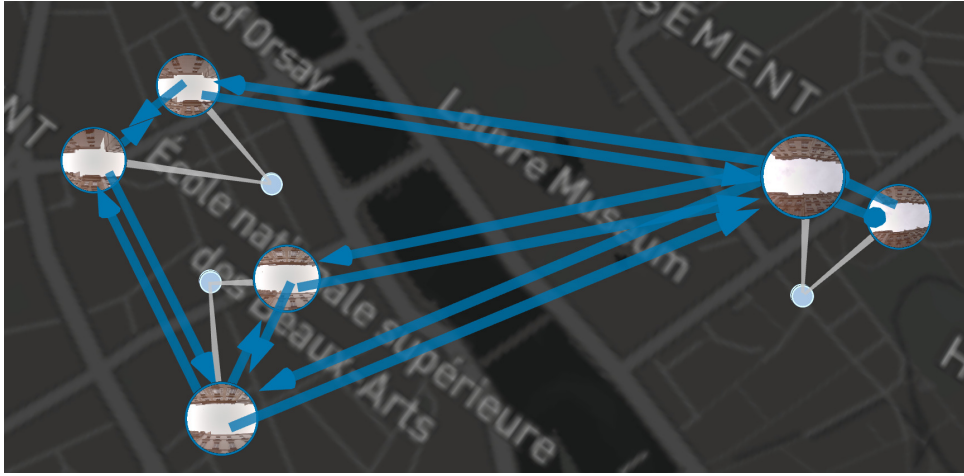


Figure 4.6: Example of contradiction between the metadata information (location) and the automatically estimated visual similarity

A theoretical example of this could be the fact that a photographic acquisition of the whole city is ordered, and a single photographer is assigned to each district in the city. Thus, two images taken by different photographers are less likely to depict the same place than two images taken by the same photographer. Weighting the visual similarity links by this information could lead to a better structuring of the dataset. Figure 4.7 could represent this situation: red and blue links represent more or less strong visual similarities, each node is a photograph and the node's color is based on the photographer that took the picture. Whereas visual similarity links are mainly coherent, some link two nodes of different color. Weighting the visual similarity based on a coherence score extracted from the structured metadata could help remove those potentially wrong links.

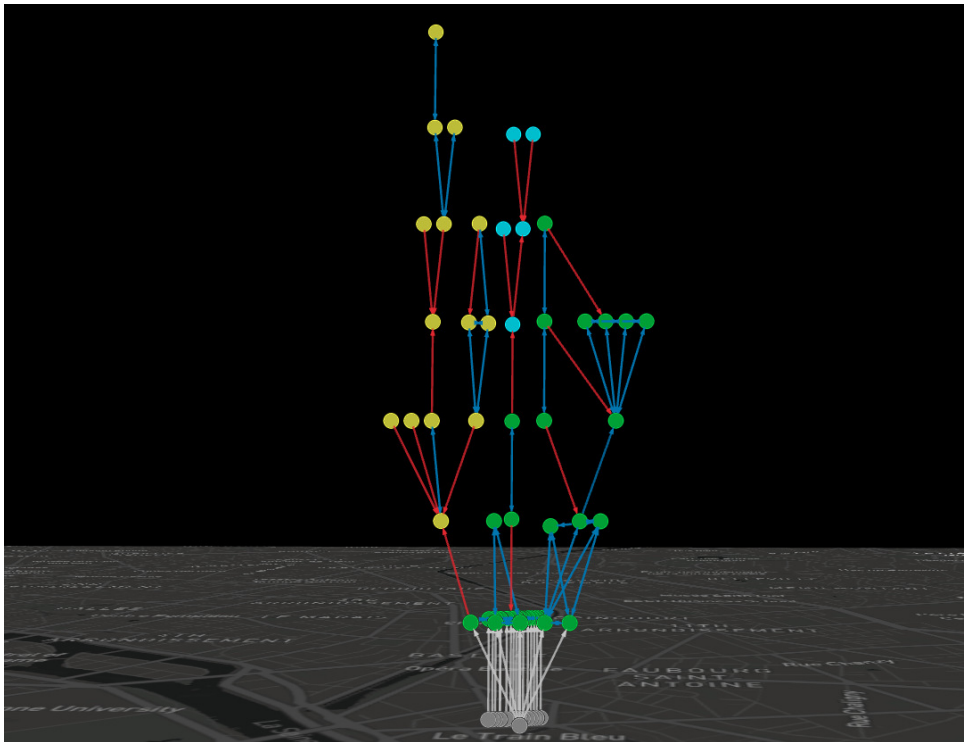


Figure 4.7: Example of structuring based on visual similarities compared to a structuring metadata

4.3.2 Structuring with location information

To evaluate our hypothesis, we focus on the position information available for some of the images in our dataset. This is universal information, but potentially of varying nature: for example, 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 the quality of the location 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). Thus, we give a different weight for the different locations, based on the confidence rate we assign to each of them. Stereopolis images, thanks to their mobile mapping acquisition, have a certain location, giving them a confidence score of 1. For the Parisienne de la Photographie locations, if they are part of the 1637 images selected in classes, their addresses were checked before geocoding, thus giving them a 0.9 confidence rate to account only for automatic geocoding errors. As for the distractor images, their addresses were not manually checked before geocoding, earning them a 0.8 confidence rate to account for both geocoding errors and errors in the addresses directly.

Then, based on the image locations available and in addition to the visual similarity score provided by image retrieval, we define a spatial proximity score $s_{I,J}^s$ between two images I and J .

$$s_{I,J}^s = \sigma(x_{I,J}) \quad (4.7)$$

$\sigma(x_{I,J})$ is a proximity score based on the spatial Euclidean distance $d_{I,J}$ between images I and J : $x_{I,J} = 1 - \frac{d_{I,J}}{d_{max}}$ (normalized over the diameter of Paris in our experiments (d_{max})). We define σ as a double sigmoid function:

$$\begin{aligned} \sigma(x_{I,J}) = & a + (b - a) \times \frac{\tanh(k_1(x_{I,J} - X_1)) + 1}{2} \\ & + (c - b) \times \frac{\tanh(k_2(x_{I,J} - X_2)) + 1}{2} \end{aligned} \quad (4.8)$$

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.

In our case, after analyzing the distribution of distances between all position information in the dataset, we chose to use values $a = 0$, $b = 1$, $c = 2$; $k_1 = 10$ and $k_2 = 50$, as we want to ensure a steep slope (k_2) when images are close in order to validate only very close images while k_1 is lower as the discriminativeness between further images has less impact in the overall process. Finally, $X_1 = 0.878 - 1.5 \times 0.066$ and $X_2 = 0.878 + 1.5 \times 0.066$ with 0.878 being the mean of the distribution of distances and 0.066 the standard deviation. The 1.5 coefficient was determined experimentally to ensure a certain plateau around 1 where distance really does not discriminate one way or another in terms of visual

similarity. Figure 4.8 displays the weighting function defined in Equation 4.7

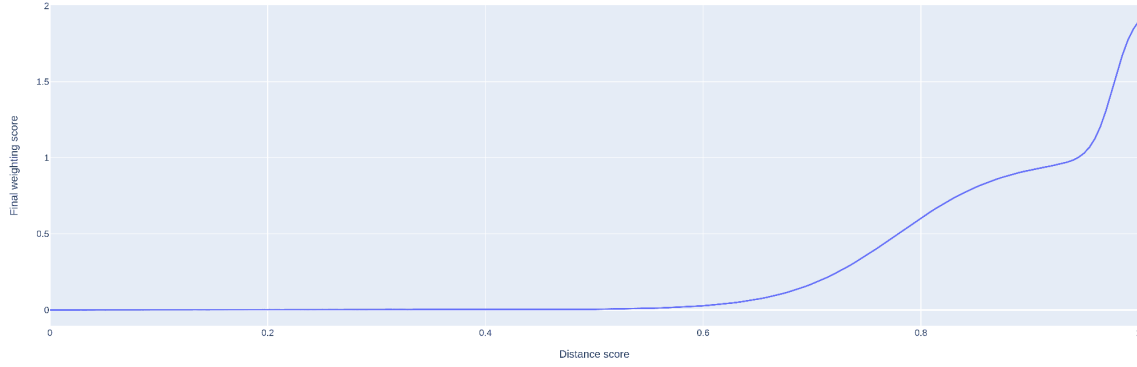


Figure 4.8: Plot of distance weighting function $w_{i,j}$

Furthermore, we weigh $s_{I,J}^s$ with the confidence rate (c_I for I and c_J for J) presented earlier to take the location quality into account, which leads to a spatial weighting score $w_{I,J}^s$ as follows:

$$w_{I,J}^s = s_{I,J}^s c_{I,J}^s \text{ with } \begin{cases} c_{I,J}^s = \frac{1}{c_I \times c_J} & \text{if } s_{I,J}^s < 1 \\ c_{I,J}^s = c_I \times c_J & \text{otherwise} \end{cases} \quad (4.9)$$

To summarize, $w_{I,J}$ ranges in $[0,2]$ and equals 1 if we do not have location information for both images.

We finally combine the proximity and visual similarity scores between couples of images, through a simple weighting of the similarity score with the weight of equation 4.9, with the objective of limiting incoherent retrieval errors due to the limitations of visual descriptors. We have thus the final similarity score between two images I and J :

$$S_{I,J} = s_{I,J}^v \times w_{I,J}^s = s_{I,J}^v \times s_{I,J}^s c_{I,J}^s \quad (4.10)$$

with $s_{I,J}^v$ the visual similarity score.

The whole process applies this weighting scheme and re-rank images by combining two similarities, preventing in no way other steps of re-ranking afterwards. In a way, this can be seen as a sort of geometric verification, using as a scene the whole area of the dataset and as geometric information the 2D location of the images.

This re-ranking proposal is evaluated in Section 5.2.3 of Chapter 5.

4.4 Conclusion

In this chapter, three new re-ranking approaches were proposed in order to alleviate issues our specific dataset raises with classical re-ranking approaches, relying on the geometry of the scene or on geometry between images, and on metadata exploitation. More precisely, with those methods we exploit the structure of the dataset. It can be after a first step

of retrieval by exploiting the global geometric structure of the scene of the query as in Section 4.2. It can be concurrent to the first retrieval step when using dataset-wide structure provided by related metadata to validate or invalidate the visual similarities, as shown in Section 4.3.

The following chapter will focus on testing and evaluating those proposed approaches and further estimate which re-ranking method or combination of methods as hinted before are best suited to obtain the best possible retrieval results.

Chapter 5

Re-ranking Strategies Evaluation

5.1	Introduction	101
5.2	Contributions' evaluation	102
5.2.1	3D geometric query expansion	102
5.2.2	2D geometric query expansion	103
5.2.3	Location proximity weighting	104
5.3	Combination of re-ranking methods	105
5.3.1	Multiple re-ranking combinations' evaluation	105
5.3.2	Insight on provider entropy impact for re-ranking	109
5.4	Runtime of methods and combinations	113
5.5	Key takeaways	114
5.6	Conclusion	115

5.1 Introduction

In Chapter 3, state-of-the-art re-ranking approaches were evaluated against our dataset. Presenting multiple specificities, such content proved challenging for classical retrieval and re-ranking approaches. Faced with those challenges and inspired by diffusion re-ranking approaches exploiting the first retrieval results in a global fashion to extract information from the structure of the retrieved results, we proposed in Chapter 4 new re-ranking strategies using some form of structure. On one side a geometric structure artificially created using a first step of content-based retrieval. On the other side, a global dataset-wide structure exploited jointly with the content-based approach to weigh in on the retrieval results; Section 5.2 will present the evaluation of those approaches individually.

The idea of using a specific structure to perform re-ranking also brought to mind the fact that after every re-ranking step, the structure of the retrieval results is changed. Thus, another re-ranking approach might not yield the same results directly after the first retrieval results or after a first re-ranking step. Added to the fact that some of the proposed strategies are complementary, this led us to evaluate combinations of re-ranking

methods in order to find the most efficient; Section 5.3 will investigate this.

Once again, the effectiveness of the strategies are evaluated based on the mAP score. Furthermore, the computational cost of the re-ranking steps must also be evaluated as it can impact the choice of methods depending on the context in which those methods are used (online, offline, more or less often, etc.); Section 5.4 will discuss this.

5.2 Contributions' evaluation

As previously presented, this section aims at evaluating the approaches proposed in Chapter 4 in the same order they were first introduced. First the 3D reconstruction-based approach in Section 4.2.1, then the method using an approximate 2D reconstruction of the scene in Section 4.2.2 and finally, the weighting scheme based on known image locations in Section 5.2.3. To be noted, for all experiments the number of re-ranked images is always $k = 135$, that is the size of the largest class in our dataset.

5.2.1 3D geometric query expansion

We first evaluate the impact of our proposal exploiting a global 3D scene, presented in Section 4.2.1 of Chapter 4 (named R3D). The evaluation of this method is performed on $DB_{large+dist}$, including distractors. The detection and matching of keypoints for the 3D reconstruction are done with two pairs of detector + matcher, namely SuperPoint + SuperGlue (DeTone et al., 2018; Sarlin et al., 2020) and SuperPoint + LightGlue (DeTone et al., 2018; Lindenberger et al., 2023). To be concise, in all following sections and tables, methods using SuperGlue will end with "-SG" and those using LightGlue will end with "-LG".

Table 5.1, presents the mAP scores for the R3D method compared to those of the simple image retrieval and the classical geometric verification approach. Before analyzing those results, it is interesting to note that out of 9834 potential 3D reconstructions (because of 9834 images queried in the dataset), R3D-SG produces 5111 and R3D-LG 5650. The difference in reconstructions is most certainly due to the higher quality of matches produced by LightGlue, as an improved version of SuperGlue.

Table 5.1: Evaluation of the 3D reconstruction-based approach

Descriptor + Re-ranking step	mAP
How-A	41.0
How-A + RANSAC-SG	41.5
How-A + RANSAC-LG	41.9
How-A + R3D-SG	44.4
How-A + R3D-LG	44.2

The mAP gain using a 3D reconstruction for geometric verification is substantial. More than 3% more than simple retrieval and more than 2% compared to classical geometric

verification. It is a great improvement, tending to prove the validity of our geometric query expansion approach.

An interesting result is the fact that even if R3D-LG manages to reconstruct more 3D scenes, its performance is slightly lower than R3D-SG. Our hypothesis is that with a stricter matching, the 3D scene is also "stricter", meaning that it includes less loosely matched information because it can reconstruct a more certain scene using only stronger matches.

Nonetheless, both experiments clearly outperform classical geometric verification. This costly proof of concept based on 3D information validate our hypothesis that adapting query expansion to geometric information rather than visual description is worth pursuing with an approximate but less costly 2D geometric query expansion.

5.2.2 2D geometric query expansion

This section presents the evaluation of the second proposed method, named R2D, presented in Section 4.2.2 of Chapter 4 as an adaptation of the previous 3D reconstruction-based method.

The evaluation is performed on $DB_{large+dist}$, using the two different keypoint matchers. Table 5.2 presents the mAP scores for both experiments.

Table 5.2: Evaluation of the re-ranking method using a 2D pseudo-reconstruction

Descriptor + Re-ranking step	mAP
How-A	41.0
How-A + RANSAC-SG	41.5
How-A + RANSAC-LG	41.9
How-A + R3D-SG	44.4
How-A + R3D-LG	44.2
How-A + R2D-SG	36.2
How-A + R2D-LG	41.9

Evidently, the results are not as promising as those of R3D. Indeed, when using Super-Glue, the retrieval is impaired compared to simple retrieval, and by a large gap of 4.6%. However, when using LightGlue, the results are similar to those of a classical RANSAC with LightGlue.

Our hypothesis for the different scores between R2D-SG and R2D-LG is actually opposite the one we had to explain the difference between R3D-SG and R3D-LG. In the case of the 3D reconstruction, the stricter matching of LightGlue hindered the expansion capabilities of the 3D scene, thus encoding less geometric information. However, in the 2D approximation, a stricter matching helps improve the approximation of the geometry, thus allowing for a better geometric verification afterwards.

Though disappointing as is, those approaches will have to be reevaluated in light of the combination of re-ranking approaches which will be evaluated in Section 5.3.

5.2.3 Location proximity weighting

Here is the evaluation of the final re-ranking approach that was proposed in Section 5.2.3 of Chapter 4. Relying on location information associated to the images of the dataset, this method was evaluated on $DB_{large+dist}$, however, different paradigms of "metadata availability" were tested.

Indeed, due to their heterogeneity and their high degree of variability between each other, heritage iconographic content collections do not often have corresponding metadata. Thus, as we tested our approach with the location information, we chose to subsample the information available for re-ranking in order to evaluate the impact of data sparsity.

The weighting scheme was tested in three different settings, with the location confidence scores described before in Section 5.2.3 of Chapter 4, as follows:

- using only certain location information provided by the Stereopolis dataset, *i.e.* 537 locations, that is about 5% of the dataset. It is indicated as (Sp) in the results. The confidence score given to those locations is 1, as they are acquired via mobile mapping, thus certain;
- using all locations from Stereopolis and the Parisienne de la Photographie, not considering the distractors, *i.e.* 730 locations, that is about 7% of the dataset. It is indicated as (No dist) in the results. As locations for the Parisienne are obtained from geocoding addresses, their associated confidence score is of 0.9, because the geocoding may be flawed but the addresses were checked beforehand;
- using all possible locations, including those of the distractors, *i.e.* 7896 locations, that is about 80% of the dataset. It is indicated as (All) in the results. Finally, the confidence score associated to the distractors' locations is 0.8 as they also come from automatic geocoding but without checking the address' correctness before.

Table 5.3 presents the results of all three testing settings for the weighting scheme, compared to all previously tested re-ranking methods.

Table 5.3: Evaluation of the location weighting re-ranking approach

Descriptor + Re-ranking step	mAP
How-A	41.0
How-A + RANSAC-SG	41.5
How-A + RANSAC-LG	41.9
How-A + R3D-SG	44.4
How-A + R3D-LG	44.2
How-A + R2D-SG	36.2
How-A + R2D-LG	41.9
How-A + location weighting (Sp)	42.0
How-A + location weighting (No dist)	40.5
How-A + location weighting (All)	42.5

Analyzing those results shows the great promise that such a weighting scheme holds. Indeed, using all possible location data improves retrieval higher than all methods except 3D-based ones. Furthermore, exploiting only the certain location information of the Stereopolis images, while not performing as well as the setting with all locations, still outperforms all non 3D-based methods.

Using all location information from the dataset but that pertaining to the distractor images led to disappointing results, even underperforming compared to simple retrieval. The explanation for this is probably the fact that the less certain location information of the Parisienne de la Photographie brings noise into the "perfect" location information of Stereopolis without bringing the advantage of disambiguating retrieval with the distractors as shown in the (All) setting.

As with all proposed and evaluated methods, this weighting scheme holds promise in terms of re-ranking. However, its full impact will be evaluated in the following section as part of a pipeline of multiple successive re-ranking steps.

5.3 Combination of re-ranking methods

Studying the impact of all previous re-ranking approaches, it has been shown that exploiting as much information as possible is beneficial for image retrieval. Indeed, whether it is dataset-wide structure information, from retrieval (diffusion) or annex data (meta-data weighting), or global geometric information at query level, the more information and structure the better.

However, as soon as a first step of re-ranking is performed, the structure available for other methods of re-ranking is different from that of a simple retrieval step. Thus, specific successive steps of re-ranking may prove beneficial to one another and improve retrieval re-ranking capabilities further than each method independently.

This will be evaluated in this section, and will offer new insights on retrieval and re-ranking for iconographic heritage content.

5.3.1 Multiple re-ranking combinations' evaluation

Going further than a single step of re-ranking, we propose to combine several approaches of re-ranking in a logical manner giving each approach the optimal information for it to perform optimally.

5.3.1.1 A first visual illustration

To illustrate this idea, we visually show the impact of combining re-ranking methods on a simple example in Figure 5.1. On each retrieval result, the image framed in black is the query, the images to its right are the retrieved images in descending order of similarity from left to right. If framed in green, the image is correctly retrieved (same class as the query), whereas if it is framed in red, it is an incorrect result (of a different class).

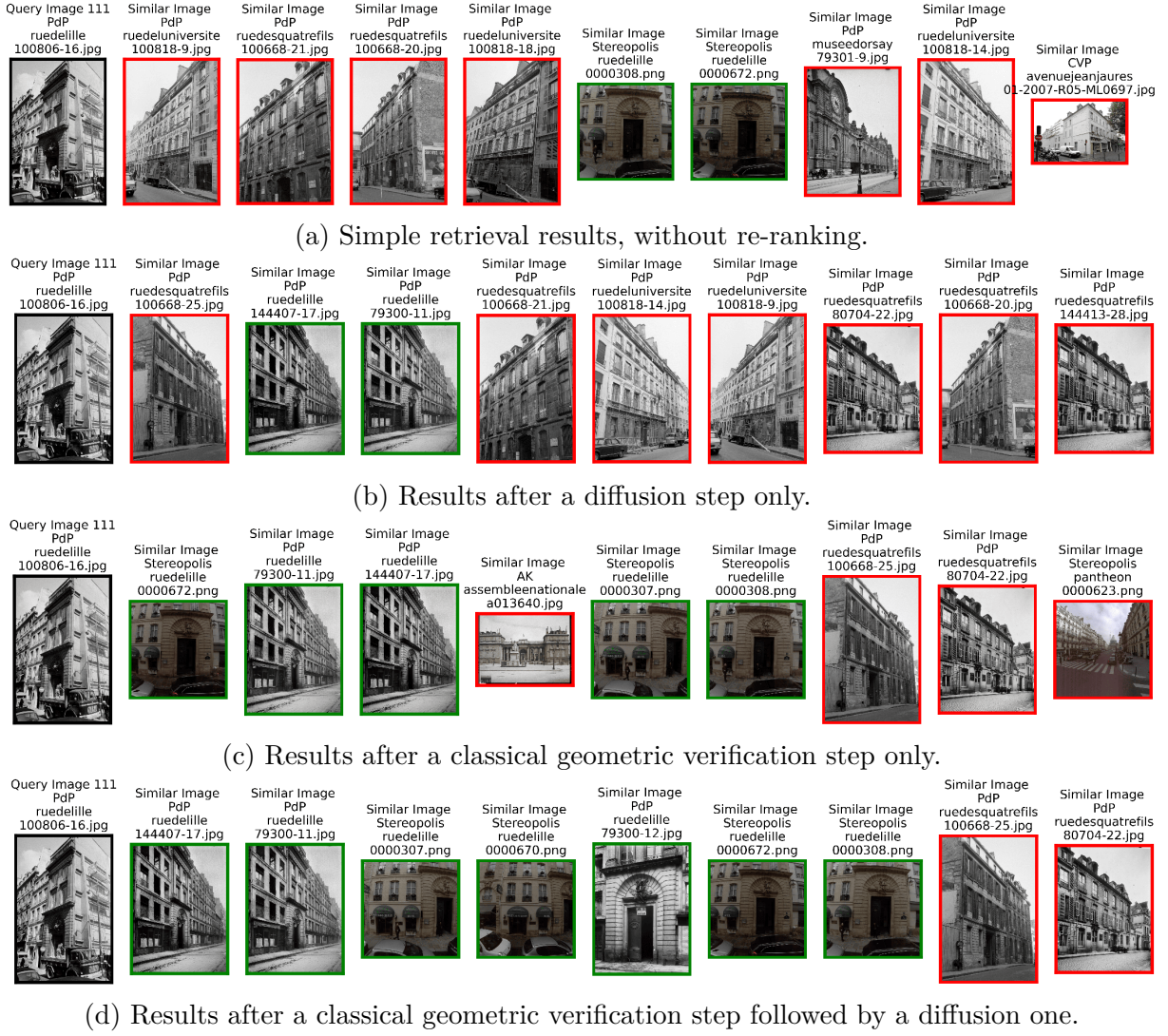


Figure 5.1: Visual example of re-ranking methods combination

This easy example illustrates the reason for combining several strategies of re-ranking. Indeed, applied to the simple retrieval results (Figure 5.1 (a)), the diffusion process (Figure 5.1 (b)) improves slightly the ranking of the first correct image. However, we visually observe that it diffuses information from incorrect results. Thus, the new first correct results now come from the same provider as the query and the initial incorrect results whereas the previously first retrieved images that were from a different provider (Stereopolis) are relegated further in the ranking list. This demonstrates that diffusing incorrect information does not improve retrieval in all aspects of interest to us. Indeed, it improves the ranking of the first image but not the provider interlinking potential of the retrieval.

From another perspective, shown in Figure 5.1 (c), using a classical geometric approach greatly improves the first retrieved results. We can observe that it still preserves the inter-provider aspect that we want, however, an incorrect image still remains highly ranked.

Figure 5.1 (d) finally shows the benefits of successively performing a classical geometric verification step followed by a diffusion one. Indeed, more correct images are highly ranked, and different providers are represented high in the ranking list.

Building on this proof of concept, we evaluate all possible combinations, using a specific

evaluation protocol described in the following section.

5.3.1.2 Evaluation protocol

To evaluate all possible re-ranking methods combinations, we first had to determine the order in which methods could be combined. Indeed, depending on each method, the information provided by another approach may be disregarded, the combination being the equivalent of simply using the second method. Thus, we evaluated each method to divide them into categories based on their input and output to combine them cleverly.

The first category regroups classical geometric verification and our location weighting approach, as approaches for which a previous re-ranking step is useless. Indeed, the first one evaluates the first k images in a pairwise manner, no matter the new order those first images might be given by a first re-ranking step. The second one combines spatial similarity and visual similarity in a new score used for re-ranking. Independently of the order of the first retrieved images, the important information are the two scores, thus once again, no matter the order of the input retrieved images, the output will remain the same.

In contrast to those approaches, we define another category for which the order of the retrieved images is paramount. This category regroups R3D and R2D methods. Indeed, as the reconstruction (3D or 2D) is based on the first ten images in our experiments, the results after a first step of re-ranking may differ from those after a simple retrieval. This category of approaches can potentially build on and exploit the improvements brought by another approach.

Finally, we let the diffusion-based re-ranking approach in another category. Indeed, with this method both the similarity scores (visual or visual weighted by location) and the order of the images in the ranking list are important. Diffusion can thus be applied after no re-ranking or multiple steps of re-ranking. Diffusion can also be applied after a first diffusion step, further improving the retrieval as will be shown in the next section. Furthermore, as shown previously, diffusing errors is not beneficial to retrieval, thus diffusion should be the last step of re-ranking.

Figure 5.2 summarizes all categories and all possible combination of methods.

5.3.1.3 Results presentation and analysis

This section presents results for most combinations of re-ranking methods. Table 5.4 summarizes mAP scores for all combination. All previous naming conventions are kept, and one is reminded: GNN-R stands for GNN-Reranking, the diffusion approach from (Zhang et al., 2020b), evaluated in Section 3.5.6 of Chapter 3.

The scores of Table 5.4 confirm our intuition that combining re-ranking methods yields better results than any one of them independently. Indeed, using a single re-ranking method yields at best a 57.2% mAP when using only diffusion, with the potential drawbacks for inter-provider retrieval already mentioned. Combining RANSAC-SG and R3D-SG, followed by two diffusion steps brings the mAP up to 65.8% which is a great

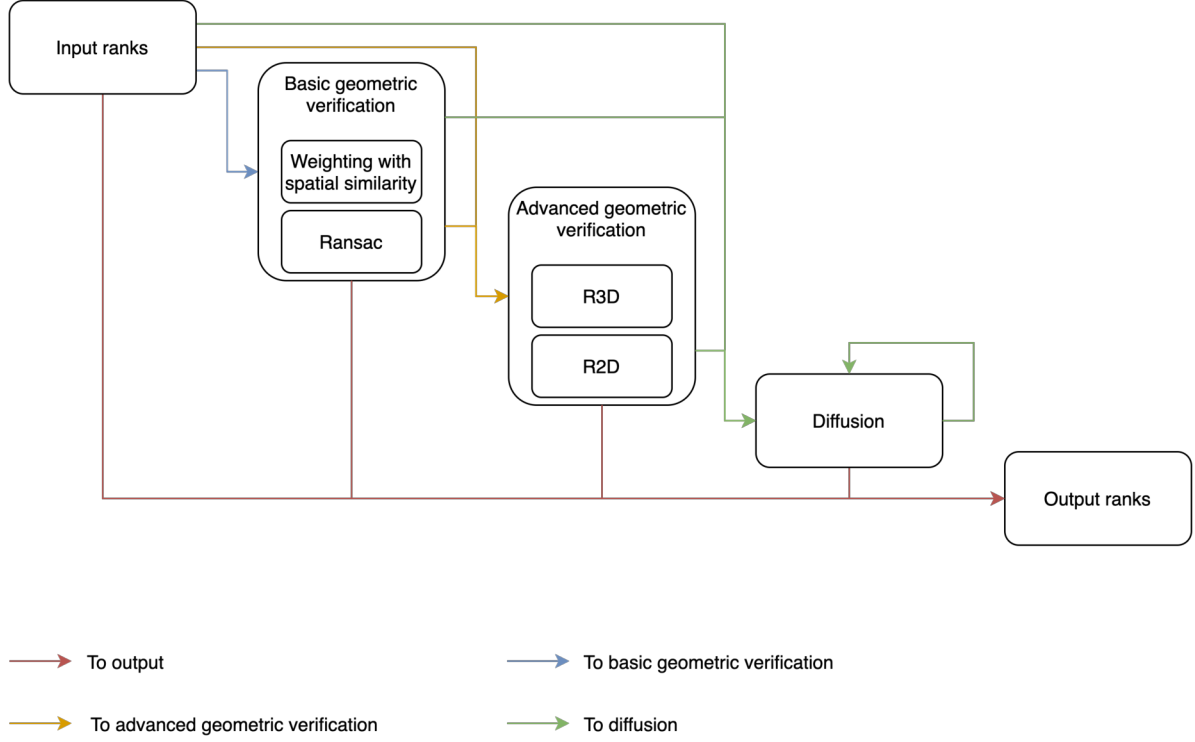


Figure 5.2: Workflow of re-ranking steps combinations

Table 5.4: mAP scores for multiple combinations of re-ranking steps. Indicated in color are the **first**, **second** and **third** best results for each column, and in **bold** the best score overall.

Descriptor + Re-ranking step	Diffusion after previous re-ranking			
	No GNN-R	GNN-R \times 1	GNN-R \times 2	GNN-R \times 3
How-A	41.0	57.2	59.3	57.0
How-A + RANSAC-SG	41.5	57.2	59.3	57.0
How-A + RANSAC-LG	41.9	61.2	65.5	63.3
How-A + R3D-SG	44.4	61.9	64.2	61.9
How-A + R3D-LG	43.2	61.1	63.2	60.7
How-A + R2D-SG	36.2	59.6	62.9	60.5
How-A + R2D-LG	41.9	61.0	64.4	62.1
How-A + location weighting (Sp)	42.0	58.9	61.8	59.5
How-A + location weighting (No dist)	40.5	57.8	61.1	59.0
How-A + location weighting (All)	42.5	60.2	63.1	61.8
How-A + RANSAC-SG + R3D-SG	44.9	62.9	65.8	63.3
How-A + RANSAC-LG + R3D-LG	43.0	61.8	64.1	61.9
How-A + RANSAC-SG + R2D-SG	36.9	60.1	63.0	60.5
How-A + RANSAC-LG + R2D-LG	41.7	61.2	64.3	62.2
How-A + location weighting (Sp) + R3D-SG	44.7	62.4	64.9	62.4
How-A + location weighting (Sp) + R2D-LG	41.9	61.1	64.7	62.1

improvement of more than 8% against only diffusion. Furthermore, those results allow us to draw other conclusions.

A first common point to every combination is that performing diffusion re-ranking in succession is relevant up to two times. The third diffusion step depreciates the results in every case.

A second lesson is that conclusions on our proposed geometric query expansion approaches in Section 5.2.1 and 5.2.2 still remain.

First, in terms of reconstructions, RANSAC-SG + R3D-SG manages to reconstruct 5479 3D scenes, a bit more than without RANSAC-SG whereas RANSAC-LG + R3D-LG reconstructs 7582 scenes. Despite a higher number of reconstructions, its mAP score is not better. This reinforces our belief that a matching too strict is detrimental for the required geometric "expansion". However, it must be noted that using a RANSAC before the reconstruction allows for a better score overall, no doubt due to more meaningful reconstructions using a better set of input images.

Second, in the case of R2D, using a stricter matching method proves once more that it helps render a better approximation of the scene. Furthermore, as with R3D, using a RANSAC before helps the method to perform better when combined with diffusion. Indeed, a very important thing to notice is that even if R2D approaches perform less than R3D ones and even RANSAC-LG, when combined with diffusion they perform adequately, with scores similar or only slightly below R3D based methods of which they are an approximation. Finally, they reach third best score overall when combined with location weighting.

A final teaching of these results is the huge impact that location weighting has when combined with other methods. Indeed, we chose the most probable case, using only information from a certain source, the Stereopolis locations and yet, all combinations with location weighting perform better than all combinations with RANSAC but one. When thinking in terms of computational cost, it is very interesting and will be discussed in Section 5.4.

To conclude on those results, combining re-ranking steps is a sure way to further improve retrieval results. However, when comparing the first three methods in each column, one can see that the final score is not only dependent on the initial mAP score before diffusion. That is particularly clear with RANSAC-LG which is the 8th best method before diffusion and reaches the 2nd place overall after diffusion. Hence, the diffusion seems to extract and diffuse some information from the global dataset-wide ranking list that the mAP score alone can not represent. This will be investigated in the next section.

5.3.2 Insight on provider entropy impact for re-ranking

As previously shown with the evaluation of combinations of re-ranking methods, the mAP score at the end of a first step of retrieval or re-ranking does not predict the results obtained after diffusion. We investigate in this section a reason explaining why diffusion is able to extract more information to diffuse through the dataset from ranking lists of

lower quality in terms of mAP and thus managing a final results higher than other list of higher mAP score.

5.3.2.1 Intuition on provider entropy

First, we remind the readers that the diffusion process, exploiting both the similarity scores and the ranking of the most similar images, extracts information from the graph of nearest-neighbors of each query and diffuse widely the retrieval context to the whole dataset.

Second, in our context of multi-provider dataset, it must be remembered that image descriptors are most performant in an intra-provider setting, meaning that the retrieval process will most likely retrieve images from the same provider as the query's. Images from the same provider will then probably be ranked higher than images from other providers.

Based on both those insights, our intuition is that in order to diffuse the most information and rank highly as many correct images as possible (and increase the mAP), the diffusion must be able to increase inter-provider content retrieval. To do this, it must have access to a maximum of providers in the graph of nearest-neighbors. Indeed, as explained before, providers retrieve themselves more easily. Thus, if a query from one provider retrieves other providers then the diffusion will be able to extract similarities from those providers' contents and then increase the presence of those contents in the final ranking list.

To conclude, aside from the initial retrieval performance, the distribution of providers in the ranking list is also paramount. We evaluate that with the entropy of provider's distribution and observe its impact on the diffusion's performance.

5.3.2.2 Artificial entropy experiment

To validate our intuition, we performed an artificial provider entropy modification experiment in order to see if this is indeed an impacting factor in diffusion performance.

To do this, we modify the ranking list in order to maximize the entropy of providers amongst correct images while conserving the same exact mAP score. We reorder the first twenty elements of the ranking list following those conditions for each query's ranking list:

- get the ranks of correct images,
- reorder the correct images in these ranks by alternating providers in order to maximize the entropy,
- the images are repositioned with regard to their original ranking order.

Figure 5.3 presents an example of entropy maximization. Here we maximize the entropy of the five first results. In black is the query, in green correct images and in red

incorrect images. The retrieved images are displayed from left to right in order of decreasing similarity. Inside each retrieved image, the initial ranking number and in correct ones, the provider is also added. We assume that between the 7th and the 12th image there are no other correct images and that there are only three possible providers for the query image.

After entropy maximization for the first five results, the mAP score is unchanged whereas all providers are now ranked high in the retrieval list.

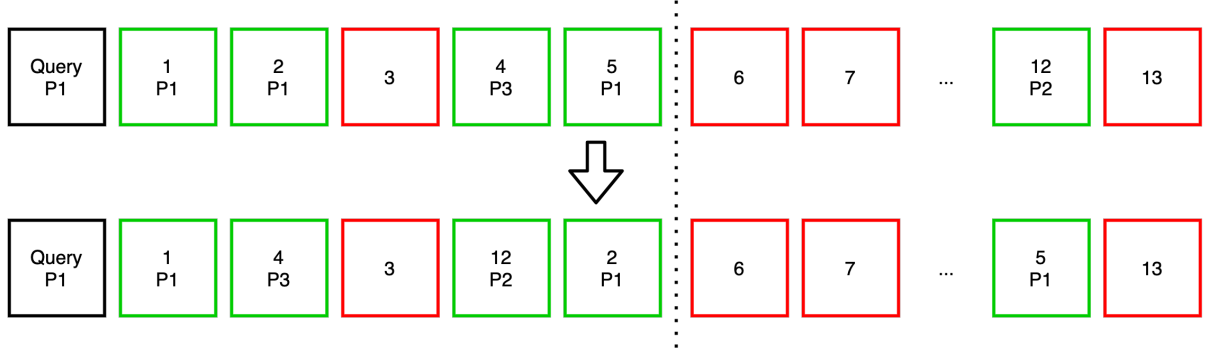


Figure 5.3: Example of artificial increase of provider entropy while keeping the same mAP

We maximize entropy amongst the twenty first images and test entropy maximization on some of the descriptors and re-ranking combinations; the results are summarized in Table 5.5 (lines with "Max entropy").

Table 5.5: Evaluation of the impact of diffusion depending on provider's entropy

Descriptor + Re-ranking step	Entropy @20	Diffusion after previous re-ranking			
		No GNN-R	GNN-R \times 1	GNN-R \times 2	GNN-R \times 3
How-A	36.4	41.0	57.2	59.3	57.0
How-A (Max entropy)	61.4	41.0	66.8	69.9	67.3
How-A + RANSAC-SG	38.0	41.5	57.2	59.3	57.0
How-A + RANSAC-SG (Max entropy)	61.5	41.5	67.2	70.1	67.7
How-A + R3D-SG	41.5	44.4	61.9	64.2	61.9
How-A + R3D-SG (Max entropy)	61.6	44.4	71.4	73.7	70.4
How-A + R2D-SG	42.4	36.4	60.1	63.0	60.5
How-A + R2D-SG (Max entropy)	57.7	36.4	69.3	72.1	69.3

In all the experiments performed, we observe that increasing the entropy artificially, while keeping the mAP score identical, improves the diffusion efficiency of almost 10%. This experiment demonstrates the high impact on the diffusion process of the visual entropy coming from different providers. This observation clearly demonstrates the importance and interest of implementing a retrieval technique based on descriptors capable of being robust to visual variation between providers' contents.

5.3.2.3 A combination of mAP and entropy for diffusion

We further analyze the re-ranking combinations obtained with this provider entropy aspect in mind. The objective is to validate the fact that an ideal first re-ranking step

must combine a good retrieval performance evaluated by the mAP score and ensure a good distribution of every retrievable provider in the ranking list used afterwards by the diffusion step.

We illustrate this idea with several examples that must be analyzed alongside Table 5.4. For each example, the entropy is computed for the first n images in the ranking list.

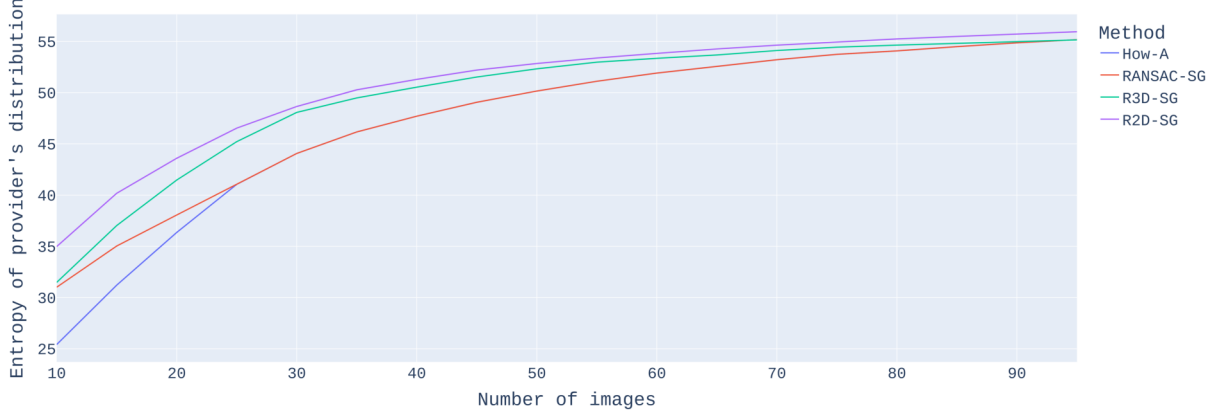


Figure 5.4: Comparison of provider entropy between How-A, How-A + RANSAC-SG, How-A + R3D-SG and How-A + R2D-SG

This Figure 5.4 compares the entropy of a simple retrieval (How-A) and three approaches to re-ranking before diffusion (How-A + RANSAC-SG, How-A + R3D-SG and How-A + R2D-SG). All re-ranking approaches use the same matching method.

First, we saw before in Table 5.4 that despite a small improvement at first, combined with diffusion, RANSAC-SG reaches the same results as the simple retrieval. In terms of provider entropy, while starting higher, RANSAC-SG’s entropy becomes similar than that of simple How-A, explaining the similar performance with diffusion.

Second, while R2D-SG started with a worst mAP than simple retrieval, diffusion brought it almost up to the level of R3D-SG which started with a mAP much higher than simple retrieval. When studying their respective entropy, we observe that R2D-SG’s entropy is on par with that of R3D-SG.

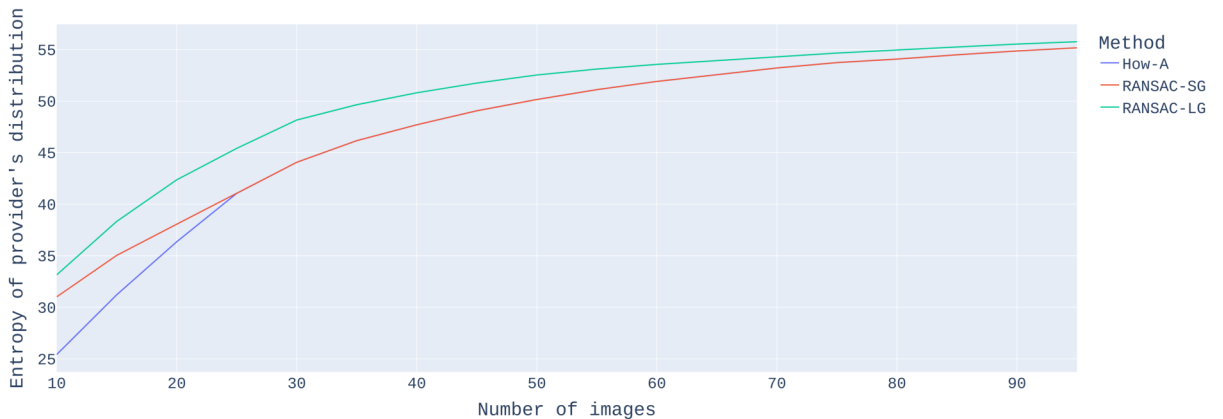


Figure 5.5: Comparison of provider entropy between How-A, How-A + RANSAC-SG and How-A + RANSAC-LG

Figure 5.5 compares the entropy of simple retrieval and both RANSAC methods, meaning using two different matching methods. While the mAP gain of RANSAC-LG compared to RANSAC-SG was only of 0.4%, after diffusion the difference is of 6%. Comparing their respective entropies shows that while RANSAC-SG does not improve its provider entropy compared to simple retrieval, RANSAC-LG greatly improves it, explaining that it reacts much better to diffusion.

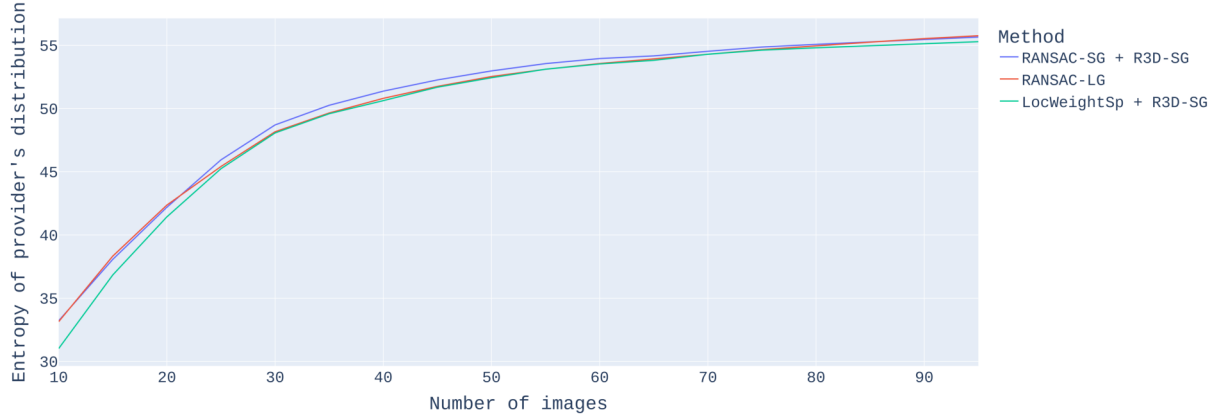


Figure 5.6: Comparison of provider entropy between How-A + RANSAC-LG, How-A + RANSAC-SG + R3D-SG and How-A + Location weighting (Sp) + R3D-SG

Finally, Figure 5.6 compares the entropy scores of the best three methods overall. While before diffusion about 4% of mAP separate the first and the third method (with the final order respected), less than 1% separate them after diffusion. Furthermore, comparing the entropies show that they are very similar, with the final order of performance once more respected.

To conclude, these examples show that more than focusing on the best re-ranking step in terms of mAP, for inter-provider retrieval in a challenging dataset like ours, we have to ensure the maximum possible entropy in the different providers' distribution amongst the first retrieved images. More than pure mAP efficiency, a good combination of mAP and entropy is essential.

5.4 Runtime of methods and combinations

In order to complete the evaluation of the strategies of retrieval with re-ranking proposed, this section provides an analysis of computation time for each solution, to be put in perspective against the mAP gain they allow for. Table 5.6 summarizes all mean computation time for the re-ranking of the first $k = 135$ images of one query image.

Those computation times indicate that, logically, the more re-ranking steps, the more computation time required. However, three main aspects should be noted.

First, RANSAC-LG appears to finally be the method that, combined with diffusion, is the best trade-off between mAP performance and computation time. Indeed, for less than a third of computation time, it performs only 0.6 % less well than RANSAC-SG + R3D-SG.

Table 5.6: Mean computation time for each re-ranking strategy, including combinations, for $k = 135$ images

Descriptor + Re-ranking step	mAP after GNN-R $\times 2$	Mean computation time
How-A	59.3	
How-A + RANSAC-SG	59.3	+120s
How-A + RANSAC-LG	65.5	+100s
How-A + R3D-SG	64.2	+220s
How-A + R3D-LG	63.2	+210s
How-A + R2D-SG	62.2	+150s
How-A + R2D-LG	62.8	+140s
How-A + location weighting (Sp)	61.8	+1/30s
How-A + location weighting (No dist)	61.1	+1/30s
How-A + location weighting (All)	63.1	+1/30s
How-A + RANSAC-SG + R3D-SG	65.8	+340s
How-A + RANSAC-LG + R3D-LG	64.1	+300s
How-A + RANSAC-SG + R2D-SG	63.0	+270s
How-A + RANSAC-LG + R2D-LG	64.2	+240s
How-A + location weighting (Sp) + R3D-SG	64.9	+220s
How-A + location weighting (Sp) + R2D-LG	63.5	+140s

Second, location weighting is a very quick way to improve the final mAP if computation efficiency is paramount.

Finally, for the diffusion process, GNN-R is very quick (less than 2s) but is performed at dataset-level, requiring for all images in a dataset to have been re-ranked first in order to exploit the full potential of the diffusion. This can be very costly and is not be ideal for online performance. However, when done in a production setting, once and for all, it is definitely worth exploiting the best combination of re-ranking methods for optimal performance.

5.5 Key takeaways

We will here succinctly summarize all key takeaways that the experiments of this chapter have brought to light.

First, exploiting more structural information is always beneficial when combined with a diffusion process. Indeed, the more information diffused, the better the final performance. Thus, on one side the geometric query expansion propositions or the location weighting scheme have proven to improve re-ranking substantially. On the other side, exploiting dataset-level information for weighting visual similarity brings structure and improves overall retrieval while being very quick.

Second, combining multiple re-ranking approaches is beneficial for overall performance. However, with costly methods on top of costly methods, a trade-off has to be performed depending on the desired frequency of the retrieval.

Finally and more importantly, rather than trying to achieve the best mAP performance before diffusion, it is essential to ensure a good distribution of all providers of images in the first retrieved results. Indeed, in the setting of collections interlinking with a

low inter-provider retrieval performance, a high entropy in the provider’s distribution at the beginning of the ranking list is paramount for the diffusion process to achieve top performance.

5.6 Conclusion

In this chapter, our proposed methods for re-ranking were first evaluated, proving interesting for improving retrieval. However, their potential was really demonstrated when combined with other re-ranking steps, especially the diffusion process. Indeed, combining multiple steps of re-ranking brings the initial retrieval results from a mAP score of 41.0 up to 65.8, a very high performance boost.

Furthermore, further analysis of the various combinations’ performances led us to confirm that performance of re-ranking steps prior to diffusion should be evaluated as a combination of two elements. First, the classical mAP score that evaluates the quality of the retrieval. Second, as we perform our study in the setting of collection interlinking, the distribution of the various retrievable providers amongst the first retrieved results is very important. Indeed, that is what allows the diffusion process to perform at its best and alleviate the main issue with retrieval amongst various providers, a poor inter-provider retrieval performance.

However, as we evaluated the computational cost of those re-ranking steps, it appeared clear that at some point, despite all the re-ranking possible, we will hit the maximum of what automatic methods can do in the very particular setting of image retrieval for iconographic heritage content. Indeed, the specificity of the data, its high visual variability and its natural unbalanced sparsity prevent from achieving perfect retrieval.

With this conclusion in mind, alongside our re-ranking setting exploiting structure for retrieval improvements and the specific paradigm of diffusion that propagates structure throughout the dataset, we thought about how to add more certain structure for diffusion purposes. For that purpose, the following Part II will focus on a graph-based semi-automatic retrieval paradigm that we propose. The main idea is to use a graph-based visualization platform for an expert to evaluate the first automatic step of retrieval. With its interventions, the expert adds structure that can then be diffused through the dataset, in order for a small intervention to have a multiplied effect.

Part II

Graph-based Semi-automatic Structuring

Chapter 6

Structuring, Spatializing and Visualizing Iconographic Heritage

6.1	Introduction	119
6.2	Image spatialization	120
6.2.1	Spatialization paradigms overview	120
6.2.2	Manual spatialization	124
6.2.3	Semi-automatic spatialization	125
6.2.4	Automatic spatialization	128
6.3	Iconographic heritage structuring and visualization	131
6.3.1	Single modality platforms	131
6.3.2	Multiple modalities combinations platforms	134
6.4	Conclusion	140

6.1 Introduction

As structuring the increasing digitized contents becomes paramount due to the potential applications becoming more and more numerous, several paradigms for organizing and visualizing those contents are developed and exploited. The choices in structuring leading to limitations in terms of usage and visualization, the various paradigms must be correctly identified.

This chapter aims at presenting the various paradigms of structuring and visualization available through different platforms and serving different goals. Furthermore, we also detail how to spatialize contents in more or less automatic ways; this is an essential step for visualizing content in a spatialized environment, and that can be complex to implement. Indeed, using a spatial structuring is a very efficient way to display and analyze visual contents, thus warranting a further study in this chapter.

To summarize, Section 6.2 presents spatialization techniques depending on their level of automation while Section 6.3 provides an overview of existing platforms for structuring and visualizing visual contents based on their structuring paradigms.

6.2 Image spatialization

Organizing and displaying contents in any given space, whether it is geographic, based on descriptors, or completely arbitrary, brings out endogenously the global structure of any dataset (no matter the data). No matter the use one may have for it, analyzing the structure of any dataset helps understand it. Visualizing contents in their context only enriches their potential in terms of comprehensive analysis and their overall value. The added value of spatialization of contents for analysis purposes has been made apparent in multiple fields and more specifically in social sciences and humanities (SSH). Starting in 1989 with the geographer Edward Soja, the term "spatial turn" describes the paradigm shift in SSH studies that is still ongoing today (Podpora, 2011; Dörfler and Rothfuss, 2023; Bartmanski et al., 2023). That led to the spatialization of multiple contents for analysis and visualization purposes, as the spatial aspect took a larger place in analyzing social processes and organizations.

In terms of iconographic heritage, displaying scenes from the past, within a global context, both spatial and temporal, is paramount, particularly in social sciences and humanities studies, as shown in (Blettery et al., 2020). When it comes to images representing past or still existing scenes, structuring them and displaying them allows for analysis of their evolution. This gives depth to both the current scene and the past representations. The most adapted representation space to do this is the geographical space, either in 2D or in 3D. However, heritage contents not always come with location information. In the case of heritage contents acquired digitally, location metadata are more often acquired at the acquisition time and stored jointly with the images (*e.g.* EXIF metadata with a GPS location). However, digitized contents are rarely digitized with integrated metadata, and the location information can be lost during this digitization process if it were even noted at acquisition time. Furthermore, when digitized heritage contents are provided with a location information, it is mostly as textual metadata representing an address, in a more or less usable formatting. Thus, to exploit those contents in the geographical space, one must spatialize them first.

This section introduces various paradigms and methods for spatializing image contents requiring more or less manual expert intervention and leading to more or less precise location information. We first present all spatialization paradigms in Section 6.2.1, in terms of levels of spatialization and reference frames available. We then focus successively first on manual spatialization methods in Section 6.2.2, then on semi-automatic approaches in Section 6.2.3 and finally Section 6.2.4 develops the full automatic spatializing paradigm.

6.2.1 Spatialization paradigms overview

This section is a preamble to the presentation of the spatialization techniques. It has two objectives: first, clarifying what kind of spatial information can be associated with an image, and second, revisiting the data sources available to assist the spatialization process.

6.2.1.1 Levels of spatialization

Depending on the method and application, the spatialization of an image may refer to finding an information of geolocalization either of the content imaged or of the sensor at the origin of the image. More precisely, this information can be:

- A **textual annotation**, providing an information of geolocalization with toponyms: department, city, locality, name of a monument, etc. It is often the case with collections from preservation institutions which are documented with standardized descriptive metadata (standardized with vocabularies and reference databases (*e.g.* CIDOC-CRM)), or AI learning algorithms dealing with the "place recognition" problem which provides a semantic label.
- A **2D or 3D position**, which can be relative (determined in a particular coordinate system, *e.g.* a map, a 3D model) or absolute (on Earth, associated with a standardized reference system, *e.g.* WGS84). Such information on images is natively provided by national mapping agencies, based on regular surveys, as well as by recent cameras equipped with GPS. It can also be provided with geocoding techniques that consist in assigning geographic coordinates to a toponym by using reference datasets (*e.g.* GeoNames geographical database) and API (*e.g.* Google's Geocoding API or OSM's API Nominatim), and also estimated by vision-based computational tools.
- A **6-DoF pose** (*i.e.* the position and orientation of the camera with 6 Degrees of Freedom), either available with professional systems (national mapping agencies, mobile mapping systems) or estimated with computational tools dedicated to vision-based localization. Depending on the specifications of the application, the algorithmic choices can go as far as the calibration of the acquisition system (*e.g.* the estimation of the focal length for rectification or dedicated visualization of the spatialized content).

We will afterwards mainly refer to a 2D or 3D position as a location and a 6-DoF pose as a pose.

6.2.1.2 Available data as spatialized reference

Whatever the approach employed, spatializing an input image supposes that the involved space is already known, in other words that we have at our disposal a spatialized representation of this area on which we can rely on to infer the localization of the image input. This reference can take various forms, from simple descriptive metadata or labels up to a 3D model of the scene, through different kinds of maps and image datasets, which we briefly revisit here.

Spatialized image datasets. A first way of spatializing non-located contents is to exploit the spatialization of similar contents either as a reference (a map for instance) or as a starting location to be refined afterwards (from similar images that are spatialized).

In the Computer Vision and Machine Learning communities, there exist several annotated datasets dedicated to landmarks, which can apply for spatialization; let's mention specifically Google Landmarks (Weyand et al., 2020), which is one of the best known (version GLDv2 is the largest with over 5M images and 200k distinct spatial instance labels). It is also relevant as training and test dataset for the CBIR as presented in Chapter 3 and place recognition tasks, which can be used as a first step of localization (seen later in Section 6.2.4).

For more specific or dedicated purposes and contents (neither mapped nor perennial landmarks), annotated training datasets and benchmarks are usually not available. However, the CBIR task is able to exploit spatialized image collections that may exist in GLAMs (Galleries, Libraries, Archives and Museums) which cover various iconographic contents, or in public and private mapping agencies which image territories at large scale. Here, the metadata associated with iconographic heritage are very heterogeneous, depending on the objectives and standards of the holding organizations. For instance, semantic descriptions of the content for preservation institutions and multimodal geographic descriptions for mapping agencies. Using this location information (whichever the form it comes in) can provide a starting location for similar non-spatialized contents.

Note that maps are by definition a rich source of referencing, with an unequaled spatial (and sometimes temporal) coverage, but if there exist some automatic solutions to align a vertical airborne view with a map (Krüger, 2001; Khokhlova et al., 2021), it is more difficult to establish a link between a map and a free viewpoint image. One alternative is to rely on semantic landmark extraction (through pattern detection tasks) and to search on maps by exploiting spatial reasoning, such as in (Weng et al., 2020) where the semantic objects seen in the street-view image serve as anchors for spatialization within OpenStreetMap.

Several public image datasets dedicated to landmarks are presented in Table 6.1, with a focus on the spatial coverage addressed, the type of localization data available as well as the time period covered to have an insight on their match with old contents. Whether the objective is to employ them to learn a description or as reference to spatialize a content, these datasets are numerous, but most of them are not dedicated to heritage iconographic heritage. They do not reflect correctly the heterogeneity representative of heritage contents as experimented with the ALEGORIA dataset (Gominski et al., 2021), in which are highlighted the difficulties encountered by state-of-the-art deep features in the context of cultural heritage content retrieval.

3D models. To gain in robustness and precision when the objective is to estimate a 3D position or a 6-DoF pose, the most recent and efficient spatialization approaches exploit all the geometrical 3D information available. Many approaches exploit 3D point clouds

Table 6.1: Overview on public image datasets dedicated to landmarks, exploited for training purposes or as spatialization reference (MMS stands for Mobile Mapping System)

Dataset	Number of images	Viewpoint and spatial coverage	Localization type	Time gap
Large Time Lags Locations (Fernando et al., 2015)	500	Street-level 25 cities of Europe and Asia	Label	150 years
Google Landmarks Dataset v2 (Weyand et al., 2020)	Over 5M	Street-level and aerial 246 countries	Label	Unspecified
\mathcal{R} Oxford (Radenovic et al., 2018)	Over 5k	Mostly street-level and some aerial Oxford	Label	Unspecified
Aachen Day-Night (Sattler et al., 2018)	7712	Street-Level city of Aachen (Germany)	Label, GPS, 3D	2 years
Extended CMU-Seasons (Sattler et al., 2018)	Over 110k	Street-level MMS camera areas of Pittsburgh (USA)	Label, GPS, 3D	1 year
RobotCar Seasons (Sattler et al., 2018; Maddern et al., 2017)	Over 35k	Street-level MMS camera city of Oxford (UK)	Label, GPS, 3D	1 year
Kitti Vision Benchmark (Geiger et al., 2012)	389	Street-level MMS camera Greater Karlsruhe (city, rural areas and highways)	Label, GPS, 3D	2012
SILDa Weather and Time of Day (Balntas, 2019)	Over 14k	Street-level and aerial London	Label	1 year
HistAerial (Ratajczak et al., 2019)	4.9M	Vertical aerial France (sparse)	GPS	1970-1990
ALEGORIA (Gominski et al., 2021)	13175	Street-level and aerial France (sparse)	Label	1920's-today

obtained with Structure-from-Motion (SfM) techniques, as in (Yang et al., 2019), (Pion et al., 2020) and (Sarlin et al., 2021a), built especially on a given area for spatialization in this area. There exist other alternatives, such as simple or sophisticated 3D building models, as well as LiDAR or RGB-D data belonging from recent scanning systems that respectively provide a 3D sparse geometrical information and a 3D depth.

For 20 years, with the purpose of autonomous driving, the Robotics community has provided a large variety of vision-based public benchmarks, involving image datasets spatialized with a very rich information (GPS, LiDAR, RGB-D, 3D models, etc.). Their benchmarks are far from the iconographic heritage spatialization problem, but it is interesting to point out that recently they have been enriched with multi-date data to tackle the problem of long-term mapping, which to some extent bring them closer to the variety found in heritage contents.

Interestingly, thanks to public or private mapping agencies, some 3D models exist at large scale and are then usable on a much less narrowed footprint than dedicated SfM clouds or public benchmarks, such as those displayed in Figure 6.1. Currently, their spatial coverage tends to be inversely proportional to their precision (in terms of levels of detail and localization), but this ratio is reducing with the implementation of massive and sophisticated acquisition protocols (*e.g.* aerial HD LiDAR will be made available by IGN on the whole French territory in 2025).

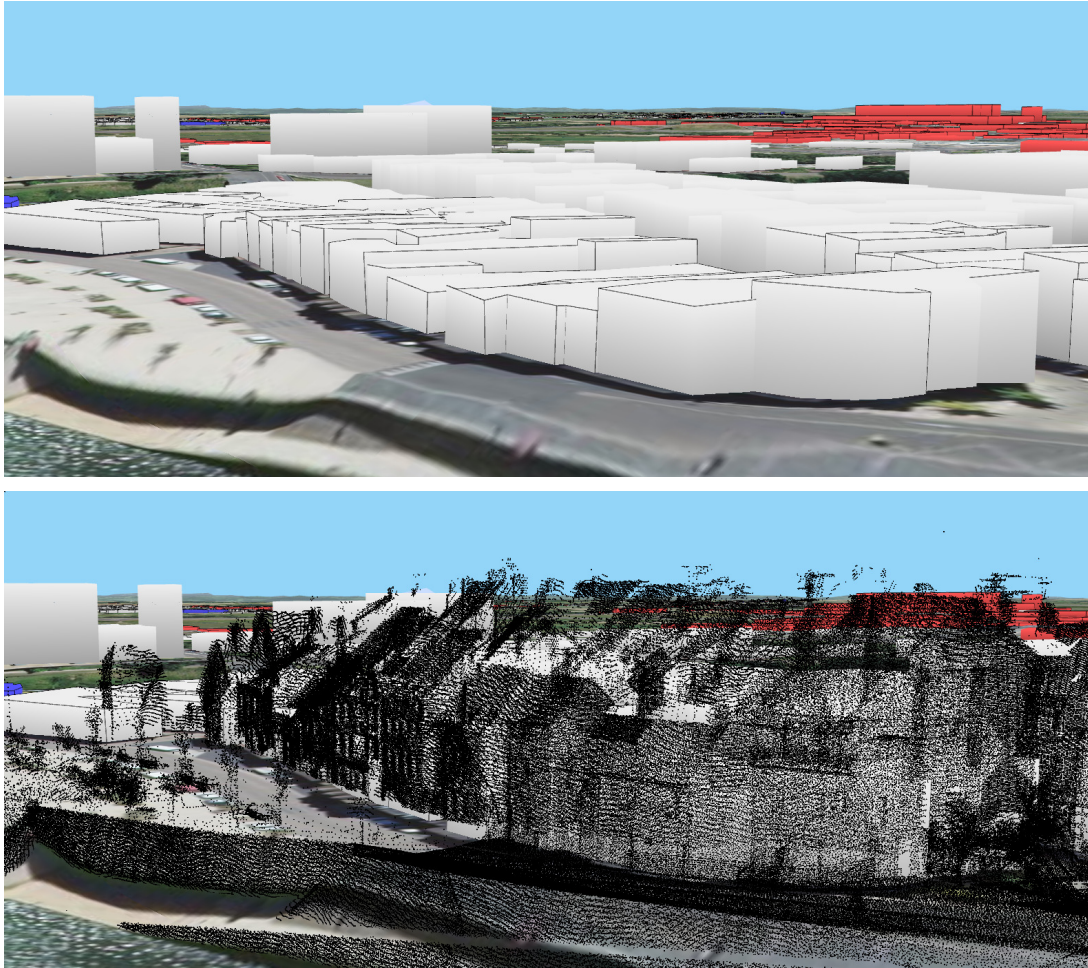


Figure 6.1: Examples of scalable georeferenced 3D models (data from IGN). 1st row: CityGML LoD1 buildings (French "Ref3DNat" reference), available on the whole territory; 2nd row: Superposition with terrestrial LiDAR point cloud acquired by Stereopolis at the scale of the city.

Note that such kind of information is very rich and has proven its relevance to improve spatialization tasks (especially considering detailed models such as 3D point clouds), but these recent acquisitions raise the question of their adequacy facing old iconographic contents potentially associated with landmarks that have evolved.

6.2.2 Manual spatialization

When it comes to spatialization of an image, most basic methods are manual ones. They rarely allow for a high precision and are very costly when faced to a large number of images.

A first, very basic way to add location information to an image is to give it an address as a textual metadata. Indeed, even without converting it to a 2D or 3D position, this textual information can be compared with other addresses to find similar or close ones.

Rather than an address, an image can simply be pinned to a map. Thus, a manual selection of a 2D point can be enough for localizing the image. This is the method used by the French national library's collaborative platform for spatializing its contents (Gallicarte project, 2019) as shown in Figure 6.2. Going one step further, at the time of the 2D point

selection, the user could also manually indicate the coarse orientation of the image (in this case simply the direction on the 2D map).

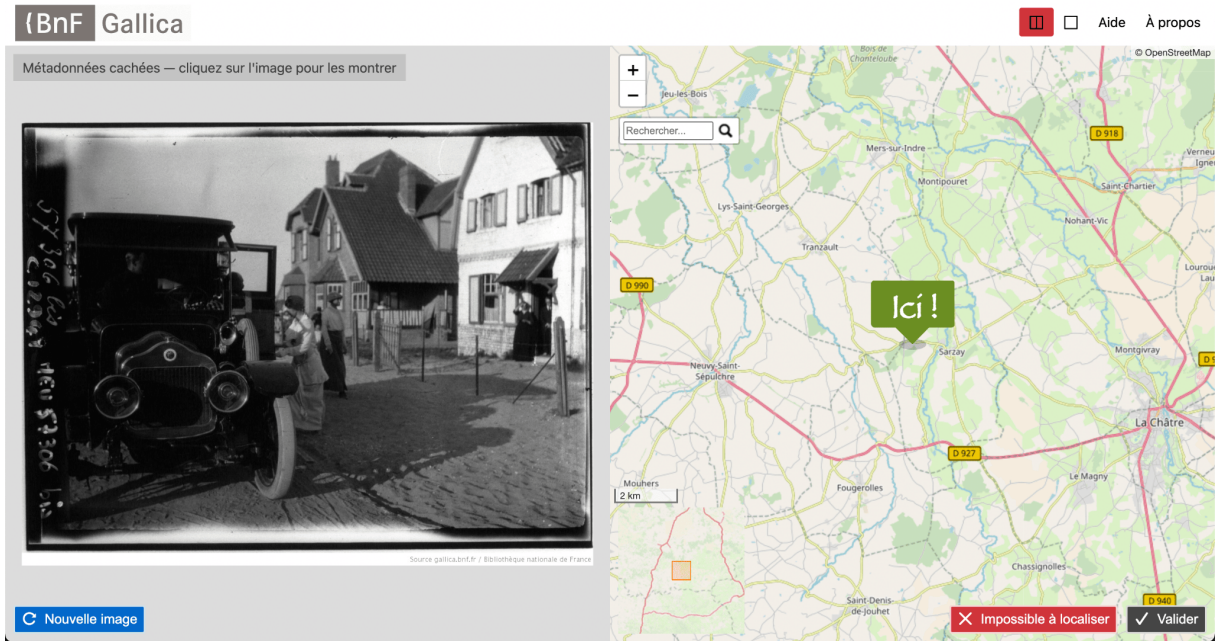


Figure 6.2: Manual 2D spatialization of contents in the French national library’s collaborative platform

A similar setting can also happen in a 3D environment. For instance, HistoryPin (HistoryPin, 2010) (see later in Figure 6.12) allows users to localize images in a pseudo 3D setting. Indeed, the image is located as a layer in an existing pseudo 3D scene, depending on a camera pose from Google Street View. Thus, a coarse position and orientation of the initial camera is proposed, but it depends on another camera, making it a very poor localization of an image. Indeed, should the user move the camera around, the image will move around as well instead of staying at its supposedly correct location.

Those approaches are very costly and lack both precision and reproducibility too much to be an ideal choice for large-scale spatialization.

6.2.3 Semi-automatic spatialization

To alleviate the drawbacks of manual spatialization, especially in terms of processing time cost, semi-automatic methods aim at automating part of the process to improve speed and reduce the cost of manual interventions.

For instance, to assign a 2D point as a location semi-automatically, instead of pointing precisely on a map, geocoding methods exist. Indeed, geocoders will take as input an address and return the corresponding 2D location. Multiple geocoders are available, either free of charge or chargeable. One can mention some like Google Maps, Open Street Map or the one from the French Mapping Agency. Once the point is computed, a user can quickly check its correctness.

The semi-automatic paradigm is more useful for a more complex localization estimation, that is the 6D pose. Indeed, this localization estimates where the camera was when

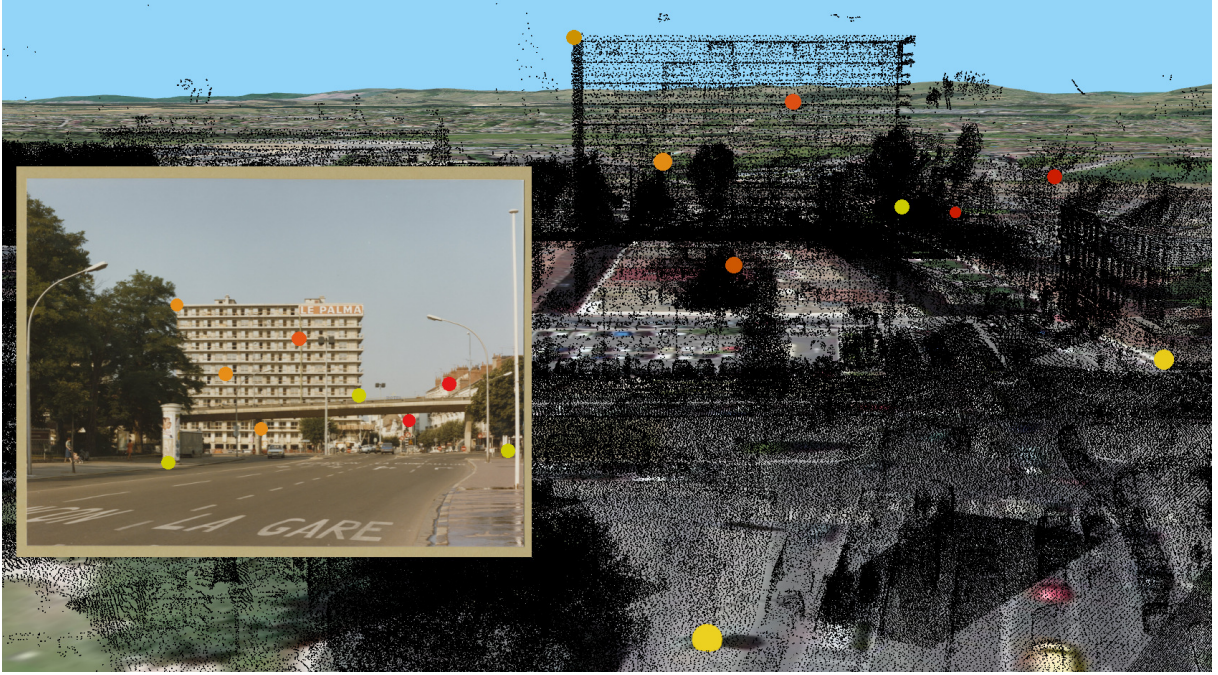


Figure 6.3: Interactive selection of 2D-3D pairs of points (colored bullets) in the photograph and in the 3D scene modeled with LiDAR points, as input of a 6-DoF pose estimation tool (iTowns web application (Blettery et al., 2020)). In this example, we observe differences between the old photograph and the recent version of the scene (disappearance of the bridge, new buildings, roadway modification), which highlights the challenge of the points selection for the pose estimation (images from Musée Nicéphore Niepce and IGN).

the picture was taken, that is a 3D position and a 3D orientation (not simply a generic 2D direction). To this end, the most commonly used method supposes to have a 3D model of the scene and an image. The idea is for the user to manually identify 2D-3D correspondences between the scene and the model (as shown in Figure 6.3) and then automatically use a geometric PnP solver. Of this PnP solver, multiple adaptations have been proposed to deal with more or less complex case. A major difference between two sets of solvers is the previous knowledge (or not) of the camera’s calibration (its intrinsic and extrinsic parameters). Hence, with a calibrated camera and 2D-3D matches (at least 3), several solvers of the PnP problem have been developed such as an efficient P3P (Kneip et al., 2011) or a method accepting more than 3 matches such as EPnP (Lepetit et al., 2009) or PPnP (Fusiello et al.). An intermediary solution when the focal length of the camera is unknown is the P3Pf solver (Sattler et al., 2014). However, when the camera’s intrinsic parameters are unknown, using the Direct Linear Transformation (DLT) (Hartley and Zisserman, 2004) solves it with a minimum of six matches to calibrate the camera, *i.e.* estimating its 6-DoF pose plus its intrinsic parameters (focal length, principal point, skew); this method is known as P6P. Furthermore, all those methods benefit from a RANSAC loop (Fischler and Bolles, 1981) to estimate the best possible 6D pose. Finally, as this takes place in a semi-automatic setting, the estimated pose can be directly visualized by the user to evaluate its quality instantly after computation.

As part of Nelson Fernandes’ internship, under the supervision of Valérie Gouet-Brunet and myself, several experiments were performed to spatialize heritage iconographic content

in a semi-automatic setting (as shown in Figure 6.3). This work, published in (Blettery et al., 2021), aimed at exploiting all available 3D information (coarse building representation and fine-grained LiDAR data) and evaluating the most suited method for 2D-3D pose registration in our context of unknown camera calibration (most common case with iconographic heritage). Several PnP algorithms were also implemented and evaluated with different configurations of intrinsic camera. Visual examples of the conclusions obtained are shown in Figure 6.4.

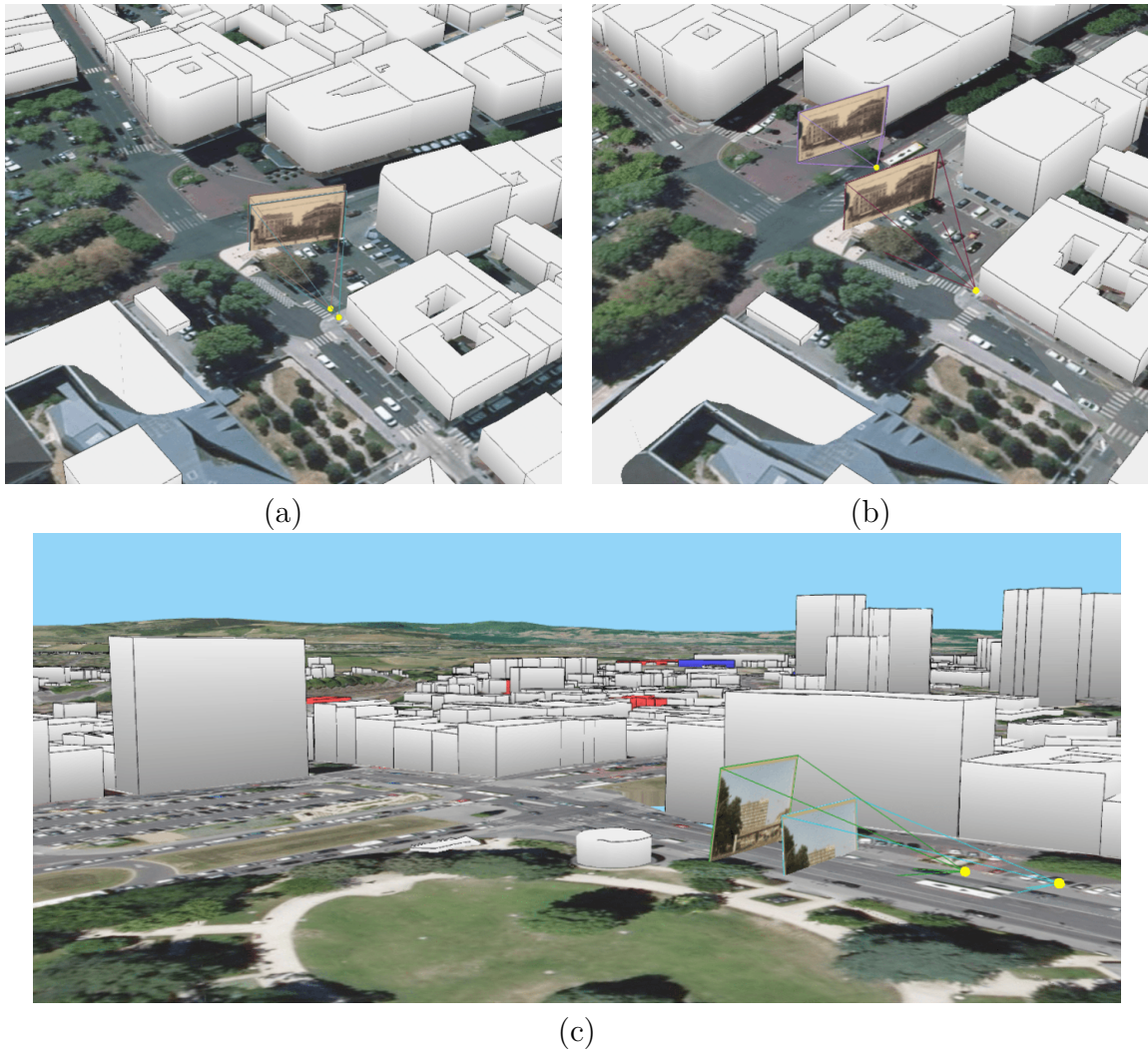


Figure 6.4: Illustration of the influence of the intrinsic parameters on the pose estimation: (a) Calibration estimated with DLT vs. 6-DoF pose estimated with PPnP and intrinsic parameters correctly chosen empirically (the two localizations are similar and correct); (b) Same estimations with different intrinsic parameters for PPnP (localization with PPnP is damaged) and (c) Same estimations as (a) on another more difficult example (localization with DLT is damaged by noisy input points while PPnP’s one remains correct) (images from IGN).

Finally, even if it allows for higher quality localization and more certainty in the outputs, it remains costly in terms of processing time and can prove difficult to do at large-scale.

6.2.4 Automatic spatialization

To fully remove any manual intervention, automatic methods of spatialization have been developed. They can output localization information starting from 2D points up to precise 6D poses. However, the automation of the localization estimation means that verification is not included in the process, and thus may lead to undetected errors, especially when confronted to highly diverse datasets in terms of metadata and visual contents.

Automatic spatialization methods, though not requiring manual input, still require a first coarse idea of where the content is located. Indeed, amongst the methods presented after, whether it is a geocoded address, a pose fused from relative poses or one created via 3D scene regression by a CNN, some information of localization was initially known. First, it was a textual address, secondly, similar images and finally the 3D scene encoded in the CNN. To obtain this starting location, most methods exploit a first CBIR step to limit the number of possible locations, known as Visual Place Recognition. Indeed, finding similar images allows for finding coarse location information, either as an address of a similar image that can be propagated or as close image poses that are a good first approximation of that of the image that is being localized. This first approximated position can be used to initiate a semi-automatic process as seen in Section 6.2.3 but is mostly used for the full automatic approaches described next.

To use once again geocoding methods, once all images of a dataset (or a large number) have a textual address information, either native, propagated after a CBIR step, or assigned manually by an expert user, the transformation of those addresses in 2D locations on a map can be performed as a batch, locating all images in one go. However, once more, as all images are located as a whole, detecting wrongly located images is hard, even using the geocoders own confidence score. It can be even harder when address information in one dataset are not expressed the same way for all images, leading to even more potential geocoding mistakes.

Another type of localization uses similar images' localizations to estimate the query image's one. This can be done for a 2D location, simply averaging the positions of the most similar images. This averaging process can also be weighed using the visual similarity score for instance. The fusion of localization can also be performed for a 6D pose provided the first retrieved images possess a 6D pose. (Song et al., 2016) estimates relative poses between the query image and each first retrieved image to find potential poses. It then minimizes the possible poses adjustments to find the final pose, as illustrated in Figure 6.5. (Pion et al., 2020) fuses candidate poses in a linear weighted combination, the weight being for instance based on the rank of the retrieved image, or based on the similarity of descriptors amongst the first retrieved images. Furthermore, relative poses could also be estimated using RANSAC-Flow (Shen et al., 2020a), which considers dense alignment between images, from which a relative pose could be extracted. New networks have also been proposed (Ding et al., 2019a) to solve this relative pose regression problem. However, due to the need of encoding all reference images poses in the network, the computational cost is high whereas the performance has not yet reached that of classical geometric

solvers.

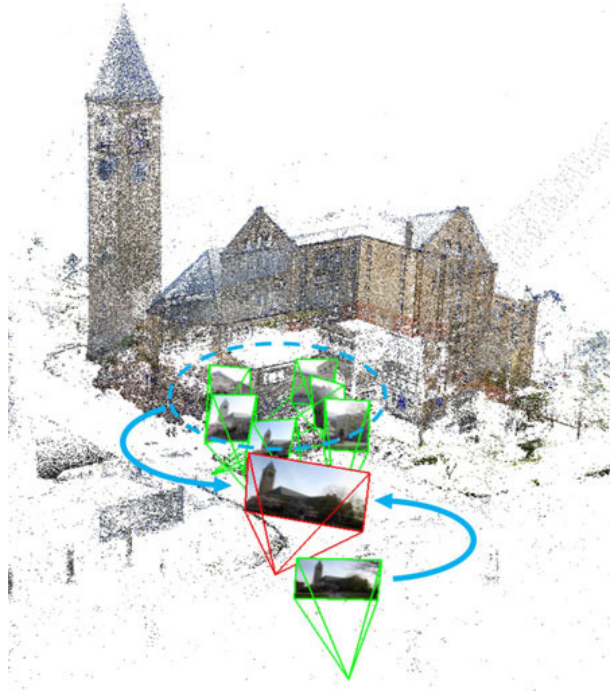


Figure 6.5: Pose estimation by fusing relative poses as proposed by (Song et al., 2016) (figure from (Song et al., 2016))

A different paradigm of localization uses explicitly or not a 3D structure to estimate the pose of an image. This 3D structure of the scene can be computed offline beforehand or computed using the first retrieved images that happen to possess a 6D pose. The idea is always to identify 2D-3D matches between the image to localize and the 3D scene and it can be done in various ways. (Sattler et al., 2016) uses visual words assigned offline to the 3D scene to easily find which part of the scene the image depicts and find 2D-3D matches there. (Brachmann et al., 2017) introduces DSAC, a differentiable RANSAC to include it in a CNN-based pipeline for estimating 2D-3D correspondences. More recently, (Sarlin et al., 2021a) proposed PixLoc which learns features end-to-end for the visual localization task, aligning deep features with a 3D model. (Brachmann and Rother, 2018) uses a CNN for 3D scene regression to estimate scene coordinates for the input image contents, that is 2D-3D matches. The final pose estimation using those matches can be performed using PnP algorithms (see Section 6.2.3), otherwise deep networks can be used to regress the pose.

Finally, to bypass the step of identifying 2D-3D correspondences and the use of a 3D structure, new methods propose to use a trained CNN to directly estimate the 6D pose of an image given as input. Indeed, the network is trained on a whole scene and encodes its geometry, using information from localized reference images. It then produces directly the 6D pose of the input image. This field of CNN-based absolute camera pose regression is well described by (Sattler et al., 2019). Introduced by (Kendall et al., 2015) with PoseNet, this approach of absolute pose regression has been improved on multiple fronts. First, in terms of encoders and decoders for the image features with (Shavit et al., 2021) replacing CNNs by transformers. Others like (Xue et al., 2020) exploit spatio-temporal constraints

to improve the pose regression. To further improve the encoding of the scene, a Neural Radiance Fields (NeRF) approach (first introduced by (Mildenhall et al., 2020)) can be used to create new synthetic views of the scene to further train the pose regression network as proposed by (Moreau et al., 2022). (Moreau et al., 2023) in turn applies absolute pose regression in the context of autonomous driving and regresses the pose hierarchically in real-time within an urban setting. These approaches show promise, but the quality of the obtained poses remains for now sub par with what structure-based approaches offer. Furthermore, it depends on the training of the network on a specific area, areas that may not exist anymore when dealing with iconographic heritage contents.

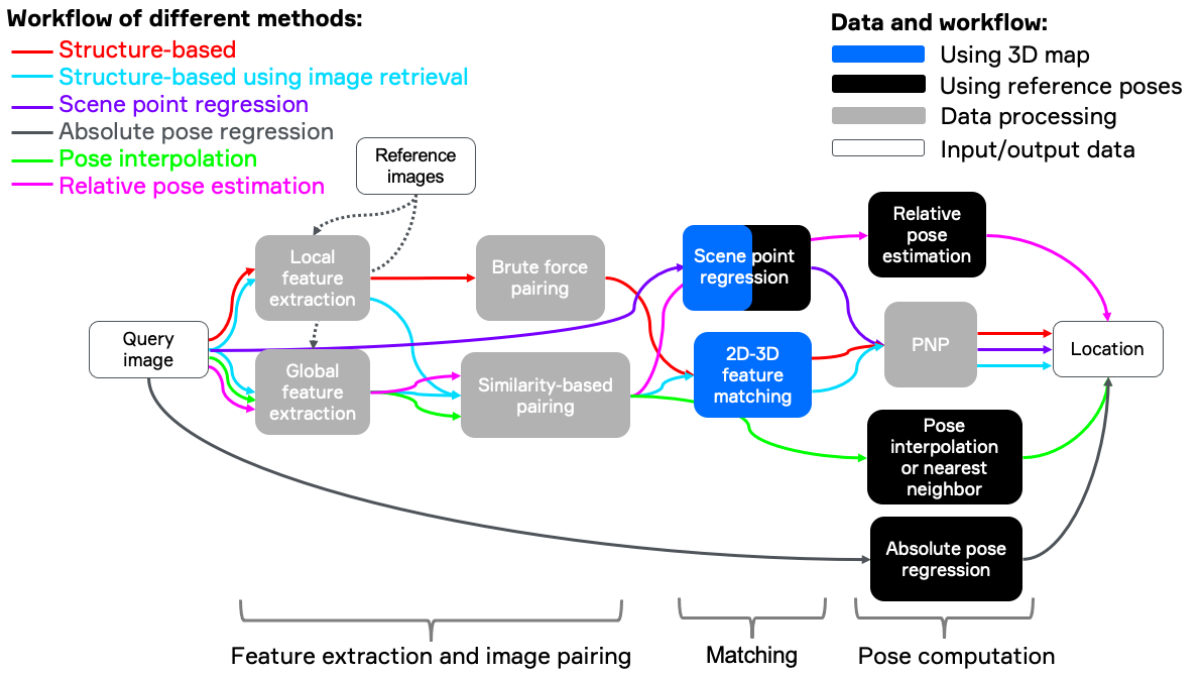


Figure 6.6: Classes of methods for pose estimation (figure from (Humenberger et al., 2023))

All those automatic approaches (whose main classes for pose estimation are summarized in Figure 6.6) are promising in terms of type of localization they offer (6D poses mainly) and in terms of scalability to large datasets. However, as many of these methods rely on visual contents, they suffer from the same drawbacks as CBIR techniques. Indeed, visual heterogeneity mainly decreases performance of those methods which are not trained with such type of data. This is the main reason why spatialization of heritage contents still mainly remains semi-automatic, exploiting collaborative platforms (some are presented in 6.3). However, with the growing spatialization of those contents, state-of-the-art methods may soon be trained to handle a wider range of visual changes between contents.

6.3 Iconographic heritage structuring and visualization

As first introduced, structuring visual contents for visualization purposes becomes paramount at times when web-based platforms are flourishing for users to browse through collections. However, unlike an index intended for specialists only, the structures required for any mundane user to understand it must be as self-explanatory as possible. Several options are available, like metadata based structuring using understandable tags or spatial structuring where the organization of the collections displays itself on a map.

Furthermore, as explained, using visual context or comparative analysis (within the collection or with ancillary data) increases the information that can be gathered through the visual contents, furthering their use in research projects for instance, but also more simply their understanding by basic users.

We present here an overview of existing web-based platforms dealing with structuring and displaying image collections at a more or less large scale. We classify them in categories depending on the data used for structuring and querying the images.

6.3.1 Single modality platforms

Many platforms allow for browsing through their contents based on a single (or at least one largely predominant modality). From metadata to spatialization via visual contents, those platforms structure and display their contents based on one type of information. This section provides an overview of those single-modality based platforms.

6.3.1.1 Metadata based structuring

Several GLAMs developed platforms for displaying their collections on the web. The structure used as support for the organization of contents in the visualization is based on metadata. Indeed, categories of contents are created based on similar metadata, allowing users to browse through data using filters based on tags/metadata. The contents are not particularly organized together and not much more information emerges from the visualization. A few examples are Gallica (French National Library, 2015), the platform of the French National Library or the Base Mémoire (French Culture Ministry, 2019) of the French Culture Ministry but also the Dallas Museum of Art platform (Dallas Museum of Art, 2020).

Going further than simply browsing tagged images, the project Inventer le Grand Paris (Consortium Inventer le Grand Paris, 2017) organizes not merely visual contents but whole research studies based on their metadata, more specifically the object of research. Visual contents are then mainly used to illustrate the research projects.

Further again, the Oronce-Fine platform (Verdier et al., 2017) organizes contents as a graph. Linking contents together using metadata and annotations makes information emerge as the graph endogenously organizes contents and let the structure and internal

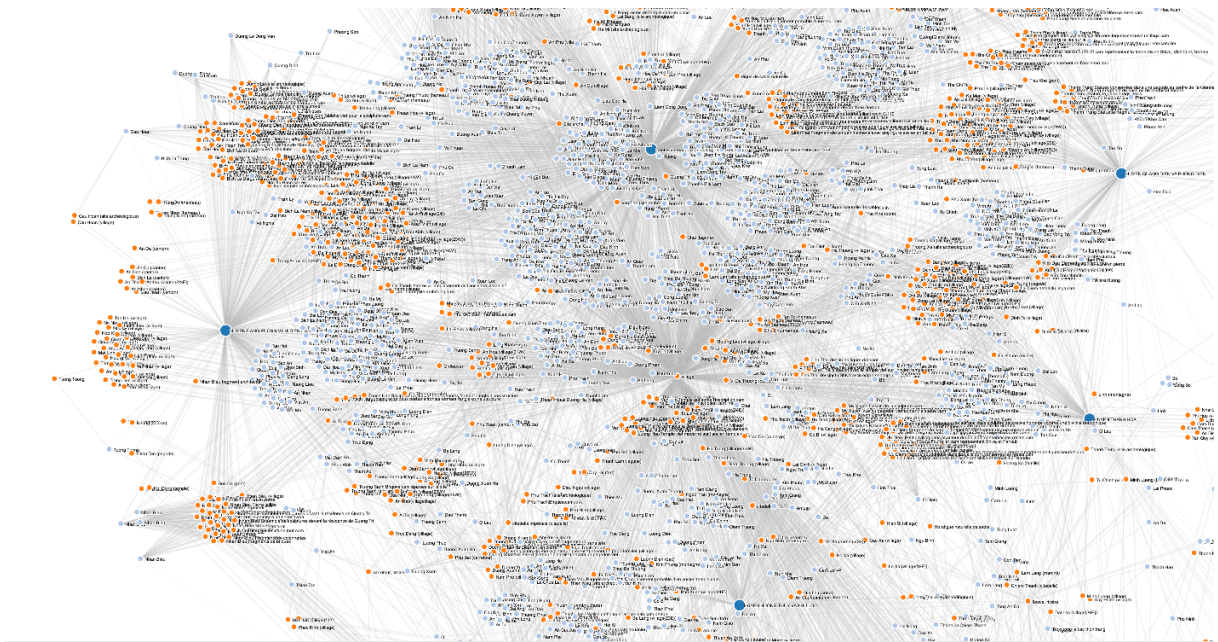


Figure 6.7: Graph of a collection organized in the Oronce-Fine platform

logics appear. Figure 6.7 illustrates the graph representing all links between contents. The dark blue nodes represent iconographic contents, orange nodes are semantic concepts and light blue nodes are the annotations added to the contents.

6.3.1.2 Visual content based structuring

Other platforms only exploit visual content to structure contents and visualize them in the space of the visual descriptors. The images are then not organized using metadata assigned to each image. Groups of images are created automatically as similarly looking images are displayed together while far away images do not look alike. Furthermore, this paradigm of organization allows for visualizing the contents and their structure concurrently. *PixPlot* (Duhaime, 2017) is an example of those platforms, as shown in Figure 6.8.

6.3.1.3 Location-based structuring

A final type of single-modality structure that can be leveraged is the spatial location of contents. Indeed, especially with contents depicting existing scenes, displaying them based on their location easily displays the structure of the dataset and gives context to each image independently and between images.

A basic setup is the platform WhatWasThere (Pup Ventures, LLC, 2021) where users can upload more or less recent photographs with two informations, a date and a 2D location. The contents are then organized as pins on a map on which a user clicks to see the image, as shown in Figure 6.9.

Similarly, the platform (Commission du Vieux Paris, 2023) of the Commission du Vieux Paris (one of our collection provider) displays the locations of the images on a map as points. Clicking on a point reveals the images at this location, alongside more of their associated metadata. The structuring of the collections is purely spatial.



Figure 6.8: Preview of a collection visualized via PixPlot

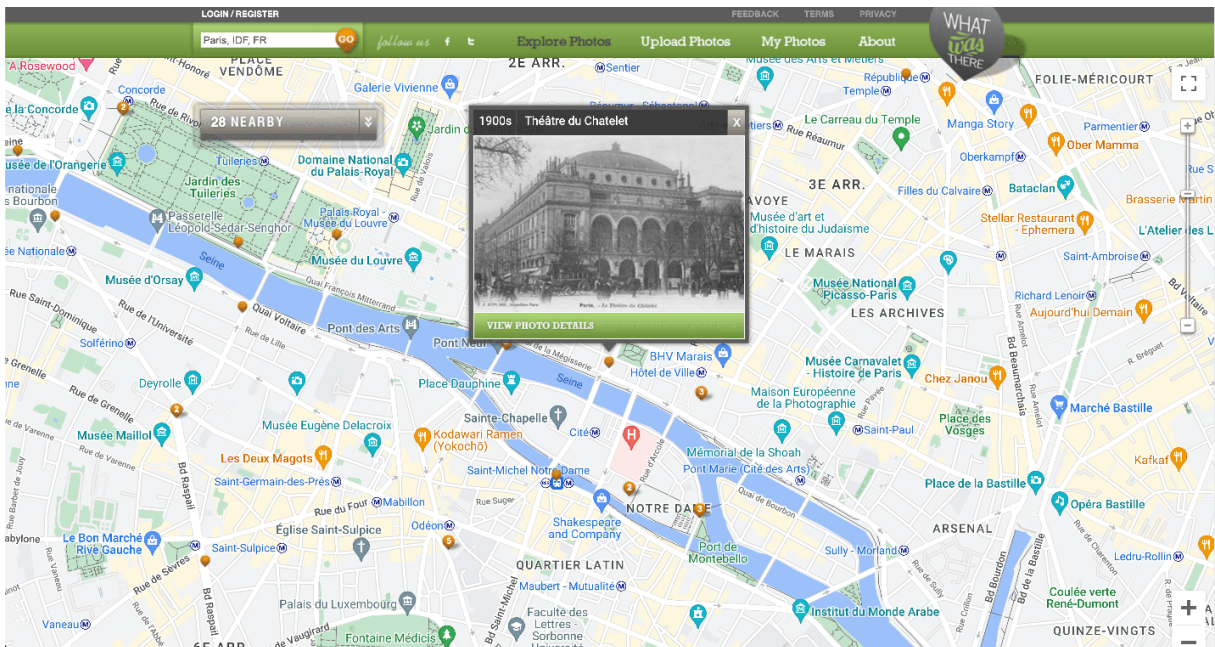


Figure 6.9: Preview of the WhatWasThere platform

Other platforms go a little further, allowing to visualize multiple types of contents jointly to compare them, but always based on a location. That is the case with the Remonter le temps platform (French Mapping Agency, 2016) where the French Mapping Agency (IGN) displays its aerial photographs and maps dating from the 18th century up to now (shown in Figure 6.10). Similarly, the Voyages dans le temps (Office fédéral de topographie swisstopo, 2020) part of the platform from the Swiss topographic agency displays heritage maps and aerial images on a map of Switzerland, alongside more recent contents, should the user wish it.

Going a bit further than simple 2D, the Mapillary platform (Mapillary, 2013) aggregates images from multiple sources and organizes the visual contents based on their 2D

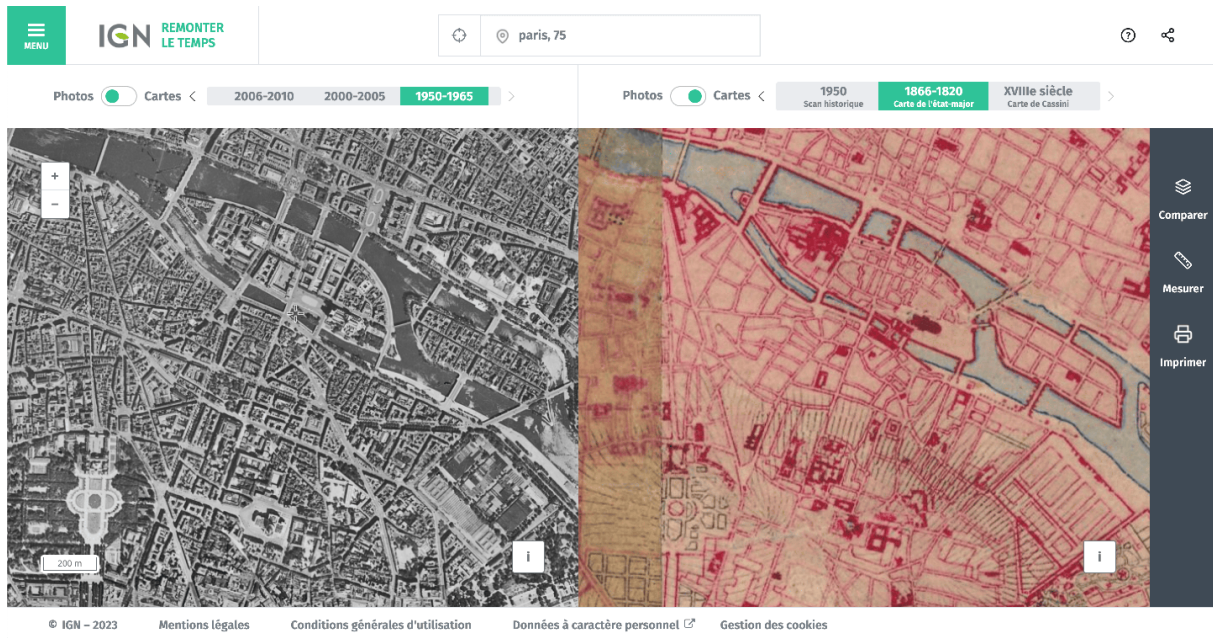


Figure 6.10: Preview of the Remonter le temps platform, comparison of a 1950 ortho-photograph and a 1850s map of Paris' center

location on a map. By clicking on a location, user can see the content located there. Furthermore, if a 360° view of the location is available, the user can move the camera all around the 360°, thus getting a better picture of the global scene.

6.3.2 Multiple modalities combinations platforms

Even though single modalities may be an adequate solution for organizing, structuring and displaying visual contents, the full potential of web-based platform is revealed when using a combination of modalities to structure and then browse through collections. Several combinations have been explored and will be presented next.

6.3.2.1 Metadata and 2D location based structuring

A first combination of modalities uses metadata and 2D location. The main advantage of this combination resides in the possibilities to filter contents in two different ways, either spatially or using metadata tags. Combining them in a different order may lead to different visualization of the global structure of the dataset and then let different conclusions emerge from browsing the dataset. For instance, one can first choose the type of data to visualize and then, based on their spatial distribution select their area of study. However, selecting first the area of study and then asking for specific data can reveal an absence of data revealing another issue worth studying.

A first simple example of dual structuring is the platform of the Albert Kahn museum (Albert Kahn museum, 2016) which uses it to promote its collections. Indeed, the 2D map displaying locations of the images is mostly useful for people to focus on places they know and spatially select them for browsing and downloading.

Navigae (Navigae, 2018) extends the previous approach to research projects and datasets

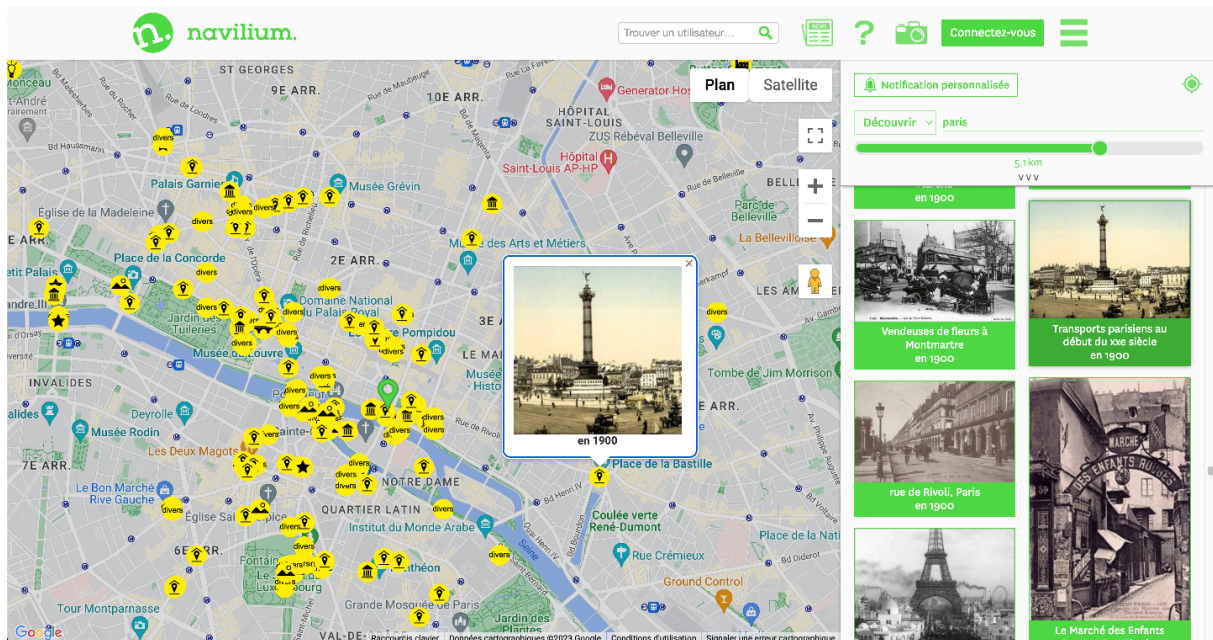


Figure 6.11: Preview of the Navilium platform, with a located image selected

in geography. Visual contents used in projects or studies are located on a map but can also be queried based on their date, on tags about what they represent or even on the subject of the study they are related to.

Other examples of this combination of information for structuring visual contents are set in a paradigm mixing collaborative platforms and social network.

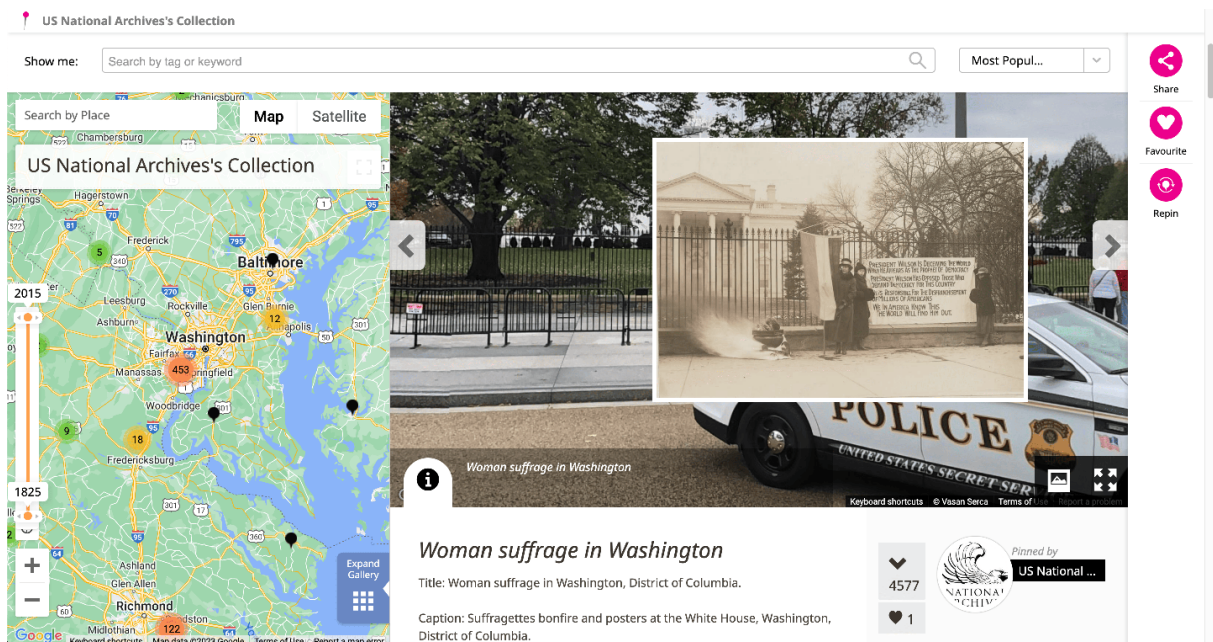


Figure 6.12: Preview of the HistoryPin platform, with an image visualized in pseudo-3D in the scene

Navilium (Navilium, 2016) first lets users upload images, assign a date, a 2D location and other descriptive tags. Users can then browse through all available images by indicating several informations. It can be a location (they can also move around on the map, changing the area of focus will change the images visible), a time span or tags. Images are

then filtered and pins are placed on the map, pins on which the user can click to visualize the image associated. One can also simply browse through all images pertaining to its search criteria (on the right-side of the web-browser). An example of the visualization is shown in Figure 6.11.

HistoryPin (HistoryPin, 2010) finally is very similar to Navilium in terms of browsing through and querying the contents. However, a specific difference is the possibility to spatialize the image in pseudo-3D (as explained in Section 6.2.2). Indeed, users can place coarsely the images in the scene of a recent camera from Google Street View, trying to give it more context. An example is shown with Figure 6.12.

6.3.2.2 Metadata and specific 3D spatialization structuring

Rather than simply using 2D location, new platforms aim at using a 3D localization in conjunction with metadata information for visualization purposes. However, obtaining a general 3D model and localizing contents within may be quite costly. Thus, some platforms first exploited local 3D models, suited to their needs, their specific collections.

A first platform in this category is the project Hist4D (Maiwald et al., 2019). It proposes a 4D web browser to visualize a whole collection of images from 1820 up to now of the city of Dresden (see Figure 6.13). A 3D model of the city has been specifically constructed for this purpose. Images are localized using a 6D pose, user can visualize the images in their context by placing the view at the camera location and multiple analysis tools are available (heatmap of image density, clusters of image orientations, etc.).

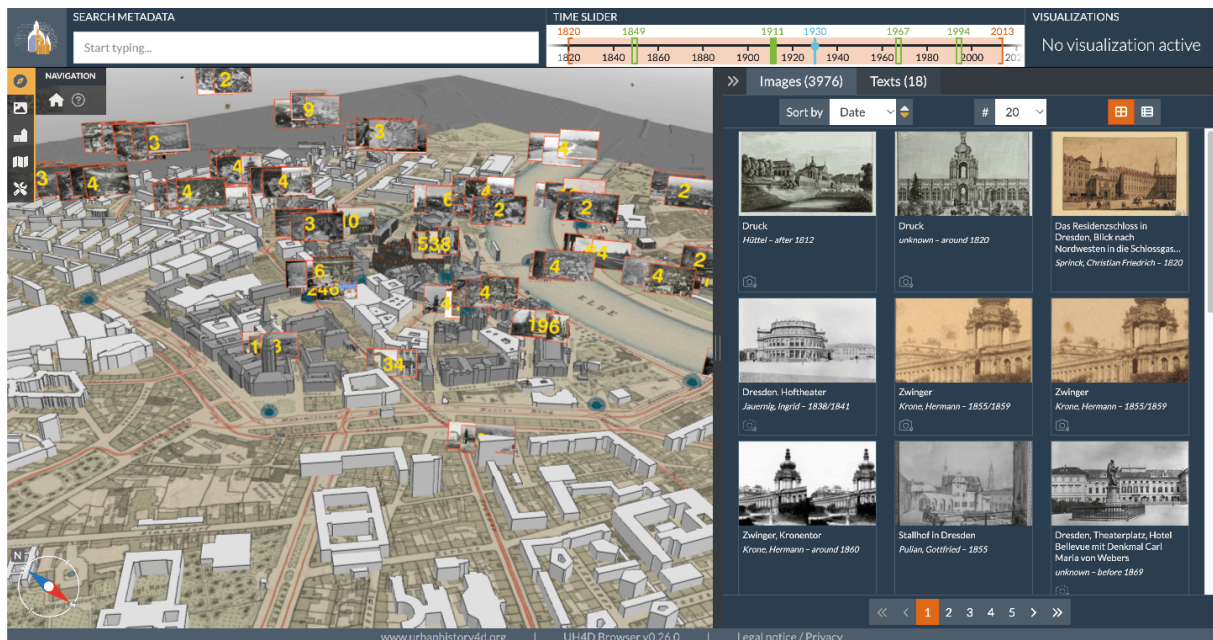
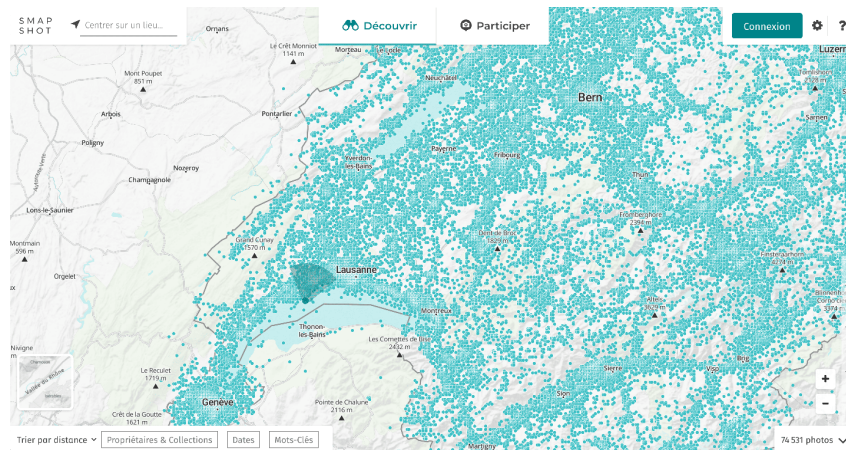


Figure 6.13: Preview of the Hist4D platform, where images are located in 6D within a tailored 3D model

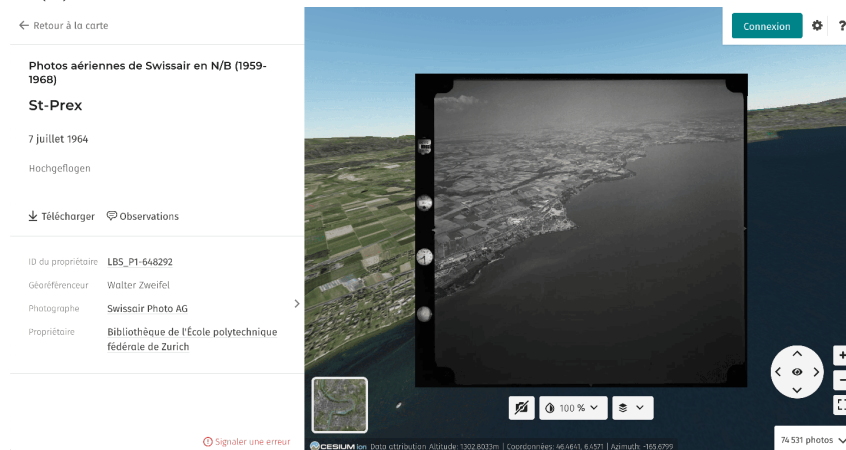
The Aioli platform (MAP laboratory, CNRS, 2017) also organizes image contents using a 3D structure, the annotations on image being propagated through the 3D to other images. However, the 3D cloud exploited is reconstructed using the images and

thus remains a specifically created 3D structure for the image data considered. Another image outside the specific project could not be located based on this structure. The structuring is done "locally".

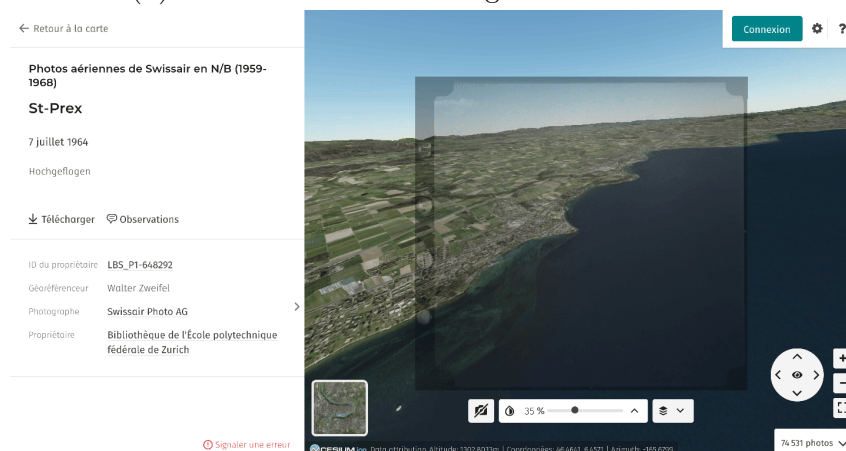
6.3.2.3 Metadata and global 3D spatialization structuring



(a) Map of all geolocated images, with their coarse orientation



(b) Visualization of the image in its 3D context

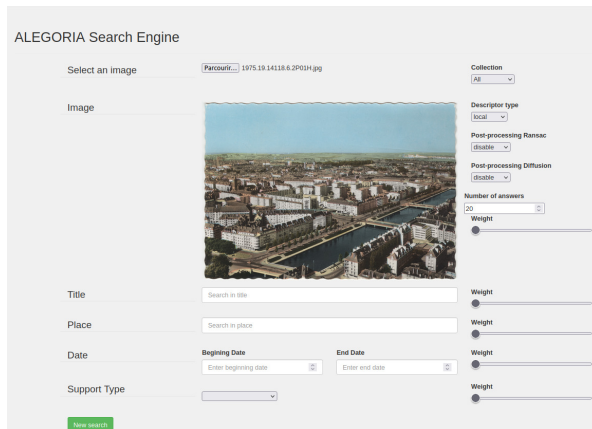


(c) Transparency-based comparison between old and recent scenes, proving the quality of the localization process

Figure 6.14: Preview of the SmapShot platform, displaying an image in its 3D context

The SmapShot platform (Blanc et al., 2018; Produit et al., 2018), integrated now in Images of Switzerland Online (Swiss Art Research Infrastructure, 2022) proposes to query contents based on metadata, or concepts but also browse through them on a map and visualize them in their 3D context, as shown in Figure 6.14. The spatialization is considered more global as it is over the entirety of Switzerland and could be easily adapted to other countries, provided a similar 3D model exists (potentially easily available from National Mapping Agency or Google Earth for instance). However, even though the images are displayed in their global environment, the user can not move around freely in the scene and visualize other close images (like in the Hist4D platform for instance). This drawback limits the full use of the context of the global scene. It prevents from visualizing other potential representations to enrich the simply visual data.

Finally, all structuring paradigms can be combined in a single platform. That is the proposition of the ALEGORIA project (ALEGORIA project, 2018). It proposes a combination of a search engine combining metadata and visual similarity (see Figure 6.15) and a 3D visualization platform (see Figure 6.16) to display the images in a global 3D context.



(a) Search of similar images

Results (by decreasing order of similarity)

Image	Title	Date	Location	Collection	Institution	Comments	Cocoref
	Vue aérienne de Caen, Les quatre ponts (pre factice)	Créat. Vers 1950	Caen	Combar	Musée Nicéphore Népce	-	show
	Vue aérienne de Caen (bte factice)	1950	Caen	Combar	Musée Nicéphore Népce	-	show
	Vue aérienne de Caen, Les quatre ponts (pre factice)	1950	Caen	Combar	Musée Nicéphore Népce	-	show
	Mission n° 864 : Caen (Colivade)	1956-09-14	Caen	MRU	Archives nationales	Reproduction libre, dans le respect des dispositions contenues dans le règlement de la salle de lecture.	show
	Vue aérienne de Caen, Les quatre ponts (pre factice)	1950	Caen	Combar	Musée Nicéphore Népce	-	show
	"B.N.30", image numérique "20K"	-	Caen	LAP 2	Archives nationales	Mention obligée : cédée Archives nationales/Fonds LAP 2. Toute réutilisation est interdite sans l'accord des ayants droit.	show

(b) List of retrieved similar images

Figure 6.15: ALEGORIA Project Search Engine, images from (Geniet et al., 2022)

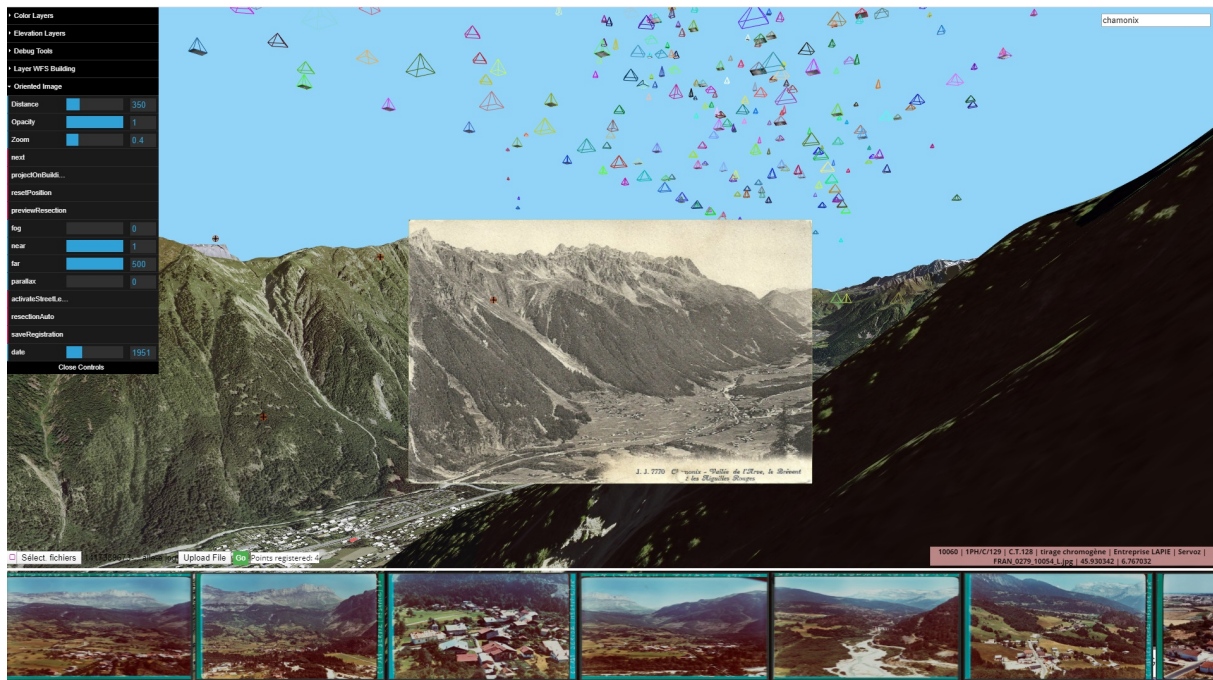


Figure 6.16: ALEGORIA project visualization platform, allowing for the visualization of multiple images at the same time (all "pyramids" correspond to an image), image from (Blettery et al., 2020)

Similarly to SmapShot, the images are displayed in a 3D scene. However, the user is much more free. First, they can upload their own images to be compared visually with existing images. Images can also be located in the 3D scene by estimating their 6D pose, using a semi-automatic approach of selecting 2D-3D matches as presented in Section 6.2.3. On the contrary, in SmapShot or Hist4D, the 6D poses are not editable. Users can also freely move around in the scene, jointly visualize and compare images. Furthermore, the modularity of Itowns¹ -the 3D platform used for visualization- allows for visualizing thematic data alongside images, further enriching the analysis that can emerge from the dataset (see (Blettery et al., 2020)).

The number of images that can be visualized at the same time even led (Paiz-Reyes et al., 2021) to investigate visualization paradigms to allow for a better browsing. This is to ensure that the global and large-scale visualization does not become detrimental to the good structuring and understanding of the dataset.

From a different perspective, with a focus on urban data and more specifically the visualization of its evolution, the virtual city project (Liris Laboratory Vcity Team, 2023) works on modelling urban data, more specifically in 4D. Furthermore, it enriches the visualization with analysis tools and semantic information, within the reproducible framework UD-SV presented in (Samuel et al., 2023). Part of the project described in (Jaillot et al., 2021) also focuses on bringing to the visualization iconographic multimedia contents to further enrich the analysis of the evolution of the urban setting considered. This is illustrated in Figure 6.17 where an image from 1856 is compared against the contemporary 3D model of the city. This type of content visualization could be possible on any scene,

¹<https://www.itowns-project.org/>

making it a global structuring approach. Their spatialization of the iconographic contents is however much coarser and their visualization options are very limited.

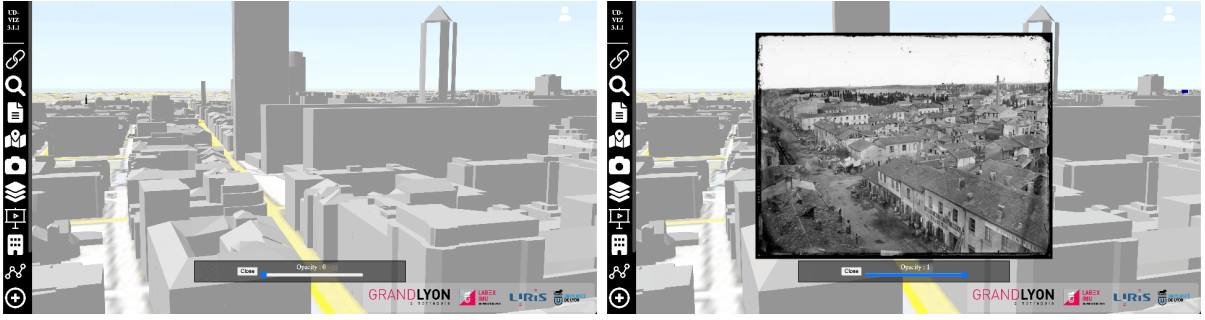


Figure 6.17: The heritage content coarsely spatialized and visualized in the 3D scene of the *UD-Viz* platform from (Jaillot et al., 2021)

6.4 Conclusion

In this chapter we provided two overviews related to spatial-based structuring and visualizing image collections.

The first one, on image spatialization, presented methods requiring more or less manual input, and resulting in various kinds of spatialization. From a simple address to a 6D pose in a 3D scene, the images can be spatialized in multiple manners, which in turn allows for various paradigms for organizing and visualizing them at large-scale.

The second one, based partly on the first one, describes existing solutions for structuring, displaying and visualizing visual contents (heritage or not) in web-based platforms. Mainly, platforms are different from one another due to the type of data they use for the structuring (metadata, localization, visual similarity, or any combination of those) and the visualization paradigm they choose. However, another important difference to note is the potential or not for the structuring to evolve. Indeed, the structuring can be done beforehand (*e.g.* GLAMs showcasing platforms or Hist4D) and then be fixed for the visualization. It can also be evolutionary, as proposed by the social network approaches (HistoryPin and Navilium) but also the platform from ALEGORIA. However, evolutionary structuring can be difficult to scale to the thousands of heritage iconographic contents available, especially if the quality of the structuring is to be preserved. Furthermore, even platforms that support evolutionary structuring and could be adapted to other areas or datasets are still designed with a specific area and a specific data in mind. This remains an obstacle to their actual generalization. Thus, the step of co-modelling data and visualization, as done by (Samuel et al., 2023) for instance, must be encouraged in order to attain generalization and reproducibility for the platforms.

In light of all those observations, both on automatic content-based approaches from Part I and on existing structuring and visualization platforms from this chapter, we focus on a new web-based visualization and structuring platform. More than a platform for visualizing structure created offline, we propose a complete workflow combining structuring and visualizing through the combination of automatic and manual processes, each

working for the other in a virtuous cycle to improve global structuring and allow for a more meaningful visualization of the dataset. This will be presented and discussed in Chapter 7.

Chapter 7

Graph-based Semi-automatic Re-ranking

7.1	Introduction	143
7.2	Graph representation of the structured dataset	144
7.2.1	Graph links considered	145
7.2.2	Presentation of the graph	147
7.3	Structuring process overview	150
7.3.1	Semi-automatic iterative process	150
7.3.2	Location propagation	152
7.4	Graph-based visualization platform	152
7.4.1	Visualization needs and technical solutions	153
7.4.2	The visualization platform	154
7.4.3	Visual clues as analysis support	159
7.5	Semi-automatic structuring process evaluation	166
7.5.1	Automatic quantitative evaluation	166
7.5.2	Qualitative visual evaluation	168
7.6	Conclusion	172

7.1 Introduction

To browse or structure iconographic collections, a natural way would be to exploit metadata describing contents with tags or location information for instance. However, the image collections we target are often organized in silo, each with their own metadata model for describing contents depending on their specific needs. This heterogeneity of the metadata led us to exploit content-based image retrieval approaches (see Part I) which prove to be state-of-the-art for relatively homogeneous contents of image collections. However, the evaluation of those automatic methods showed that they can not perform optimally due to the visual heterogeneity of the considered contents. Nonetheless, we demonstrated that exploiting structuring information at query-level or at dataset-level proves useful for

re-ranking, especially the graph-based diffusion process.

This leads us to two main conclusions, exploiting the largest possible structure is informative for re-ranking purposes and automatic structuring is often flawed for the most complicated cases that only an expert can solve.

Hence, we introduce in this chapter a graph-based semi-automatic structuring proposal leveraging automatic approaches to both create a first structure to be evaluated by an expert user and then to propagate the certainty of the user's interventions to firm up or modify the existing structure, overall multiplying the impact of the targeted expert interventions. The evaluation of the dataset's structure is performed in a 3D, graph-based, web visualization platform illustrated in Figure 7.1.

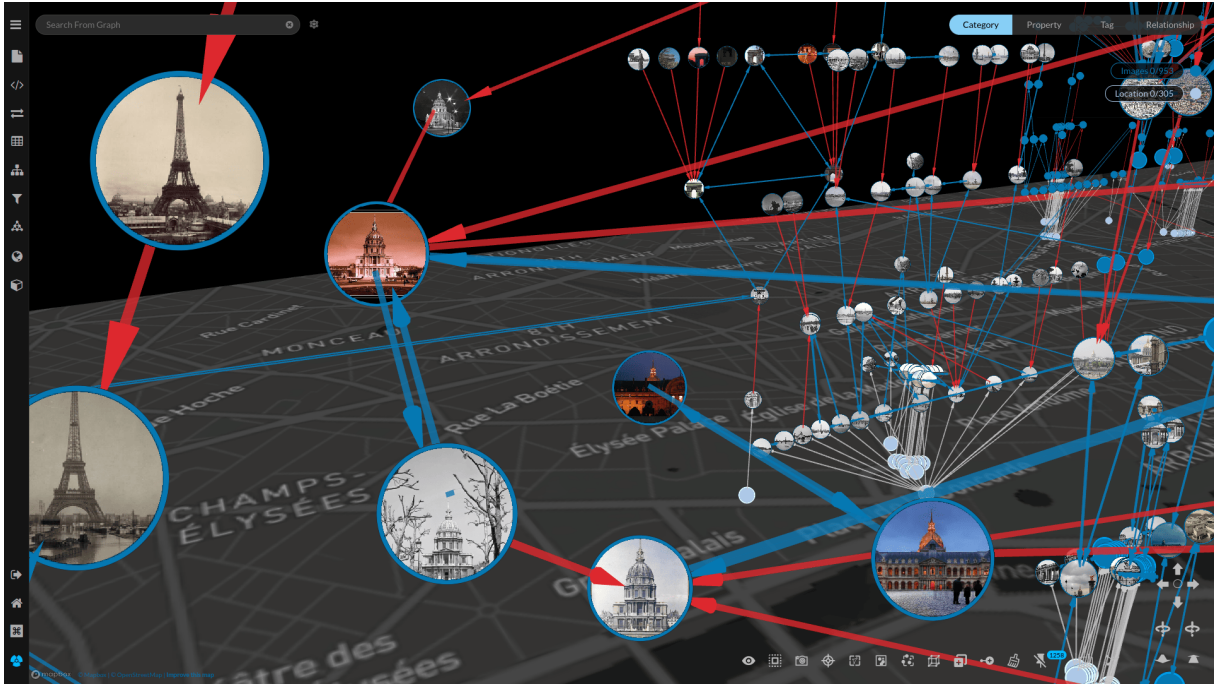


Figure 7.1: The 3D graph-based visualization platform

This chapter will first introduce in Section 7.2 our paradigm of representation of the structure of the dataset as a graph. The semi-automatic structuring process proposed is then detailed in Section 7.3. Section 7.4 then presents the visualization needs and the platform and visual clues proposed to serve the structuring process. Finally, Section 7.5 evaluates quantitatively and visually the relevance of our structuring process.

7.2 Graph representation of the structured dataset

On the one hand, diffusion-based re-ranking approaches proved very efficient for structuring contents. On the other hand, our previous conclusion was that "larger-scale" structuring information were beneficial to the global structuring process. Thus, we decided for this part to set ourselves in a paradigm where the different categories of links built are organized into graphs (including various structuring information), and to exploit this representation of the structure for re-ranking purposes. Looking at our structuring prob-

lem as being set within a graph indeed allows us to use graph-based algorithms (like diffusion-methods for instance) but also to visualize it using tools designed for graph visual analytics.

7.2.1 Graph links considered

To provide structure to the image collections, we build and exploit five kinds of links between items, some are shown in Figure 7.1. We detail next all five types of links and the similarity they encode.

Location links

They connect a location to any image at this location. Two images with matching location information coming from a different source (*e.g.* GPS or geocoded address) can be connected to the same location. The confidence $c_I \in [0, 1]$ given to the link between and image I and its location depends on the way the location was estimated, as explained in Section 4.3 of Chapter 4.

Visual similarity links

Those links connect two images similar in terms of visual content. Such links do not exploit the metadata and are processed automatically with the best approaches for image retrieval (in our case the descriptor How-A) described in Section 3.4 of Chapter 3. They encode the visual similarity score $s_{I,J}^v \in [0, 1]$ between two images I and J .

Spatial similarity links

They encode the pairing of two images according to their spatial proximity in the environment, if available. Other criteria could have been chosen (*e.g.* semantic similarity) but the geolocation criterion is a useful one in several domains, *e.g.* the geographical heritage. They can represent simply an information of spatial proximity that can validate or contradict visual similarities but also be used to propagate locations as shown in Section 7.3.2.

Those links can be automatically computed, as proposed in Section 4.3 of Chapter 4, but also manually created. They encode the spatial similarity score $s_{I,J}^s \in [0, 2]$ defined in the aforementioned Section. Furthermore, they are also assigned a spatial confidence score $c_{I,J}^s \in [0, 1]$, either manually defined or automatically computed (see Section 4.3 of Chapter 4). The similarity score is the weighted by the confidence score, which leads to the previously presented (in Section 4.3 of Chapter 4) spatial similarity weight $w_{I,J}^s = s_{I,J}^s c_{I,J}^s \in [0, 2]$.

Expert similarity links

Those are manual connections added by an expert to connect two images according to its own structuring paradigm, visual, spatial, semantic or other. This similarity thus encodes more information than a simple similarity link like visual or spatial links. Those links

encode an expert similarity score $s_{I,J}^s \in [0, 1]$ associated to an expert confidence score $c_{I,J}^e \in [0, 1]$ whose value depends on the level of expertise of the user.

Global similarity links

To exploit all types of similarity during the diffusion re-ranking step presented in Section 7.3, we aggregate all three previous similarities to create the strongest possible similarity links, exploiting all available information.

Computing the final global similarity score $S_{I,J}$ between two images I and J , used for re-ranking all images, using all possible information, is done as follows:

$$S_{I,J} = \mathcal{N} \left(s_{I,J}^v \times s_{I,J}^s c_{I,J}^s + (s_{I,J}^e)^{c^e} \right) \quad (7.1)$$

It consists of the five similarity and confidence scores presented before and developed further in the following:

- the visual similarity score $s_{I,J}^v \in [0, 1]$, computed between the two image descriptors during retrieval;
- the spatial similarity score $s_{I,J}^s \in [0, 2]$, that can be estimated two ways: either by estimating a proximity score (see Section 4.3 of Chapter 4) or set to the maximum when reflecting a spatial similarity manually added (*e.g.* 2 views of the same monument without considering any geolocalization information). A score over 1 confirms the probable visual similarity, under 1, it denies it, and if it equals 1 it reflects a lack of spatial information or a distance that is not meaningful to weight the visual similarity. The importance given to this score relies on the fact that we focus on spatialized and non-movable objects for whom spatial similarity is obvious while the visual aspect can change through time or viewpoint;
- the spatial confidence score $c_{I,J}^s \in [0, 1]$, that reflects either the quality of the estimated proximity score (based on source locations' quality (see Section 4.3 of Chapter 4), the user's confidence or the confidence in the location propagation process). When automatically computed based on location quality, it is obtained as:

$$c_{I,J}^s = \begin{cases} \frac{1}{c_I \times c_J} & \text{if } s_{I,J}^s < 1 \\ c_I \times c_J & \text{otherwise} \end{cases} ; \quad (7.2)$$

- the expert similarity score $s_{I,J}^e \in [0, 2]$, that reflects the opinion of the user as to similarity between the two images based on their visual evaluation. The similarity reflects the specific structuring the expert wants to bring to the dataset, it can aggregate visual, spatial or any other type of similarity specific to the dataset. Set to 0 if no expert similarity exists between the images, it is added to the previous scores because such validation of similarity must increase the global similarity score (if no similarity should exist, the link would be deleted by the expert);

- the expert confidence score $c^e \in [0, 1]$, whose value depends on the level of expertise of the user.

Finally, we normalize $S_{I,J}$ in $[0, 1]$ using \mathcal{N} , a min-max normalization over S . Figure 7.2 illustrates the nodes and links presented before.

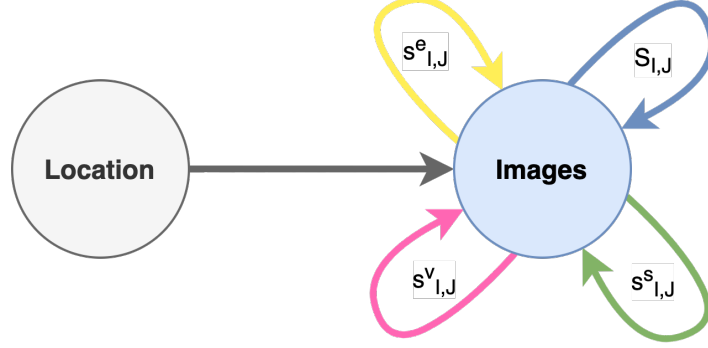


Figure 7.2: Minimal representation of the graph-based representation of the dataset

One can notice that should no expert nor spatial similarity be present, $S_{I,J}$ represents simply the visual similarity (meaning no re-ranking of any sort), and should simply no expert similarity be present, $S_{I,J}$ encodes the visual similarity weighted by the spatial similarity as defined in Section 4.3 of Chapter 4. Our global score only builds on existing structuring information to encode the most information possible.

7.2.2 Presentation of the graph

Exploiting the links considered in the previous section, we define a graph, described in details, formally and visually, in this section.

The links presented in the previous section lead to a multi-edge graph-based representation of the collections' structure, where the nodes are either image contents (and metadata) or a location. We formally define here the global graph G :

- Let us call G the global graph, $G = \{V, E\}$;
- We highlight two subgraphs:
 - $G_I = \{V_I, E_I\}$ the subgraph of image nodes,
 - $G_L = \{V_L, \emptyset\}$ the subgraph of location nodes,
 - with $G = G_I \cup G_L$;
- Furthermore, $V = V_I \cup V_L$ and $E = E_I \cup E_{LI}$;
- Amongst the edges, there are first $E_{LI} \subset [0, 1] \times V_I \times V_L$, edges linking location nodes to image nodes. One such weighted edge is defined as (c_I, v_I, v_L) with v_I the image node, v_L the corresponding location node and c_I the confidence score associated;

- Second, the edges $E_I \subset V_I \times V_I$ linking two image nodes. As described before, those edges can be of four types, leading to $E_I = E_I^v \cup E_I^s \cup E_I^e \cup E_I^g$. They are detailed below with an example of each type of edge linking two image nodes v_I and v_J :
 - $E_I^v \subset [0, 1] \times V_I \times V_I$ linking images with a visual similarity. An example edge is $(s_{I,J}^v, v_I, v_J)$, with $s_{I,J}^v$ the visual similarity score;
 - $E_I^s \subset [0, 2] \times V_I \times V_I$ linking images with a spatial similarity. An example edge is $(s_{I,J}^s, v_I, v_J)$, with $s_{I,J}^s$ the spatial similarity score;
 - $E_I^e \subset [0, 1] \times V_I \times V_I$ linking images with an expert similarity. An example edge is $(s_{I,J}^e, v_I, v_J)$, with $s_{I,J}^e$ the expert similarity score;
 - $E_I^g \subset [0, 1] \times V_I \times V_I$ linking images with a global similarity. An example edge is $(S_{I,J}, v_I, v_J)$, with $S_{I,J}$ the global similarity score;
- Using those different edges, we can extract multiple subgraphs out of the global graph, depending on the similarity exploited:
 - $G_{LI}^v = \{V, E_I^v \cup E_{LI}\}$ the subgraph of image and location nodes with location links and visual similarity ones;
 - $G_{LI}^s = \{V_I, E_I^s \cup E_{LI}\}$ the subgraph of image and location nodes with location links and spatial similarity ones;
 - $G_{LI}^e = \{V_I, E_I^e \cup E_{LI}\}$ the subgraph of image and location nodes with location links and expert similarity ones;
 - $G_{LI}^g = \{V_I, E_I^g \cup E_{LI}\}$ the subgraph of image and location nodes aggregating all previous links;

After some iterations of the semi-automatic process we propose in Section 7.3, the graph-based representation may contain data (in nodes and edges) of three natures that all serve the understanding and structuring of the dataset in and out of the visualization:

- *source* data (from providers), *e.g.* annotations or location information;
- *automatic* data (computed), *e.g.* similarity scores between images, or new propagated locations;
- *manual* data (added by user) *e.g.* new manually added similarities or image annotations.

Some of those different informations within the graph structure are presented in the diagram in Figure 7.3.

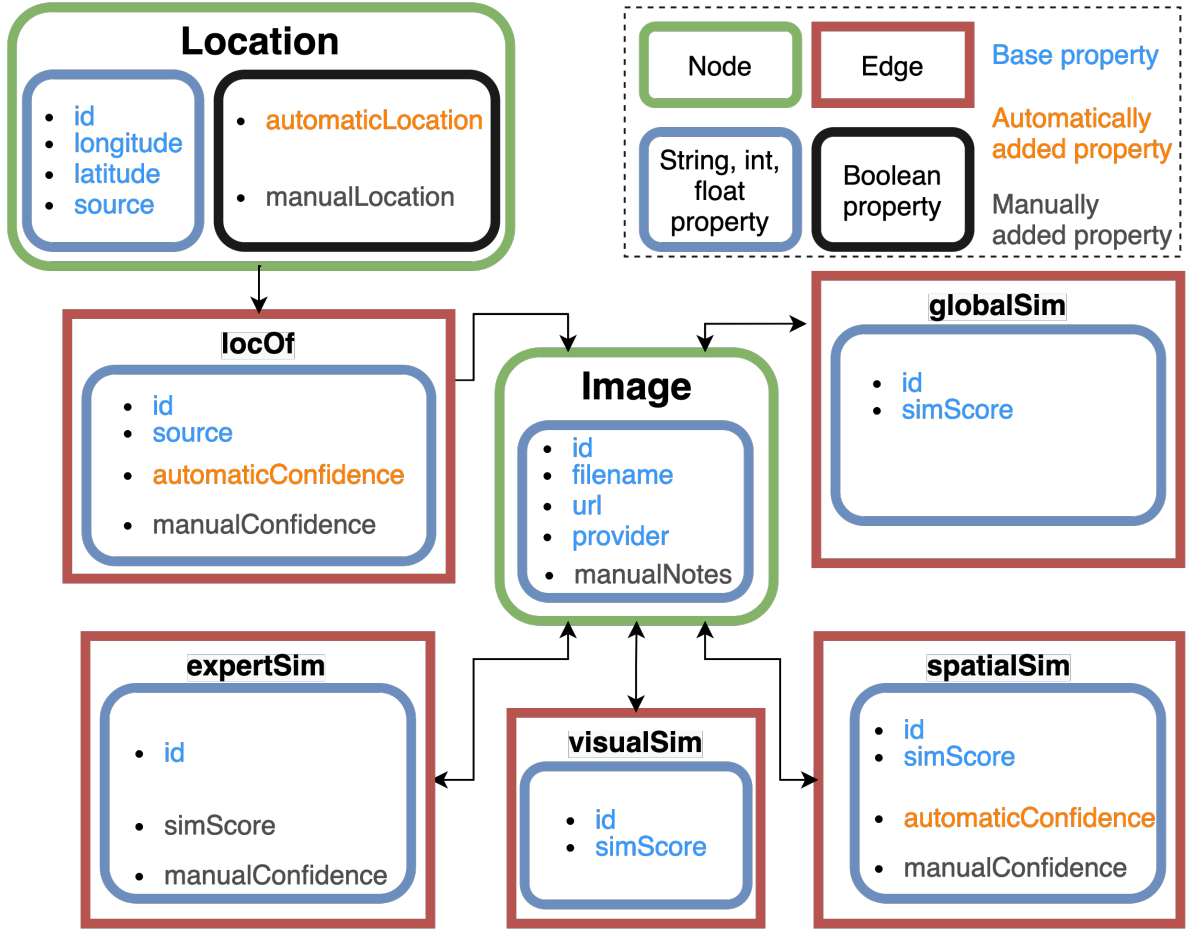


Figure 7.3: Detailed representation of the graph-based representation of the dataset

As explained in Section 7.2.1, similarities or locations can be associated to a confidence degree -depending on user’s skills and/or on some parameters of the automatic processes- which counts during the automatic structuring steps of global similarity creation or location propagation (see details in Section 7.3).

To further present our graph-based representation, we illustrate below in Figure 7.4 the structure of the dataset as various graphs based on different similarities. Blue nodes are image nodes, gray ones are location nodes. The links between them are each time one of the previously defined similarity links.

Once again, the main advantage of a graph-based representation is that existing methods of graph analysis can be leveraged in our image retrieval setting to improve our structuring process. We present next the semi-automatic structuring process that we propose.

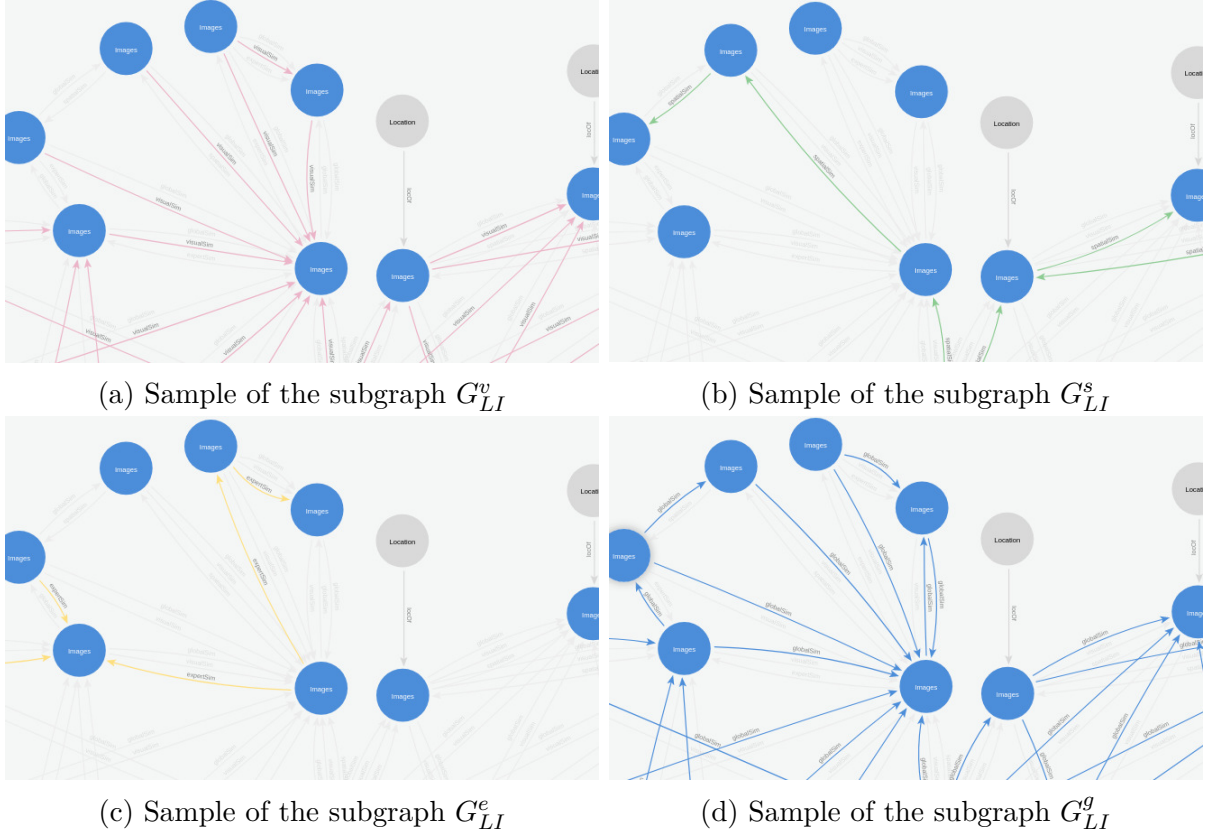


Figure 7.4: Visualization of the subgraphs obtained using different similarities

7.3 Structuring process overview

The dataset’s graph-based representation was first introduced in Section 7.2. This section presents the semi-automatic process we propose to improve the structuring of the dataset, based on the graph paradigm introduced before. We first detail the global process and then focus on the specific location propagation step.

7.3.1 Semi-automatic iterative process

Building on the diffusion-based re-ranking paradigm and the graph-based representation presented in Section 7.2, we propose a semi-automatic process where key links in the dataset’s structure are manually evaluated (created or deleted) and then the updated structure is diffused throughout the whole graph to update links. As it will be shown in Section 7.5, the diffused impact is much greater than the simple impact of the manual modifications.

Structuring is considered as a three-step iterative and semi-automatic process:

1. Automatic building of similarity links, based on CBIR or the approach described in Section 4.3 of Chapter 4 for instance. These links feed the graph-based representation of the global structure at large scale, as explained in Section 7.2;
2. Creation of the graph-based representation based on the previous similarity links;

3. Manual assessment and improvement of some complex configurations automatically highlighted with visual clues in the 3D visualization environment (examples in Section 7.4). The graph-based representation of the collection's structure is updated with these inputs. Then repeat step (1) if necessary.

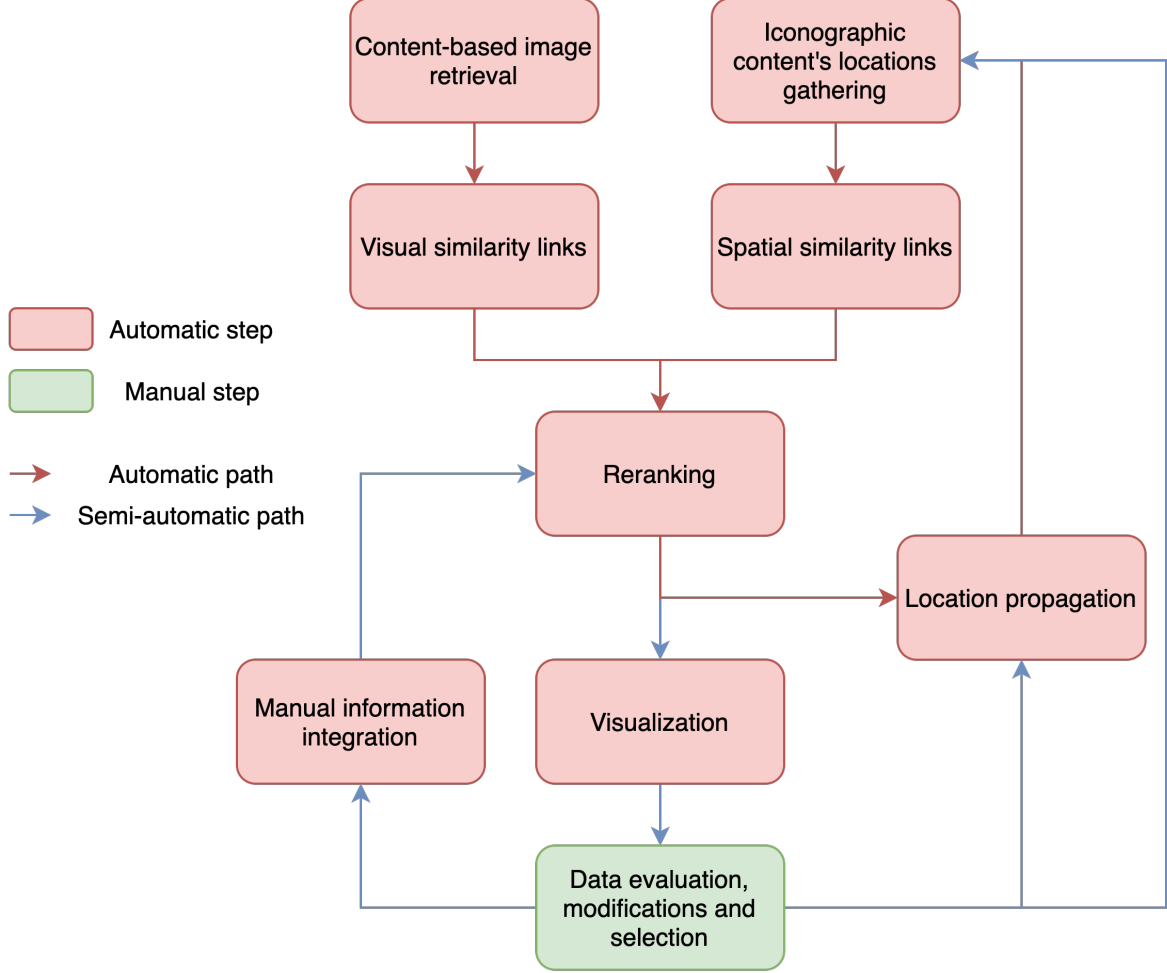


Figure 7.5: Overview of the semi-automatic structuring process

In the process depicted in Figure 7.5, an expert first visualizes the results of a first step of automatic processing. The automatic process produces links between all images based on various information (and ranked based on the global similarity score of Equation 7.1). However, as it will be presented later in Section 7.4, displaying all links is detrimental to the correct visualization and analysis of the proposed structure. Thus, only some links are displayed. It can be the first k links or links ranked between k_1 and k_2 , as desired by the expert. The expert then correct mistakes or add new expert similarities using guides described later in Section 7.4. Those new informations are used to recompute the global similarity score $S_{I,J}$, the images are then re-ranked based on those updated similarity scores. Those ranks and scores are then fed to the automatic diffusion-based re-ranking process or the location propagation process (presented after), running offline. The images are once again re-ranked by those automatic processes and the visualization of the first k links for instance will display new structures to be evaluated and enriched by the expert,

and so on. This approach exploits the best of both worlds, on the one hand automatic processes which handle a large number of data at once and on the other hand an expert knowledge that ensures strong results but is way more time-consuming.

As said earlier, for any image I in the dataset, after any modification, either automatic or manual, the N retrieved images are ranked based on their similarity with I , with the score from Equation 7.1. However, if a similarity link between images I and J is deleted through the visualization platform, J is then ranked last in the ranking list and the images initially ranked after J are moved closer to I by 1. And vice-versa for I in the ranking list of J . Furthermore, $S_{I,J}$ is set to 0, leading all "aggregated" similarity links (visual, spatial or potentially wrong expert ones) to be deleted, with their respective similarities set to 0 too.

7.3.2 Location propagation

Once enough global (or simply spatial in our case) similarity links are established, some information can be propagated through them. Here, we focus on localization: some images can be localized from the locations available with the first similar images retrieved. There exist many techniques for estimating a location from several candidate locations as presented in Section 6.2 of Chapter 6; here we simply choose to average the 2D position of the first candidate locations which are spatially coherent together. To evaluate the spatial coherence, the locations are averaged and for a location to be kept for the propagation, it needs to be closer than 30 meters from the averaged location. Furthermore, the locations are assumed spatially coherent enough for propagation if the mean distance between every possible location and the averaged one is less than 15 meters.

This process can be repeated multiple times until the requirements for propagation are not met anymore (number of linked located images, confidence over the global similarity or the location, etc.). Propagating locations, in order to exploit them for spatial similarity linking and for weighing the automatic retrieval process, proves to be an efficient way to improve the structuring as shown quantitatively in Section 7.5.1 and visually in Section 7.5.2.

7.4 Graph-based visualization platform

The final part of our semi-automatic process is the visualization platform that enables the expert to display the structure of the dataset, to analyze and evaluate it in order to make corrections. This section first presents the requirements we defined for this platform and the most-suited solution we selected. It then focuses on particular visual clues available through the platform in order to help the expert focus on the most impacting structuring issues.

7.4.1 Visualization needs and technical solutions

To visualize the structure of the dataset based on our representation choices, several requirements must be met. We focus here on defining them and then present the platform that meets most of them and that we selected for our process.

7.4.1.1 Main requirements

For our semi-automatic process, the requirements are multiple, both in terms of visualization and in terms of potential interventions by the expert on the structure in an interactive way, which we detail further here.

Visualization paradigms required

The first needs in terms of visualization are quite basic and derive straight from the graph-based structure paradigm chosen and defined in Section 7.2.

The platform must allow to visualize different types of nodes (images and locations) and several types of links at once. The organization of the nodes and links must be automatic, based on different styles of representation the user can choose from.

Furthermore, as the spatial aspect of the structuring is important to us as we work with geographical iconographic heritage, the possibility of spatializing the location nodes would be a major benefit for analyzing and understanding "naturally" the structure of the dataset and the coherence of the various similarity links.

Finally, as our objective is to exploit an iterative semi-automatic process, the native link of the platform with a graph database would ease the back and forth between the visualization and the structuring process offline.

Possible expert actions desired

In order for the expert to intervene on the visualized structure, several actions must be available to him. The two basic ones are the deletion and the creation of links (of any type). Furthermore, as previously presented, additional information could also be added by the expert on the newly created or already existing nodes and links. From an expert similarity score to added information on an image caption, both types of information have to be able to be added to the graph.

The user must also be able to focus on specific types of nodes, of links, to filter based on a specific information or to focus on specific areas of the graph to perform a finer analysis. Furthermore, as the structure is perceived as a graph, algorithms and analysis tools designed for graphs would be a bonus for helping to identify and comprehend the key elements of the structure at a glance. Indeed, modifying the representation of the graph based on several criteria could reveal different and new key elements for which an expert intervention is necessary.

7.4.1.2 GraphXR, the most-suited solution

For visualizing the graph-based structure, we decided to exploit GraphXR¹, a web-based visualization platform dedicated to graph data. We use a free version of the platform that suits our needs in terms of volume of visualization. Combined to a Neo4j graph database, it enables the visualization of the image collections and their link-based structure; see the example of Figure 7.1 at the beginning of the chapter. The platform also allows for spatialization by pinning localized nodes to a map in the 3D environment. That way, our three first main requirements are met. Furthermore, as a web-based platform, the structuring could be performed from anywhere by anyone, as long as the data is made available on a server running the Neo4j database and also making the images available via the web. However, for the current implementation and experiments, we stuck to a local solution.

The different categories of links established (visual similarity, spatial similarity, expert similarity, location linking) produce graphs, which may be connected to the map, otherwise they float in space. The interface allows a user to control the visualization of this representation: focus on one neighborhood, display sub-graphs according to a level of similarity or to a particular link type, access to images and metadata in nodes, etc. They also have the power of updating, removing or adding data manually in nodes and edges of the graphs, creating new edges, as well as launching the update of the whole graph-based representation via automatic retrieval and re-ranking.

Due to the large size of the dataset, the visualization of all types of links at the same time is not an option. Indeed, only the visual similarity links between all images amounts to more than 2 million links, many of them representing very low and meaningless similarities. To prevent the cluttering in the visualization we chose to only display a part of the links at any given time. It can be the k first (most similar) ones, but also could be the links further down the list of similarities, as long as k remains less than 10 (in our case at least).

To further help to obtain a meaningful visualization, visual clues are made available automatically in order to help users focus on more unusual or impacting areas. They can either come from graph analysis or from simple representation choices. Several of them are presented next.

7.4.2 The visualization platform

Relying on GraphXR, we present here in detail the visualization choices and the tools and actions available to the user within the platform.

The platform

The whole visualization platform, displayed in Figure 7.6, provides a 3D visualization area for the graph and/or the map but also multiple tools and visualization solutions.

¹<https://www.kineviz.com/graphxr> (free version)

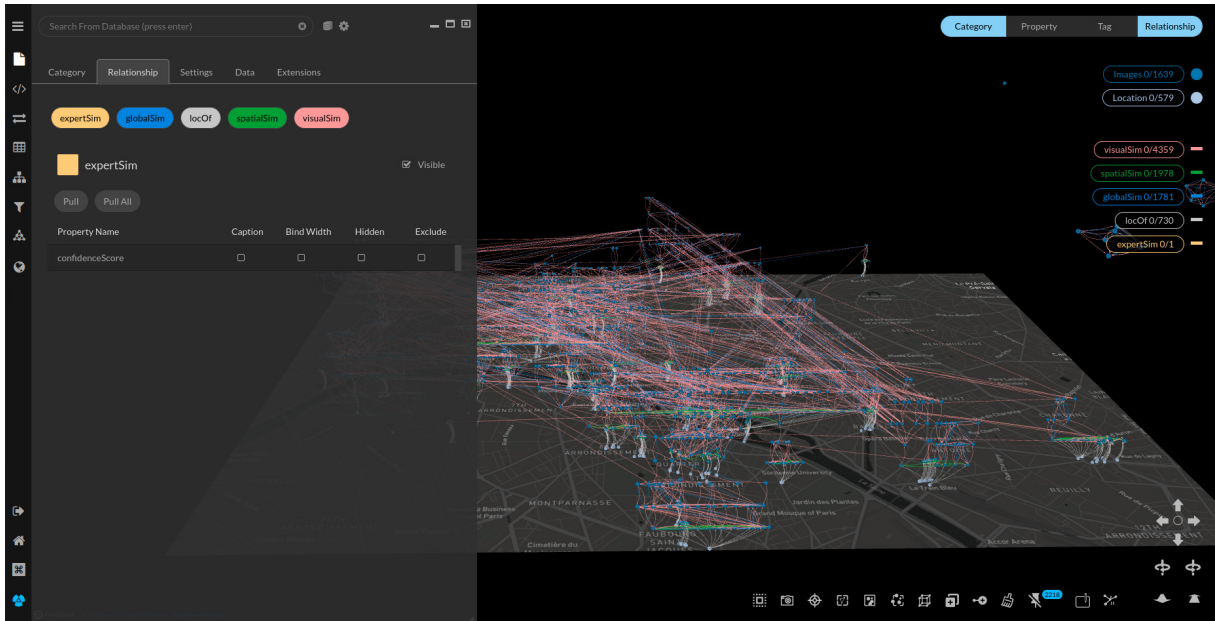


Figure 7.6: Overview of the platform's interface

Several parts of the platform are detailed further.

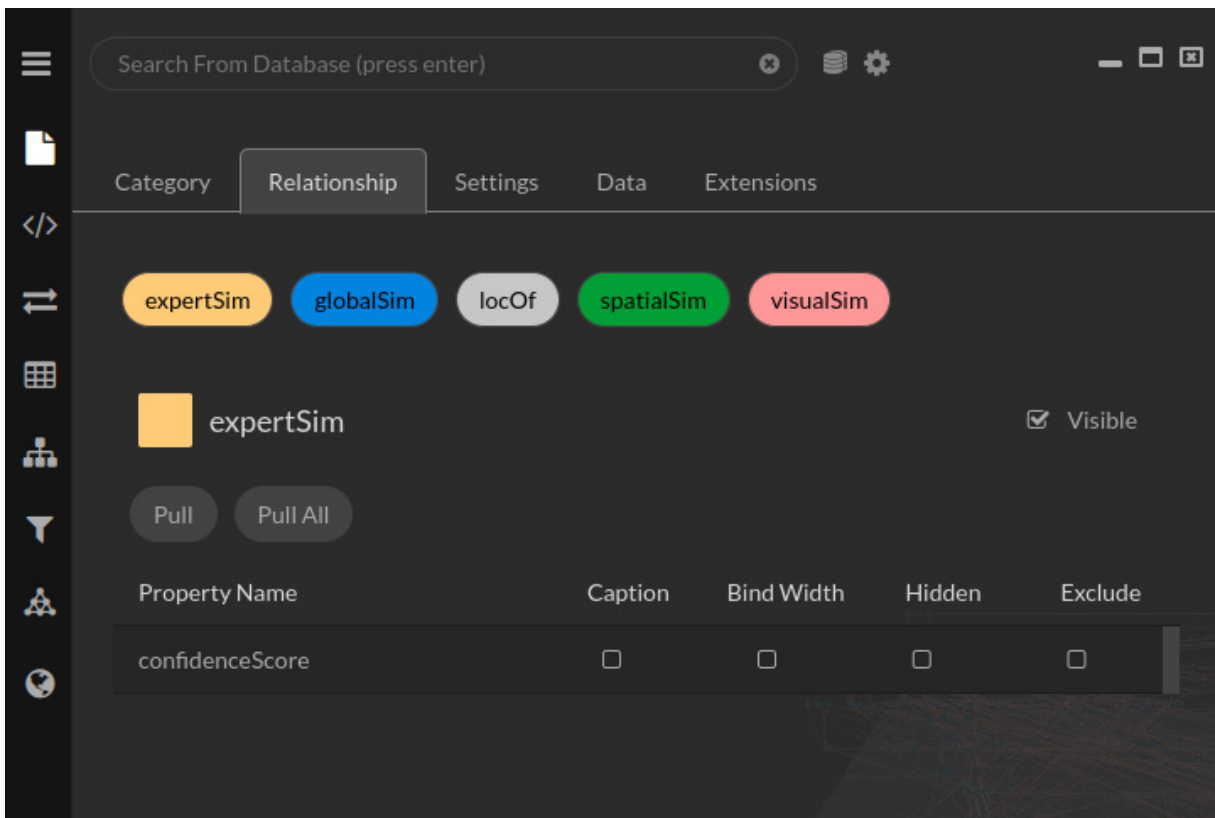


Figure 7.7: Overview of the platform's menu

On the left, as illustrated by Figure 7.7, several menus are accessible. They allow multiple actions of which we list several below:

- loading data and querying both graph and database;
- apply transformations to the graph;

- compute graph algorithms;
- select from multiple layouts;
- filter the displayed data;
- add a map;
- exploit extensions of GraphXR.

On the top right, nodes and links are summarized. The user can select the nodes or links by category but also easily modify their representation, rather than going through the menu. Figure 7.8 illustrates this.

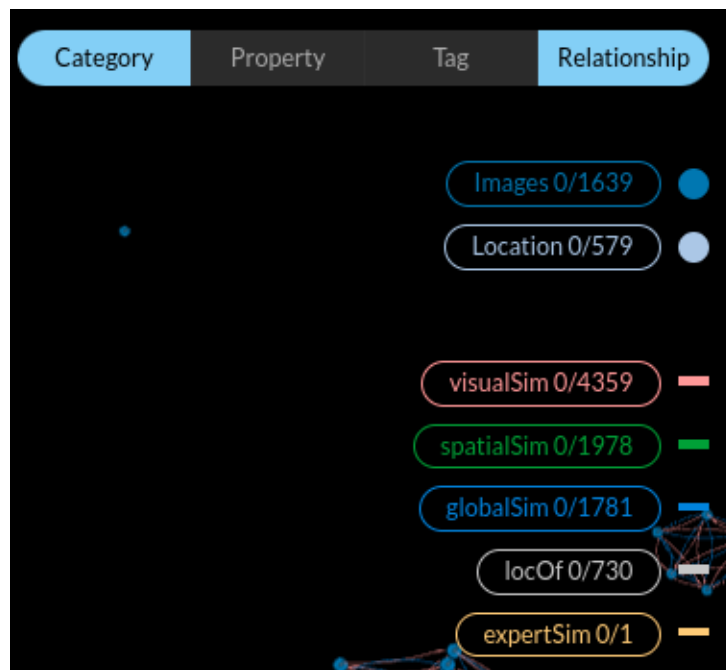


Figure 7.8: Overview of the nodes and links displayed

On the bottom right (see Figure 7.9), several visualization tools are available, to select neighbors of a selected node, to invert the selection or hide the selected nodes for instance.

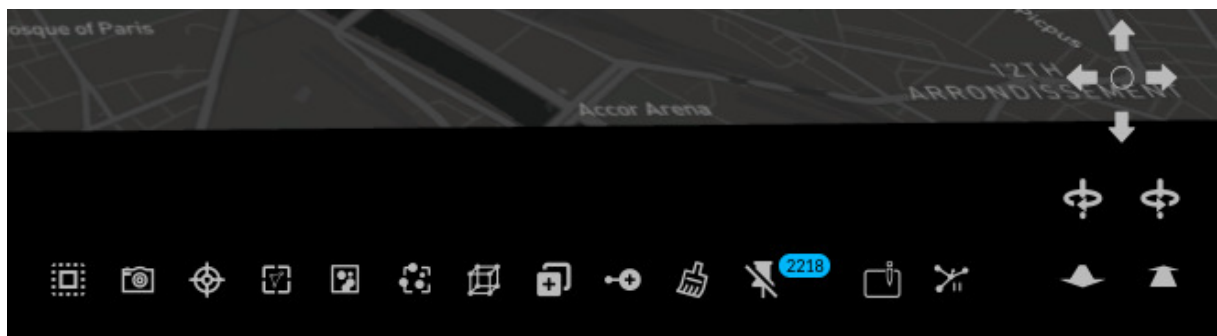


Figure 7.9: Overview of the platform’s tools

Furthermore, a right-click after selecting nodes or links on the graph gives access to many more actions. A simple one is the display of the node's information, as illustrated by

Figure 7.10. Other actions like the creation or deletion of links or nodes are also available for instance.

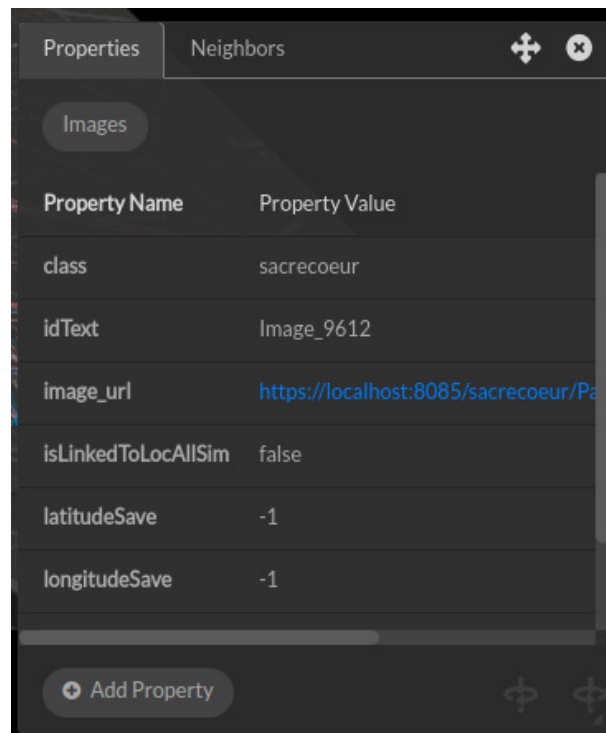


Figure 7.10: Node information display

Finally, one of the most useful aspect of the platform to create a usable platform for any expert is the possibility to create macros with their specific tool, Grove. That is create a button that once clicked will perform successive actions in a specific order. This allows to load the data easily, but also perform algorithms computation or apply specific layouts to the graph. This is illustrated in Figure 7.11 and proves essential to ensure two things. On the one hand, the user does not have to navigate through the different menus and remember the specific order in which to perform the actions. On the other hand, this ensures the fact that the platform can be used by less tech-savvy users (important in GLAMs) but also ensures consistency in the work performed. Indeed, with the same input data, the same macro will get to the same output, which is paramount if several users work jointly on the same data.

To summarize, the platform offers the expert the possibility to perform multiple actions of which we list the most useful below:

- choose the visualization paradigm,
- exploit graph algorithms to automatically compute new information on the graph that can modify the visualization,
- remove an incorrect link,
- create a link of any similarity (visual, spatial or expert),
- add information on a link or a node,
- update the graph database to launch new offline re-ranking steps.

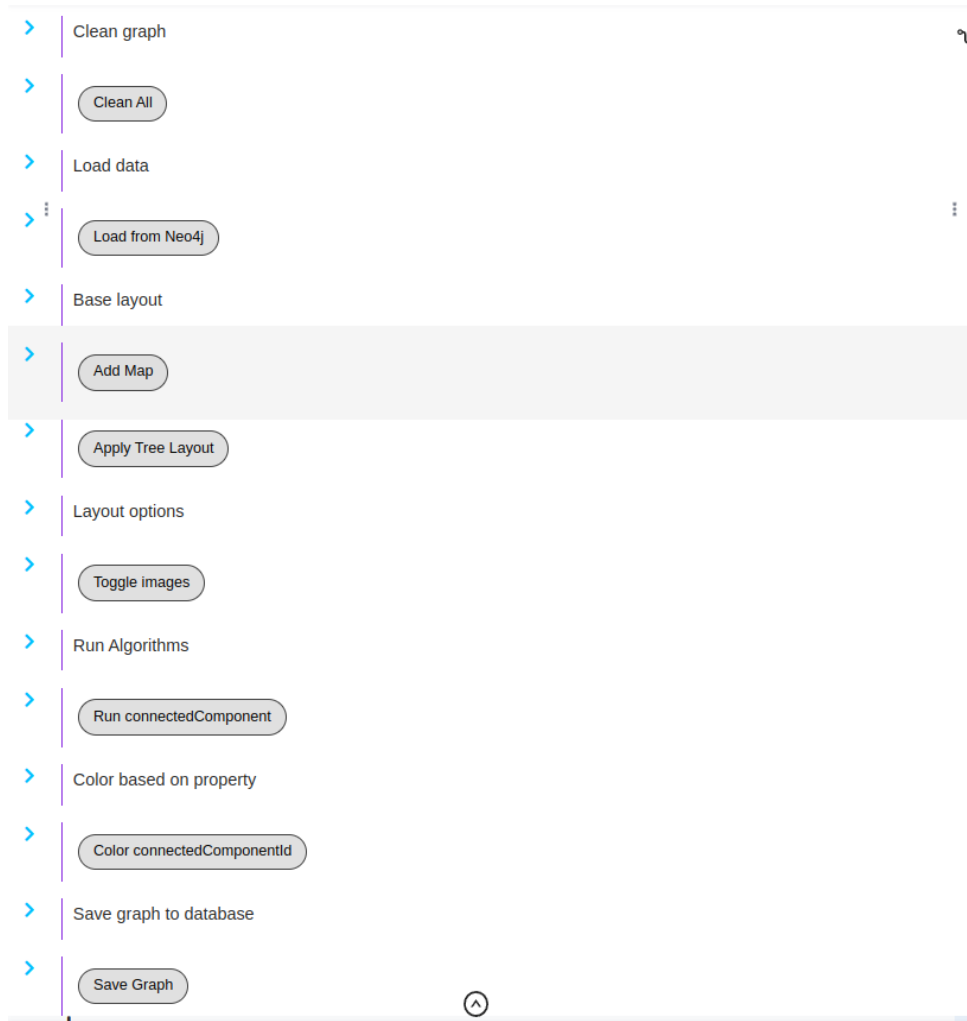


Figure 7.11: Grove extension macro buttons examples

It also allows to create macros to automate several processes in order to speed up the correction process and go smoothly from one visualization to another in a smooth fashion, giving all users, even beginners, the whole range of potential actions.

Visual choices

To ensure consistency in our experiments and illustrations, we made visual choices that we detail here.

The nodes are distinguished by their visual representation. First, the location nodes are always pinned to the map, and in light blue color. For the image nodes, the nodes are never pinned to the map and three potential visualizations are possible:

- in dark blue, the nodes simply indicate their type (that is Images);
- node properties can be displayed in various colors (scaled or not), for instance the community of the node or its betweenness coefficient (see Section 7.4.3);
- the thumbnail of the image that the node represents can also be represented in the node, allowing for a quick check of the scene depicted. Furthermore, a link to visualize the image in full is available in the node's information panel.

The links in turn are distinguished by color, detailed next:

- **Salmon pink** ones represent the **visual** similarity between two images;
- **Green** links means a **spatial** similarity between images;
- **Yellow** similarity links are the **expert** similarity links;
- For **global** similarity links, to further display information on the structuring, we differentiate them in three categories:
 - **Blue** similarity links represent **global, strong, reciprocal** similarity links;
 - **Red** ones represent **global, strong, single-sided** similarities;
 - **Purple** links express a **global, low, single-sided** similarity between images.

The difference between **strong** and **low** similarity links is that strong links are amongst the k links we want to display (as explained before), that is for instance the first k links which are thus probably correct. Low similarity links however represent a lower score, not amongst the top- k links. They are thus more uncertain.

7.4.3 Visual clues as analysis support

The following sections, from Section 7.4.3.1 up to 7.4.3.5, present several visual clues specifically useful for global dataset structuring. From simple visualization to graph algorithms, all are meant to make it easier for the expert to set their focus on impacting areas, thus maximizing the impact of their work while limiting the time cost.

7.4.3.1 From visualizing multiple similarities to a single one

As images can be linked via multiple similarities, the structuring must take all of them into account. That is done in the structuring process detailed in Sections 7.2 and 7.3. However, aggregating similarities also proves beneficial for visualization purposes. Indeed, as illustrated in Figure 7.12 (d), co-visualizing multiple similarities (individually shown in (a), (b) and (c)) leads to a strenuous visualization, not ideal for analysis. Aggregating all similarities in a global link (with the global score introduced before) leads to an uncluttered visualization (shown in (e)), more suited for visual analysis by any user, and at large scale.

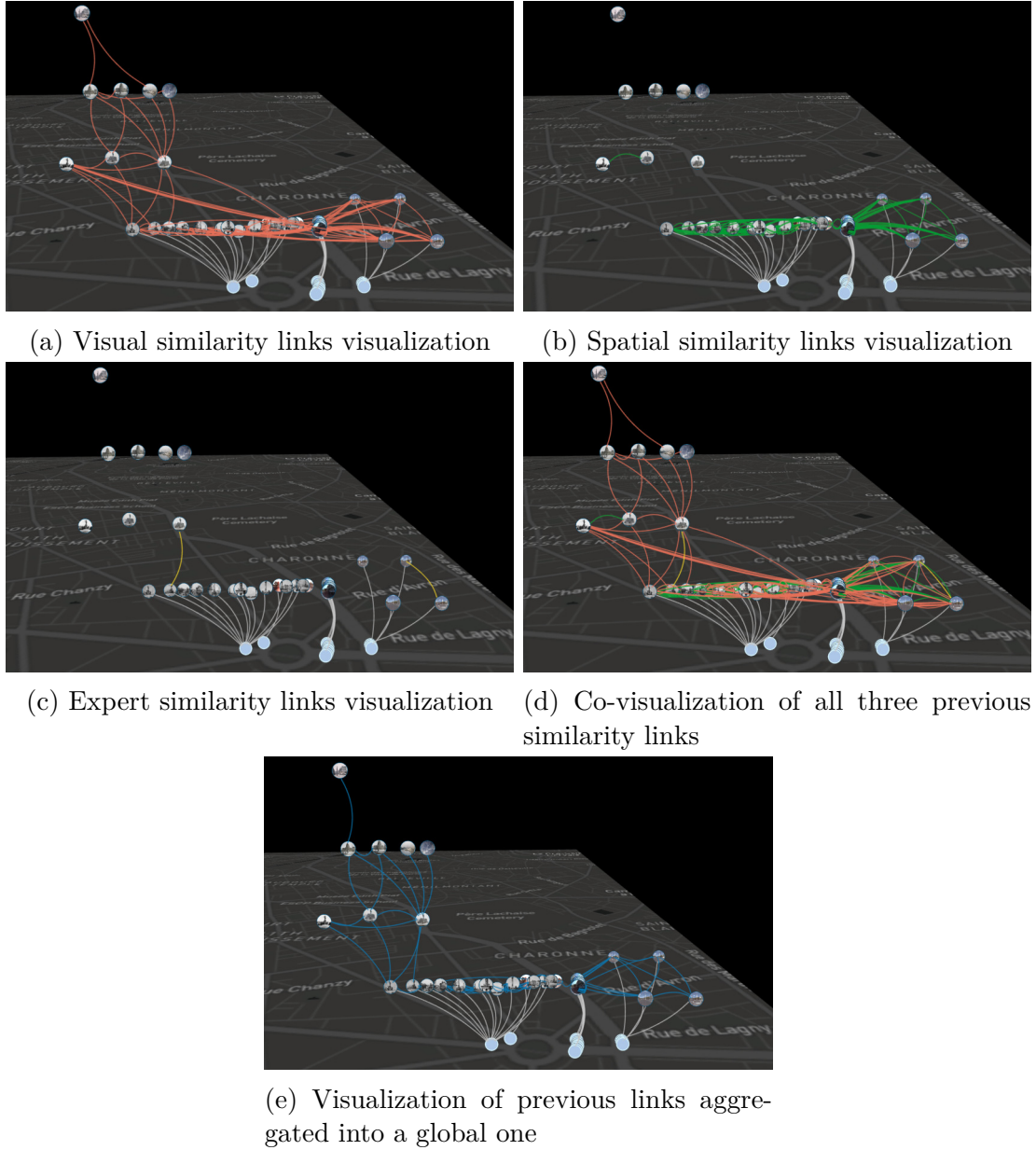


Figure 7.12: Aggregated similarity links visualization

7.4.3.2 Cross-community links

Within graphs, communities can be exhibited based on the different links between nodes. In our case, as we structure a dataset composed of multiple classes, an ideal structure representing a classification in terms of depicted object would be a single community for each class and no links between communities. This "ideal" case never presents itself due to the difficulties automatic methods have with such data (see Part I), however exploiting communities can still prove useful.

Indeed, as said before, links between two different communities are potentially false as similar object should belong together in a community, thus there should be almost no links between communities. Highlighting those links focuses the user's attention on their verification. Once the expert focuses on a specific edge, they can visualize image thumbnails and make the decision to confirm the edge, to simply delete the edge (indicating



(a) Visualization of cross community link



(b) Visual check

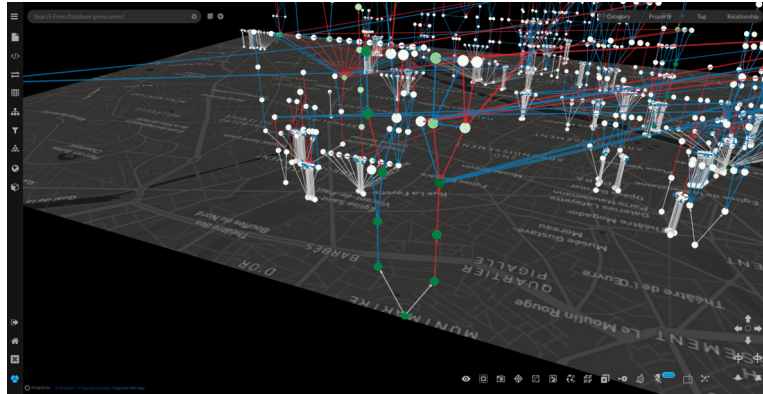
Figure 7.13: Example of cross-community link validation

that it was a false matching) or to create a new edge between the images representing the "expert similarity" they estimate, based on their own criterion.

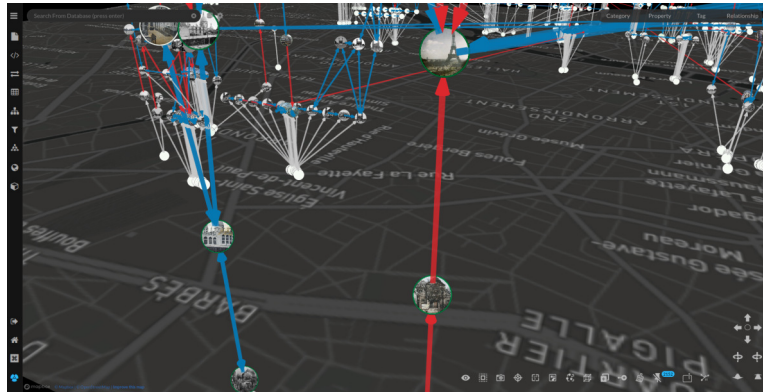
This highlighting process for deleting a link is depicted in Figure 7.13. The first image shows only inter-community links, with one highlighted. The second image shows a zoom on this link with the thumbnails of the images, allowing the expert to evaluate whether to confirm the link with a new expert similarity link or to delete the link. Either creating or deleting links proves beneficial for overall structuring, especially after diffusion, as actions #3-#4 and #6-#7 of Table 7.1 of Section 7.5 illustrate.

In our platform, communities can be exhibited in the graphs by exploiting the global similarity links using the Louvain algorithm (Blondel et al., 2008) or its improvement, the Leiden algorithm (Traag et al., 2019).

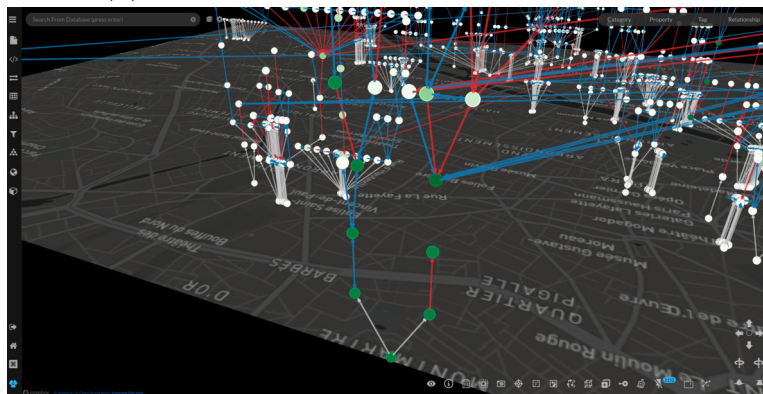
7.4.3.3 Highly central nodes



(a) Identification of highly central nodes (dark green)



(b) Visual check of the visual similarity links



(c) Deletion of the incorrect link



(d) Recomputation of the betweenness coefficient, showing that different communities are no longer wrongly connected

Figure 7.14: Highly central node edge clearing

Exploiting once again the fact that in an ideal setting, no links should exist between communities, thus, problematic links can link an image node to several communities (in this case groups of similar nodes). Those links are problematic because if a single node is linked to multiple communities, then during the diffusion process it creates noise and confuses the re-ranking.

To identify them, the betweenness coefficient is quite efficient as it estimates in how many shortest paths between any two nodes a specific node is. Thus, a high betweenness score reveals a node linking multiple groups of nodes, creating potentially wrong paths for the diffusion process.

In our platform, visualizing for each node its betweenness coefficient, computed with Brandes' algorithm (Brandes and Pich, 2007), helps the user in identifying central nodes linked to multiple clusters/communities. Checking the edges of these central nodes is highly beneficial for global structuring of the dataset. Indeed, clearing the edges of those central nodes, that is deleting or strengthening links with spatial or expert similarities, ensures a clearer frontier between communities.

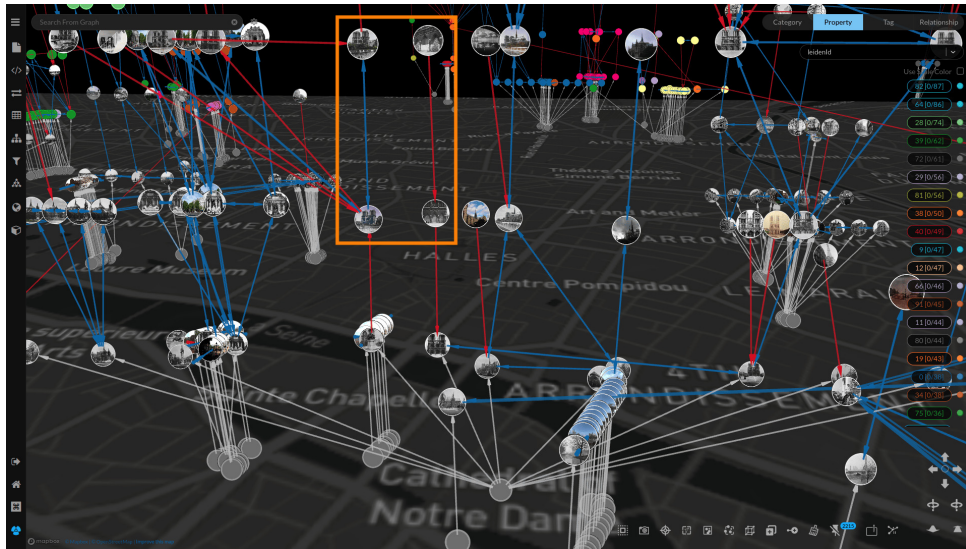
This is illustrated in Figure 7.14 where a chain of nodes seems to link two communities because their betweenness coefficient is high. Cleaning those links removes this high betweenness, indicating an improvement in the structuring. Highlighting nodes based on their centrality coefficient thus focuses the user's intervention on highly impacting evaluations.

7.4.3.4 Spatialized tree representation

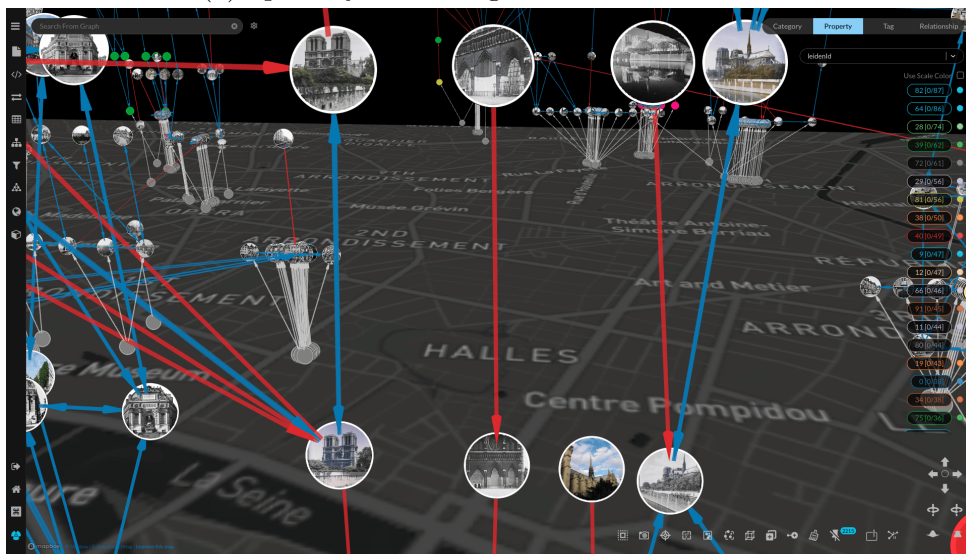
Creating trees based on similarities with location nodes as roots displays naturally together nodes which are spatially close. Indeed, the graph representation will organize nodes in a tree fashion, considering all the similarity links to place the nodes in the scene. This display setting endogenously regroups together non-located nodes not linked to each other but linked to nodes that are spatially close.

It easily allows users to densify the linking between these nodes as finding visually similar or spatially close images supposes to look only at nearby nodes. This is what Figure 7.15 represents. First, not-linked but close nodes are identified (in the yellow rectangle in (a)). Their thumbnails are then visualized (b) and it appears that they depict two parts of the same scene. Thus, there are no visual similarity links (at least not strong ones) but the user can create spatial similarity ones (in green in (c)). Actions #5 and #8 of Table 7.1 (see Section 7.5) show that adding spatial similarities that are then used to improve similarities and overall structuring via diffusion is as efficient as intervening on visual similarities, indicating that both interventions should be used jointly for best performance.

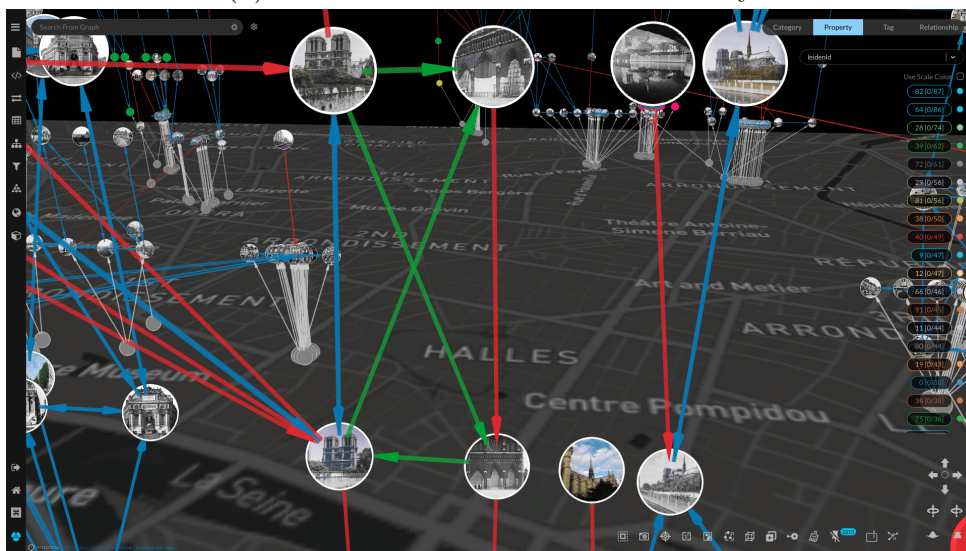
Furthermore, it also helps identifying clusters prime for location propagation as shown in Figure 7.17. That is clusters that can be selected in order to propagate automatically the location information of located images in the cluster to non-located images in the cluster, based on their visual similarity links.



(a) Spatially close images in the visualization



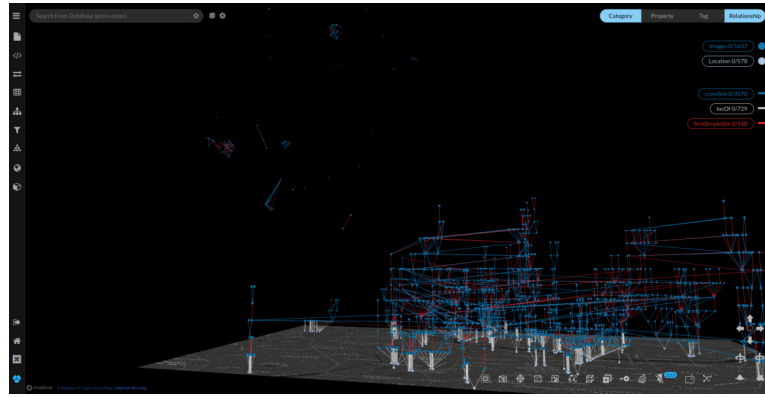
(b) Visual check of their actual similarity



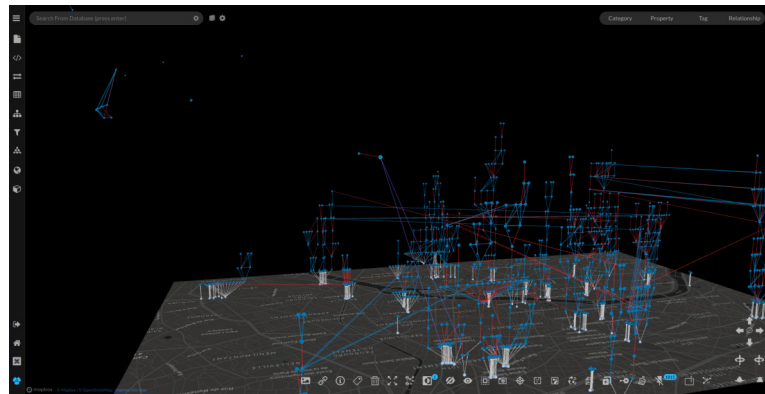
(c) Creation of spatial similarity links (green) for the following automatic diffusion process

Figure 7.15: Spatial similarity links creation process aided by the tree representation

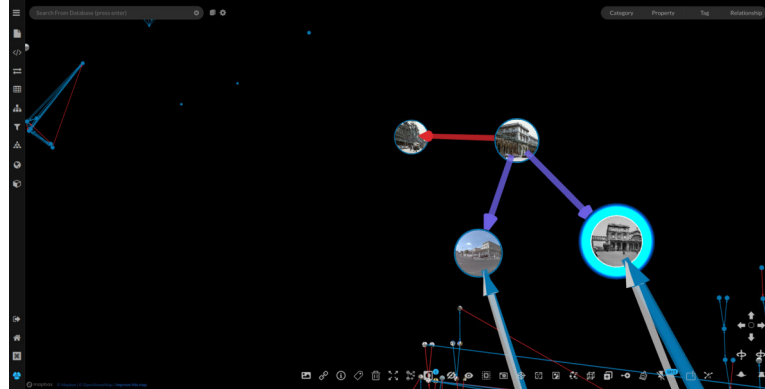
7.4.3.5 Isolated nodes reconnection



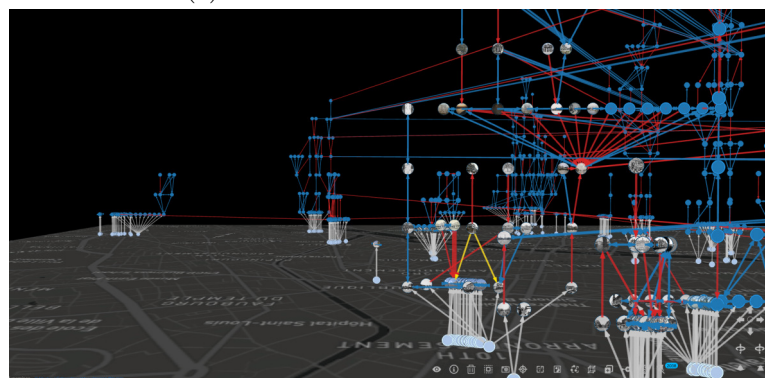
(a) Visualization of floating nodes



(b) Selection of a floating node and visualization of its lower similarity links (purple)



(c) Visual validation of the links



(d) Validation of the links and reconnection of the node to the main graph with expert similarity graphs (yellow)

Figure 7.16: Unconnected nodes' reconnection process

Sometimes, single nodes or small groups of nodes are not connected to the map and to the main graph, they float "in space". This happens when no strong global similarities link them to the main graph. To reattach them to the main graph is complicated for the expert if they do not know which scene the floating nodes represent. To help them, the expert can decide to visualize links of lower global similarity. It at least gives them putative links to evaluate in order to validate them and strongly reconnect the node (and the other nodes of its group) to the map and the main graph using a spatial or an expert similarity link. This process both helps the overall structuring when diffusing the strengthened similarities throughout the graph, it also helps in visualizing the complete dataset in a more organized fashion.

7.5 Semi-automatic structuring process evaluation

After presenting the new graph representation paradigm in Section 7.2, necessary for our semi-automatic structuring process (Section 7.3) in an adapted 3D web-based visualization platform introduced in Section 7.4, this final section evaluates the performance of our proposed process on our dataset.

First, a quantitative evaluation will show the improvements in the structuring when using the semi-automatic process and the visual clues aforementioned for several interventions on the graph. Second, some visual examples will show the evolution of the structuring after each step of intervention and diffusion.

7.5.1 Automatic quantitative evaluation

The impact of the iterative process can be assessed in terms of mAP score (mean Average Precision), previously used to evaluate the automatic retrieval and re-ranking approaches. Table 7.1 illustrates how this score evolves through iterations involving automatic linking (without and with a diffusion step) enriched with automatic and manual inputs (actions #2 to #8 in the Table 7.1).

In this experiment, the number of image nodes is 1637, the number of located images is 537, and the total number of links is around 7,000. The experiments are lead using different starting structures, created using three different sets of links.

First, we start with the structure using Stereopolis locations as a weighting scheme, to remain coherent when using automatic location propagation and subsequent weighting with the new locations. The results of these automatic steps correspond to actions #1 and #2 of Table 7.1.

The second part of our experiments was on the structure provided by the first 5 links of the structure from the previous steps. Those links being the strongest, they have a high probability of being correct, leading to a quite certain structure. Various manual interventions are performed, with results corresponding to actions #3 to #5 in Table 7.1.

Third, we performed manual interventions on the structure created with the 5 to 10 first links of the previous manual structuring steps. Those links may thus be more

Table 7.1: mAP scores evolution through iterations combining automatic and manual linking

Action #	Intervention type	Automation level	Number of added information	mAP before diffusion	mAP after diffusion
1	Image retrieval	Automatic	-	41.97	61.77
2	+ Location propagation	Automatic	85	42.32	62.20
Interventions on the first 5 links					
3	+ Deletions (visual)	Manual	70	42.36	62.32
4	+ Creations (expert)	Manual	30	42.40	62.46
5	+ Creations (spatial)	Manual	33	42.43	62.58
Interventions on the 5th to 10th links					
6	+ Deletions (visual)	Manual	78	42.44	62.59
7	+ Creations (expert)	Manual	26	42.48	63.83
8	+ Creations (spatial)	Manual	27	42.51	64.21

uncertain than the first five, leading to a potentially noisier structure. The results of these interventions correspond to actions #6 to #8 in Table 7.1.

The information added to automatic retrieval (location propagation, targeted manual interventions on similarity links) is quite small, representing each time a volume of about 2% of the total information, while the mAP scores reveal that it notably improves the structuring, with a multiplied impact after the automatic diffusion of this new knowledge.

While the automatic location propagation is quite quick, each of the two manual intervention steps takes about an hour of expert time, accounting for the actual intervention time and the time necessary to save them to the database and run the diffusion process. Several conclusions can be drawn from those results.

The overall improvement (after the three steps) of 2.44% of mAP (from 61.77 to 64.21) is significant and proves the impact of targeted structuring coupled with a diffusion process. The overall mAP score reaches the level of the second-best approach combining a single step of re-ranking before diffusion (R3D-SG) and the fourth score overall.

First, with action #2 of Table 7.1, we observe an improvement of the mAP from 61.77 (using starting locations from Stereopolis only, *c.f.* Section 5.2 of Chapter 5) to 62.20 when adding 85 new propagated locations. The improvement is quite substantial in terms of global structuring but also in terms of added information, 85 more locations represent an increase of 16% in terms of localized images.

Second, the manual interventions improve the mAP in a very limited way (0.19% overall) when not coupled with diffusion. Indeed, the diffusion process brings it up to a 2.01% improvement.

Furthermore, while the first manual interventions on the first 5 links improved the structuring by 0.38%, working on the deeper links improves the mAP by 1.63%. This shows that correcting the deeper and more uncertain structure has much more impact

when diffused afterwards.

Finally, these improvements must be put in perspective to the automatic approaches evaluated in Section 5.3 of Chapter 5. Indeed, the maximum mAP improvement using combined re-ranking approaches is 6.5%. However, it requires around 150 hours of computing to re-rank the first 135 images.

Table 7.2 compares the semi-automatic process with some automatic approaches in terms of number of images re-ranked and computation time.

Table 7.2: mAP improvements comparison against no re-ranking at all, depending on the number of re-ranked images and the computation time

Re-ranking approach	k re-ranked images per query	Computation time (hours)	mAP improvement after diffusion (%)
Automatic approaches			
RANSAC-SG + R3D-SG	135	150	6.5
RANSAC-LG	135	45	6.2
RANSAC-SG/RANSAC-LG	5	1.5	0.01
Location weighting (Sp)	135	1/60	2.5
Semi-automatic process			
Automatic location propagation	-	-	2.9
+ Interventions on the first 5 links	-	1	3.3
+ Interventions on the 5-10 links	-	1	4.9

This comparison shows that the ratio between computation time and mAP improvement is quite favorable for our semi-automatic process.

Indeed, 1 hour of automatic process on the whole dataset, such as RANSAC, does not allow to reach such improvement, as it allows for re-ranking only about 5 images per query, which is very low compared to the 135 we evaluated it on in Chapter 5. Furthermore, re-ranking only 5 images improves the mAP after the diffusion process of only 0.01%, and that using either SuperGlue or LightGlue.

Furthermore, due to the large computation time of more complex re-ranking approaches like the best performing one RANSAC-SG + R3D-SG, in 2 hours, the re-ranking could not be performed for all 1637 images in the dataset.

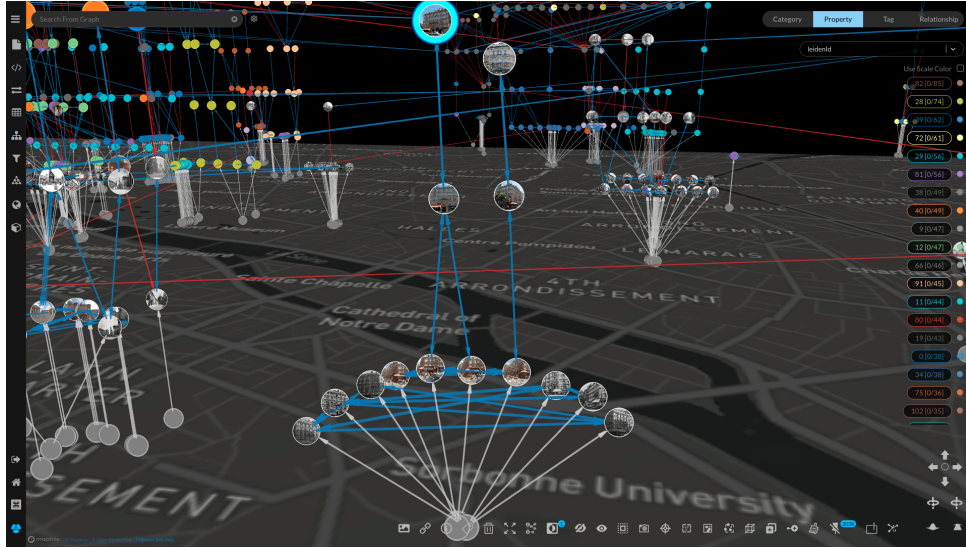
Finally, with our semi-automatic process, the visualization step ensures a correction of the structure that is coherent and validated by the expert, even (and mostly) for complicated cases. On the contrary, automatic methods are somehow black boxes and if we can deduce part of their behaviors, in our case of highly variable contents, some re-ranking may actually be detrimental in some cases, which does not happen with our semi-automatic process.

7.5.2 Qualitative visual evaluation

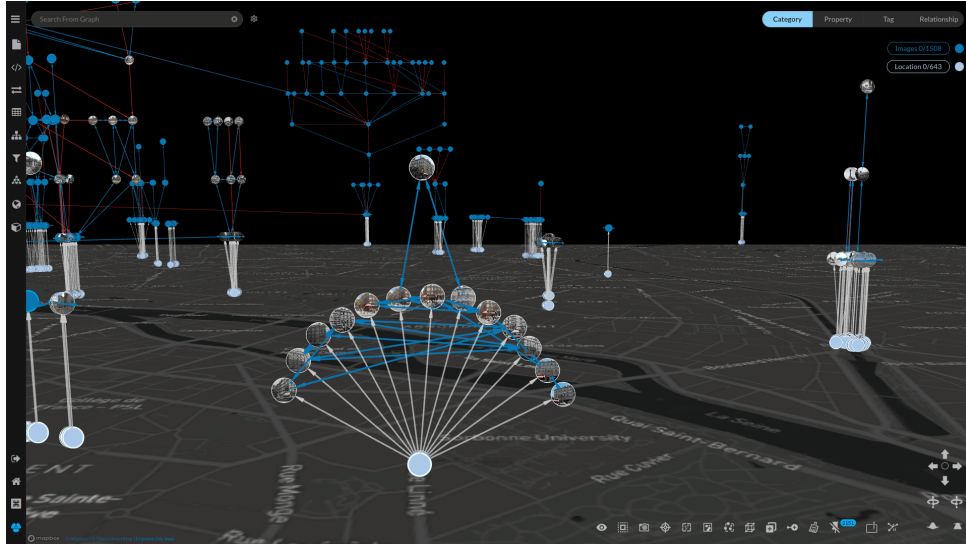
Besides the mAP score, a visual evolution of the structuring can be perceived by the expert as its corrections take place. In addition to the visual examples illustrating the visual clues

already shown in section 7.4.2, here we want to show the global visual structuring that can be displayed within the platform.

First, a visual, local example of location propagation is displayed in Figure 7.17: at first (a), in the cluster of similar images, 10 are located (linked to a location node on the map) and 4 are not. Leveraging the global similarity links between them (in blue), the locations are propagated, resulting (b) in 13 located and 1 non-located images. The location propagation brings structure and visually links images to the map, refining the global visualization.



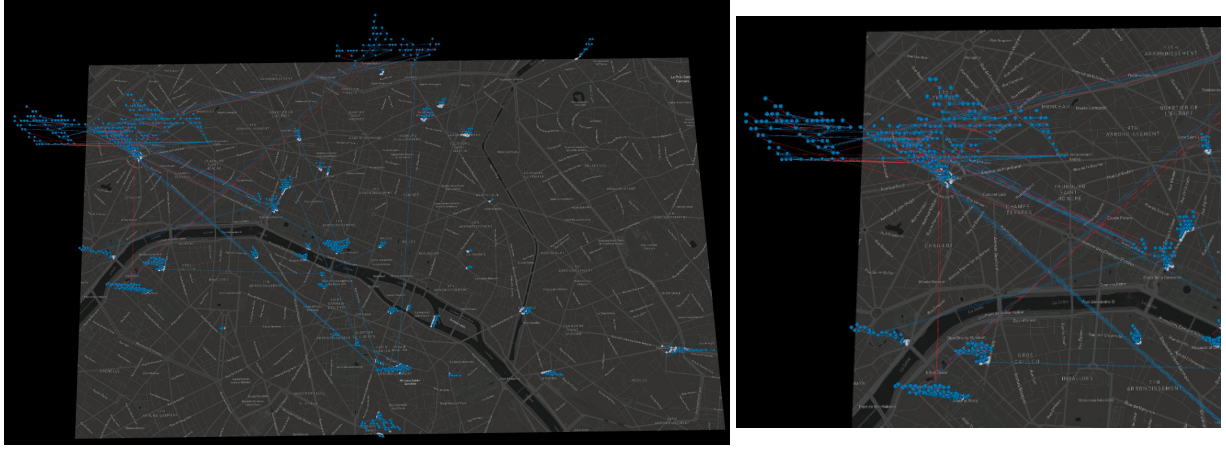
(a) Detection of a cluster of highly connected localized and non-localized images



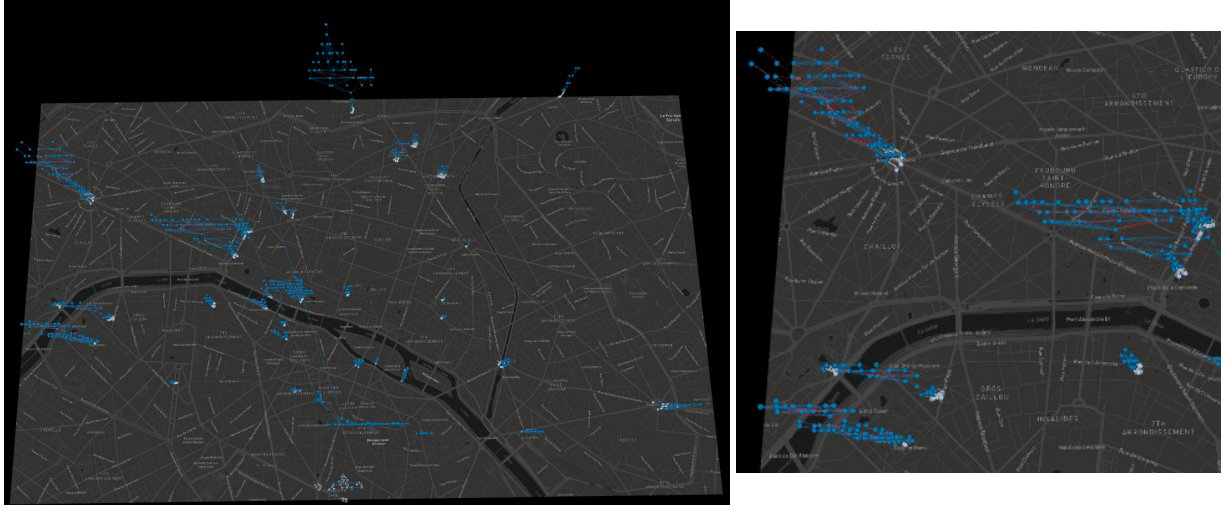
(b) Propagation of the locations to non-localized images

Figure 7.17: Example of the location propagation process

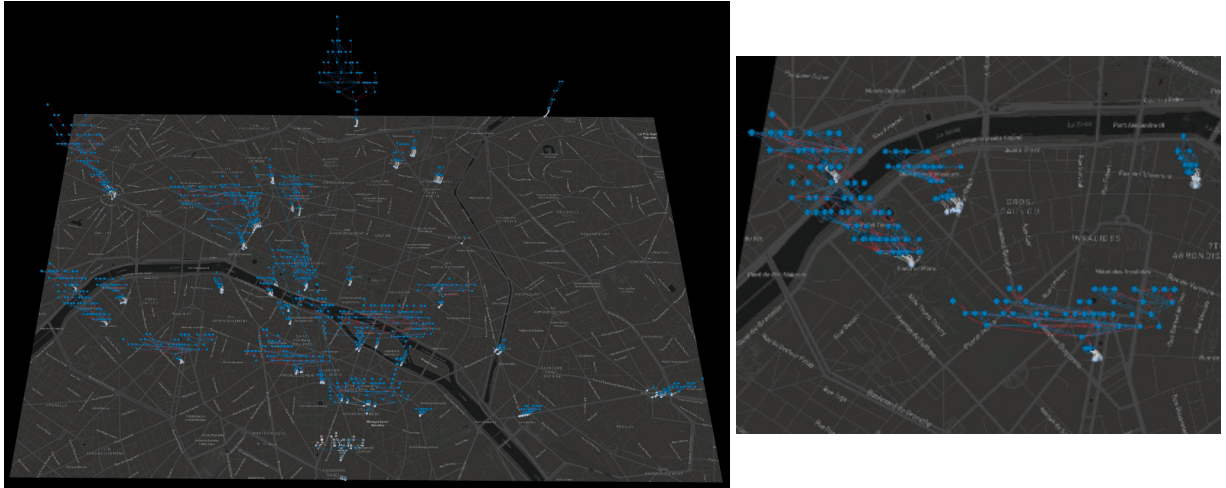
The first Figure 7.18 represents the visual aspect of the dataset's structure during the first steps presented in Table 7.1.



(a) Structure before the first deletion step (action #2), detail on the right



(b) Structure after the first deletion step (action #3), detail on the right



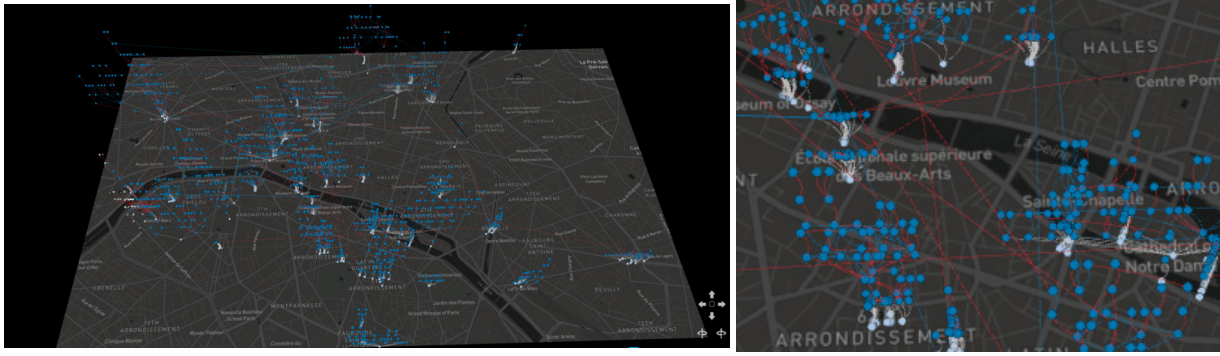
(c) Structure after the two link creation steps (actions #4-#5), detail on the right

Figure 7.18: Visual evaluation of the dataset's structure on the 5 first global links

Visualizing the structure's evolution after each step shows a better structure simply by looking at the amount of long links between images (for instance between the details of (a) and (b)), reflecting links between communities of images that should be spatially distinct.

However, the amount of images linked to the main graph (and the map) remains quite low (between 1218 after the deletion step (b) and 1379 after the final creation step (c) out

of 1637 images in total), showing a better structuring within the first 5 links but limiting the diffusion's impact to find harder images to link, thus limiting the global structuring.



(a) Structure before the second deletion step (action #5), detail on the right



(b) Structure after the second deletion step (action #6), detail on the right



(c) Structure after the two final link creation steps (actions #7-#8), detail on the right

Figure 7.19: Visual evaluation of the dataset structure, using the 5 to 10 best global links

When working with the 5 to 10 strongest global links, the structure is more noisy with more "long links" across the visualization (see Figure 7.19 (a) and especially the red links in the detailed view). However, the deletion limits those problematic links (see Figure 7.19 (b) and the detail on the same area as (b) where no more long red links exist). Finally, after adding new links (expert and spatial), the structure gets noisy again (see Figure 7.19 (c)). The zoom of (c) illustrates that the similar images remain quite correctly organized but new long, potentially wrong links appear, mostly strong single-sided global links (red ones).

However, after each step, the amount of images linked to the main graph increases.

From 1379 in (a) to 1452 in (b) after deleting incorrect links, and finally 1603 out of 1637 after adding new links. This shows that almost all images are linked to the main graph and the map using the 5 to 10 strongest global links. It shows a better global structuring of the dataset and also offers new areas for the expert to intervene on (the "long" and potentially problematic links as shown in the detail of (c)). Indeed, evaluating links within the main graphs proves easier than reattaching floating nodes to the main graphs. It also gives a better global overview of the dataset and its structure.

7.6 Conclusion

In this chapter, we introduced a new semi-automatic structuring process for image collections. Building on our evaluation of automatic retrieval and re-ranking approaches, we propose to exploit targeted manual interventions on an automatically created graph-based structure then diffused throughout the whole graph, exponentially improving the dataset's structuring.

Using a 3D graph-based visualization platform, several visual clues are proposed to help the expert focus its manual interventions on impacting areas of the graph. Jointly exploiting automatically computed similarities and manually corrected or created links, the process iteratively improves the global structure, and using the diffusion process, the impact is much higher than simply that of the few interventions of the expert.

Though costly in expert time, this proposed approach appears to be a solution to the performance ceiling that fully automatic methods seem to reach. Even though they are not able to deal with extremely complicated cases introduced by heritage content, they can use and multiply the impact of a manual input on those extremely difficult cases.

This semi-automatic setting allows to create the most complicated links between contents that automatic approaches can not. Furthermore, after our extensive evaluation in Chapter 5, it has been shown that the diffusion process is most efficient when the entropy of providers is at its maximum. Indeed, it thrives on exploiting links between different collections to further link those collections together. Knowing this, the expert can use visual clues to focus on *a priori* problematic areas but also focus on linking contents inter-collections. This allows for much more control on the subsequent diffusion process.

However, some limitations must be highlighted. First of all, for some very complicated cases, links deeper and deeper would have to be used, and it could prove problematic as the visualized structure could become noisy with too many incorrect links. Indeed, depending on the various contents, the initial automatic linking, either visual or spatial can be very poor, thus preventing the expert to see putative links easily.

For this reason but also the fact that the expert similarity must denote a certain expertise on the dataset to be coherent throughout the various evaluations, it appears for now necessary that the user be acquainted with the data considered, or at least be given strong guidelines.

Furthermore, this proposed process within this platform does not deal with a simple

problem that would need to be solved for this approach to be used at large scale, that is the concurrent work of multiple experts. What to do when two experts set a different expert similarity between contents ? How to exploit information from multiple experts in the updating and diffusion process ? What if two experts disagree ? When to update the visualization ? All those questions would have to be answered for this approach to be used efficiently by collection managers.

Chapter 8

Conclusion and Perspectives

This chapter summarizes the contributions of this thesis and outlines some future perspectives to further our work.

8.1 Contributions

The key takeaways from this thesis are twofold: (i) automatic image retrieval and re-ranking methods can be leveraged and adapted for structuring and interlinking iconographic heritage contents from multiple collections, and (ii) the specificity of this type of data still requires manual intervention to solve the most complicated cases, which can be done in a framework integrating visualization of the structure, manual corrections and relevant re-ranking methods. In this thesis, we make contributions on both of those aspects.

Our **first contributions** focus on automatic retrieval and re-ranking approaches and were developed in Part I. A first, **extensive evaluation of the state-of-the-art** methods for image retrieval and re-ranking led us to identify inadequacies between iconographic heritage content and existing methods and propose three new, more-suited re-ranking approaches. The main idea behind those three approaches is to exploit structuring information to enrich the retrieval and re-ranking. This information can be extracted at query level or at dataset level.

The **two first proposed approaches exploit the geometry of the scene**, extending the geometric verification paradigm, in an effort to alleviate the large changes in viewpoint and level of detail hindering the classical approach. The first one, **R3D, reconstruct a 3D scene** from the first retrieved images and relocates the first retrieved images in this scene to estimate their coherence in the scene. The second one, **R2D, aggregates 2D geometric information** from the first retrieved image to extend the 2D geometry encoded in the query image. This approach can be seen as an approximation of R3D, less costly in computation time while still largely improving classical geometric verification approaches.

The **third proposed approach** does not extract data from the first retrieved images

but rather **leverages a structuring information provided alongside images**, in our case a location information. The spatial proximity between images is used to weigh the visual similarity, estimating that closer images are more likely to depict the same scene than far away ones.

Furthermore, we also evaluated the relevance of combining multiple re-ranking approaches, mainly a first re-ranking step query-wise and another exploiting the manifold of data and exploiting similarities throughout the whole dataset for re-ranking (diffusion-based methods). This led us to realize that in our context of cross-collection interlinking, the first step of re-ranking must not especially be aimed at retrieving many similar images but rather at retrieving similar images from many providers. Indeed, the diffusion step leverages this provider entropy to further retrieve images from multiple providers, overall increasing the retrieval performance.

These automatic approaches used for image linking appear suited to tackle several problems for the collections managers, and more specifically those of iconographic heritage collections like the City of Paris. They are the ones concerned with the digitization process and suffer challenges from both the original manual organization of the data and the new, adapted, digitized structuring. Thus, when they were faced with examples of the potential of automatic approaches, several recurrent uses have emerged. First of all, the detection of duplicates within their collections. It can first be used to clean the collections of digitization artifacts. It can also be used cross-collections to either perform a consistency check on the metadata but also enrich each collection with the metadata from the other (as each collection often has its own metadata due to the use it has of its data). A final use for automatic, visual-based linking is to apply it to images with no metadata available, thus non-exploitable. These images whose metadata were lost can not be used as such as no one can say what they represent. Instead of having to compare these images manually to the whole collection in hope of finding a similar image, those could be set as queries for automatic retrieval against the whole collection, thus giving potential similar images for manual comparison, which would greatly reduce the time necessary to re-identify those images.

Our **final contribution focuses on semi-automatic structuring of heritage image collections** and is developed in Part II. We propose a **framework exploiting a graph representation of a collection’s structuring**. The process for structuring leverages automatic retrieval methods to create an initial structure and to diffuse manual corrections iteratively. The links between images thus represent similarities that can be visual, spatial, or simply defined by an expert. This structure is visualized in a **graph-based web platform and visual clues** are proposed to guide the expert towards highly impacting areas of the graph. Indeed, a few corrections on those areas see their effect multiplied once diffused automatically throughout the structure. This approach proves interesting in two main aspects. First, it allows for a global visualization of the collection in a structured way. Second, it helps unlock difficult structuring challenges that automatic

methods often can not solve.

This proposed framework proved to be interesting for iconographic heritage collection managers (of the City of Paris) and could be helpful for several existing works tackled manually for now. First, as an organized visualization of the data, it allows for a better understanding and analysis of the collection. It could also be used to annotate images in the collection easily, as similar images are displayed close to each other, tags applying to one can easily be added to another. Furthermore, automatic propagation of those annotations could also be considered in this graph paradigm. Finally, the visualization part of the framework could also be considered as an adequate solution for the general public to browse through the collection in its entirety. However, one must remember that either for the general public or for collection managers, the ergonomics and the simplicity of the tools is paramount. The objective is for experts on the data to use the platform for structuring the data, not for experts on the platform to try to organize the data.

8.2 Perspectives

To extend this thesis, some aspects of our research could be developed further, based on insights from this work or recent trends offering new solutions.

8.2.1 Retrieval and reranking

On the first part of our work, several aspects could be improved, either on the retrieval part or the re-ranking one. We detail them further below.

Fine-tuning models for heritage datasets

A first area on which improvements should be sought is the training of models dedicated to heritage contents. This would further the performance of the initial retrieval step, leading to a better overall performance. Several avenues could be explored to this end.

First, retraining networks on iconographic heritage data seem to be a good place to start to further improve the automatic retrieval approaches. Indeed, image descriptors' performance could only be improved with new training on a data resembling more the iconographic heritage we work with. However, as explained before, it could prove difficult to create such a dataset, but several options could be considered.

Our proposed semi-automatic approach could serve as a tool for creating such datasets, starting from some collections and gradually adding new collections while easily creating a ground truth.

To increase the number of images in the training dataset, novel view synthesis approaches could also be explored. This specific data augmentation indeed proved useful when used on recent image data for applications like pose localization for autonomous driving (Moreau et al., 2022). Either NeRF approaches (Mildenhall et al., 2020) or 3D

Gaussian splatting (Kerbl et al., 2023) approaches could help create new, artificial, heritage views to alleviate the sparsity of the real data.

Another avenue to explore is in the model itself, rather than the dataset. Recent works have shown that exploiting geometric information at training time rather than as a re-ranking step trains the image descriptor to differentiate extremely similar images (doppelgängers as they call them in (Cai et al., 2023)). As this can prove to be a challenge for regular, non-monumental images (Parisian facades for instance), and because we have shown that geometric information can afterwards discriminate between them, it could be interesting to exploit this information at training time.

Goal-oriented retrieval

More and more retrieval approaches are designed in a goal-oriented fashion (Pion et al., 2020; Sarlin et al., 2021b; Humenberger et al., 2022). The idea is that to train the retrieval step using an evaluation metric suited to the final task, whether it is a 3D reconstruction task, an inter-collection retrieval task, or any other. In our thesis, the retrieval step was similar for all subsequent tasks, independently of the re-ranking step for instance. It could be pertinent to differentiate several steps of retrieval, whether it is for 3D geometric reconstruction for re-ranking purposes for instance or for collection interlinking. Indeed, we showed that the retrieval results required for both tasks are not exactly the same. Suiting each retrieval for each task could improve both steps independently but also their combination and thus the global performance.

Improving re-ranking

To experiment further on the re-ranking steps, a first lead would be to exploit 3D point clouds, which are now more and more available with some image datasets, such as the Stereopolis one related to mobile mapping, rather than performing a 3D reconstruction with the first images. Novel approaches are quite efficient for the registration of an image into a point cloud (Li and Lee, 2021; Ren et al., 2022), allowing to estimate the geometric coherence of the image in the scene. However, two main challenges arise. First, the point cloud exploited for evaluation should be selected based on the most similar image linked to a 3D point cloud. However, in some cases, due to the variety of heritage contents, the first of such images retrieved is not correct, leading to an incorrect scene for the geometric verification step. Second, those approaches are trained on recent data and would probably need specific fine-tuning (like image descriptor networks) for them to perform on heritage data.

A second lead to follow is based on the performances of the diffusion process. As we have evaluated that the performance of the diffusion for a specific task is based on the initial data (especially the difference between inter and intra-providers retrieval), it could be interesting to evaluate the impact on combining (for instance with late fusion) two

concurrent diffusion steps performed on different initial retrieval results, thus potentially combining both performances on several "tasks" into a better global one.

Finally, evaluating our approaches of retrieval and re-ranking to solve actual linking problems faced by collection managers (metadata consistency, unknown image linking, ...) could help define further improvements based on their efficiency on real test cases.

8.2.2 Semi-automatic structuring platform improvements

The semi-automatic process we proposed remains a proof of concept and the visualization platform could be improved in multiple ways. New visual guides could be proposed. One example could be to visually aggregate densely linked clusters (supposedly correct ones), thus clearing the view for visualizing suspicious links.

Furthermore, more ergonomics could be required to seamlessly go from one visualization paradigm to another, to make it easier for non-expert to apprehend the data without being expert on the platform. Indeed, to consider making it a platform exploited by collections managers, archivists and so on, some work would be required to ensure a smooth use, to make it foolproof and most importantly to ensure concurrent work on the platform, should multiple users work on the same collection at the same time.

Finally, a more advanced platform could be designed, combining the aspects already present with a more advanced 3D visualization, for instance with a 3D representation of the area, rather than just a map. The graph-based representation could thus potentially be visualized jointly with a more immersive visualization of the contents in their context, provided they are given a 6D pose.

8.2.3 Generalization to other types of data

Another general aspect that should be investigated for both parts of our work is the use of other types of data. In our case, we exploit the spatial information of the collection. However, most collections also query and organize their collections based on other annotations which could be leveraged as structuring information. This information could be leveraged during the automatic retrieval and re-ranking process, but also within the visualization platform.

First, in the retrieval step, we have shown that more information is always helpful for structuring. Thus, exploiting textual annotations and their structure would most certainly lead to an improved structuring, provided that the coherence between annotations from different collections is ensured.

Second, within the platform, querying based on annotations could help create new informative visualizations, new visual clues helpful to the user. Furthermore, visualizing these annotations in the platform alongside other types of similarities could also be a solution for ensuring their coherence and even modify them manually directly, those modifications being then potentially diffused throughout the graph of similarities.

To conclude, limiting ourselves to a spatial and visual-based display paradigm suited our dataset but might not be the best for other collections.

Bibliography

Albert Kahn museum. Albert Kahn museum’s collections browsing platform, 2016. URL <https://opendata.hauts-de-seine.fr/explore/dataset/archives-de-la-planete/information/?disjunctive.operateur>.

ALEGORIA project. ALEGORIA research project website, 2018. URL <https://www.alegoria-project.fr/>.

Guoyuan An, Yuchi Huo, and Sung-Eui Yoon. Hypergraph propagation and community selection for objects retrieval. *Advances in Neural Information Processing Systems*, 34:3596–3608, 2021. URL <https://proceedings.neurips.cc/paper/2021/hash/1da546f25222c1ee710cf7e2f7a3ff0c-Abstract.html>.

Guoyuan An, Juhyun Seon, Inkyu An, Yuchi Huo, and Sung-Eui Yoon. Topological verification for image retrieval without fine-tuning. In *Conference on Neural Information Processing Systems*, 2023. doi: 10.48550/ARXIV.2309.05438.

Relja Arandjelovic and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *Conference on Computer Vision and Pattern Recognition*, pages 2911–2918, 2012. doi: 10.1007/978-981-10-0934-1_50.

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In *Transactions on Pattern Analysis and Machine Intelligence*, volume 40, pages 1437–1451, 2018. doi: 10.1109/TPAMI.2017.2711011.

Archival City Project. Archival City research project website, 2019. URL <https://archivalcity.hypotheses.org/>.

Artem Babenko and Victor Lempitsky. Aggregating local deep features for image retrieval. In *International Conference on Computer Vision*, pages 1269–1277, 2015. doi: 10.1109/ICCV.2015.150.

Song Bai, Peng Tang, Philip H S Torr, and Longin Jan Latecki. Re-ranking via metric fusion for object retrieval and person re-identification. In *Conference on Computer Vision and Pattern Recognition*, pages 740–749, 2019. doi: 10.1109/CVPR.2019.00083.

- Vasileios Balntas. SILDa, 2019. URL <https://medium.com/scape-technologies/silda-a-multi-task-dataset-for-evaluating-visual-localization-7fc6c2c56c74>.
- Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=p-BhZSz59o4>.
- Dominik Bartmanski, Henning Füller, Johanna Hoerning, and Gunter Weidenhaus. *Considering Space: A Critical Concept for the Social Sciences*. 2023.
- Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346—359, 2008. doi: 10.1016/J.CVIU.2007.09.014.
- Tim Berners-Lee. Linked Data, 2006. URL <https://www.w3.org/DesignIssues/LinkedData.html>.
- Nicolas Blanc, Timothée Produit, and Jens Ingensand. A semi-automatic tool to georeference historical landscape images. *PeerJ*, 6:1–7, 2018. doi: 10.7287/peerj.preprints.27204.
- Emile Blettery, Paul Lecat, Alexandre Devaux, Valérie Gouet-Brunet, Frédéric Saly-Giocanti, Mathieu Brédif, Laetitia Delavoipière, Sylvaine Conord, and Frédéric Moret. A spatio-temporal web application for the understanding of the formation of the parisian metropolis. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 6(4/W1):45–52, 2020. doi: 10.5194/isprs-annals-VI-4-W1-2020-45-2020.
- Emile Blettery, Nelson Fernandes, and Valérie Gouet-Brunet. How to Spatialize Geographical Iconographic Heritage. In *Proceedings of the 3rd Workshop on Structuring and Understanding of Multimedia HeritAge Contents*, pages 31–40, 2021. doi: 10.1145/3475720.3484444.
- Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
- Eric Brachmann and Carsten Rother. Learning less is more - 6D camera localization via 3D surface regression. In *Conference on Computer Vision and Pattern Recognition*, pages 4654–4662, 2018. doi: 10.1109/CVPR.2018.00489.
- Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. DSAC-differentiable ransac for camera localization. In *Conference on Computer vision and Pattern Recognition*, pages 2492–2500, 2017. doi: 10.1109/CVPR.2017.267.

- Ulrik Brandes and Christian Pich. Centrality estimation in large networks. *International Journal of Bifurcation and Chaos*, 17(07):2303–2318, 2007. doi: 10.1142/S0218127407018403.
- Ruojin Cai, Joseph Tung, Qianqian Wang, Hadar Averbuch-Elor, Bharath Hariharan, and Noah Snavely. Doppelgangers: Learning to disambiguate images of similar structures. In *International Conference on Computer Vision*, 2023.
- Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. BRIEF: Binary robust independent elementary features. In *European Conference on Computer Vision*, volume 6314, pages 778–792, 2010. doi: 10.1007/978-3-642-15561-1_56.
- Bingyi Cao, André Araujo, and Jack Sim. Unifying Deep Local and Global Features for Image Search. In *European Conference on Computer Vision*, volume 12365, pages 726–743, 2020. doi: 10.1007/978-3-030-58565-5_43.
- Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, volume 12346, pages 213–229, 2020. doi: 10.1007/978-3-030-58452-8_13.
- Wei Chen, Yu Liu, Weiping Wang, Erwin Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S. Lew. Deep Image Retrieval: A Survey. 2021. URL <http://arxiv.org/abs/2101.11282>.
- François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Conference on Computer Vision and Pattern Recognition*, volume 2017-Janua, pages 1800–1807, 2017. doi: 10.1109/CVPR.2017.195.
- Ondřej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval. In *International Conference on Computer Vision*, 2007. doi: 10.1109/ICCV.2007.4408891.
- Commission du Vieux Paris. Map locating the collections of the commission du vieux paris, 2023. URL <https://fnp.huma-num.fr/adws/app/515ec27b-90ce-11ec-a660-af5a22dfde2b/>.
- Consortium Inventer le Grand Paris. Plateforme du consortium Inventer le Grand Paris, 2017. URL <http://www.inventerlegrandparis.fr/>.
- Dallas Museum of Art. Dallas Museum of Art website, 2020. URL <https://collections.dma.org/art/collection/>.
- Agni Delvinioti, Hervé Jégou, Laurent Amsaleg, and Michael E Houle. Image retrieval with reciprocal and shared nearest neighbors. In *International Conference*

- on *Computer Vision Theory and Applications*, volume 2, pages 321–328, 2014. doi: 10.5220/0004672303210328.
- Cheng Deng, Erkun Yang, Tongliang Liu, and Dacheng Tao. Two-Stream Deep Hashing with Class-Specific Centers for Supervised Image Search. *Transactions on Neural Networks and Learning Systems*, 31:2189–2201, 2020. doi: 10.1109/TNNLS.2019.2929068.
- Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018. doi: 10.1109/CVPRW.2018.00060.
- Thamotharan Dharani and Laurence Aroquiaraaj. A survey on content based image retrieval. In *International Conference on Pattern Recognition, Informatics and Mobile Engineering*, pages 485–490, 2013. doi: 10.1109/ICPRIME.2013.6496719.
- Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, and Ping Luo. CamNet: Coarse-to-fine retrieval for camera re-localization. In *International Conference on Computer Vision*, pages 2871–2880, 2019a. doi: 10.1109/ICCV.2019.00296.
- Zhengyan Ding, Lei Song, Xiaoteng Zhang, and Zheng Xu. Selective deep ensemble for instance retrieval. *Multimedia Tools and Applications*, 78:5751–5767, 2019b. doi: 10.1007/S11042-018-5967-8.
- Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022. doi: 10.1109/CVPR52688.2022.01181.
- Thomas Dörfler and Eberhard Rothfuss. The geography of the life-world—spatialising the social theory of alfred schütz. *Erdkunde*, (H. 2):149–162, 2023.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. URL <https://openreview.net/forum?id=YicbFdNTTy>.
- Shiv Ram Dubey. A decade survey of content based image retrieval using deep learning. *Transactions on Circuits and Systems for Video Technology*, 32:2687–2704, 2022. doi: 10.1109/TCSVT.2021.3080920.
- Douglas Duhaime. Pixplot visualization platform, 2017. URL <https://dhlabs.yale.edu/projects/pixplot/>.

- Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In *Conference on Computer Vision and Pattern Recognition*, pages 8092–8101, 2019. doi: 10.1109/CVPR.2019.00828.
- Sovann En, Caroline Petitjean, Stephane Nicolas, and Laurent Heutte. A scalable pattern spotting system for historical documents. *Pattern Recognition*, 54:149—161, 2016a. doi: 10.1016/J.PATCOG.2016.01.014.
- Sovann En, Caroline Petitjean, Stéphane Nicolas, Laurent Heutte, and Frédéric Jurie. Region proposal for pattern spotting in historical document images. In *15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 367—372, 2016b. doi: 10.1109/ICFHR.2016.0075.
- Sovann En, Stéphane Nicolas, Caroline Petitjean, Frédéric Jurie, and Laurent Heutte. New public dataset for spotting patterns in medieval document images. *Journal of Electronic Imaging*, 26(1):011010, 2016c. doi: 10.1117/1.JEI.26.1.011010.
- Basura Fernando, Tatiana Tommasi, and Tinne Tuytelaars. Location recognition over large time lags. *Computer Vision and Image Understanding*, 139:21–28, 2015. doi: 10.1016/j.cviu.2015.05.016.
- Martin A Fischler and Robert C Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM*, 24(6):381–395, 1981. doi: 10.1145/358669.358692.
- French Culture Ministry. Base Mémoire, French Culture Ministry’s platform for heritage content, 2019. URL <https://www.pop.culture.gouv.fr/>.
- French Mapping Agency. Remonter le temps, french mapping agency platform for heritage data, 2016. URL <https://remonterletemps.ign.fr/>.
- French National Library. Gallica, french national library website, 2015. URL <https://gallica.bnf.fr/>.
- Andrea Fusiello, Eleonora Maset, and Fabio Crosilla. Reliable Exterior Orientation By a Robust Anisotropic Orthogonal Procrustes Algorithm. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-5/W1 (February):81–86. doi: 10.5194/isprsarchives-xl-5-w1-81-2013.
- Gallicarte project. Arpenteur collaborative platform for locating contents, 2019. URL <https://arpenteur.bnf.fr/>.
- Noa Garcia and George Vogiatzis. Learning non-metric visual similarity for image retrieval. *Image and Vision Computing*, 82:18–25, 2019. doi: 10.1016/j.imavis.2019.01.001.

- Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. doi: 10.1109/CVPR.2012.6248074.
- Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research*, 32(11):1231–1237, 2013. doi: 10.1177/0278364913491297.
- Florent Geniet, Valérie Gouet-Brunet, and Mathieu Brédif. ALEGORIA: Joint multi-modal search and spatial navigation into the geographic iconographic heritage. In *ACM International Conference on Multimedia*, pages 6982–6984, 2022. doi: 10.1145/3503161.3547746.
- Dimitri Gominski, Martyna Poreba, Valérie Gouet-Brunet, and Liming Chen. Challenging deep image descriptors for retrieval in heterogeneous iconographic collections. In *Proceedings of the 1st Workshop on Structuring and Understanding of Multimedia heritAge Contents*, pages 31–38, 2019. doi: 10.1145/3347317.3357246.
- Dimitri Gominski, Valérie Gouet-Brunet, and Liming Chen. Connecting Images through Sources: Exploring Low-Data, Heterogeneous Instance Retrieval. *Remote Sensing*, 13(16):3080, 2021. doi: 10.3390/rs13163080.
- Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. End-to-End Learning of Deep Visual Representations for Image Retrieval. *International Journal of Computer Vision*, 124(2):237–254, 2017. doi: 10.1007/s11263-017-1016-8.
- Albert Gordo, Filip Radenovic, and Tamara Berg. Attention-based query expansion learning. In *European Conference on Computer Vision*, volume 12373, pages 172–188, 2020. doi: 10.1007/978-3-030-58604-1_11.
- Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing XU, and Yunhe Wang. Transformer in transformer. In *Advances in Neural Information Processing Systems*, volume 34, pages 15908–15919, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/854d9fca60b4bd07f9bb215d59ef5561-Paper.pdf.
- Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004. doi: 10.1017/CBO9780511811685.
- Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patchnetvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 14141–14152, 2021. doi: 10.1109/CVPR46437.2021.01392.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

- HistoryPin. HistoryPin collaborative platform, 2010. URL <https://www.historypin.org/en/>.
- Conghui Hu, Can Zhang, and Gim Hee Lee. Unsupervised feature representation learning for domain-generalized cross-domain image retrieval. In *International Conference on Computer Vision*, pages 11016–11025, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Hu_Unsupervised_Feature_Representation_Learning_for_Domain-generalized_Cross-domain_Image_Retrieval_ICCV_2023_paper.pdf.
- Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.
- Zechao Hu and Adrian G.Bors. Conditional Attention for Content-based Image Retrieval. In *British Machine Vision Conference*, volume 0021, pages 1–13, 2020. URL <https://www.bmvc2020-conference.com/assets/papers/0356.pdf>.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.
- Martin Humenberger, Yohann Cabon, Noé Pion, Philippe Weinzaepfel, Donghwan Lee, Nicolas Guérin, Torsten Sattler, and Gabriela Csurka. Investigating the role of image retrieval for visual localization: An exhaustive benchmark. *International Journal of Computer Vision*, 130(7):1811–1836, 2022. doi: 10.48550/ARXIV.2205.15761.
- Martin Humenberger, Gabriela Csurka Khedari, Nicolas Guerin, and Boris Chidlovskii. Methods for visual localization, 2023. URL <https://europe.naverlabs.com/blog/methods-for-visual-localization/>.
- Indian Ocean iconographic heritage network. URL https://ec.europa.eu/regional_policy/en/projects/france/mise-en-reseau-des-patrimoines-iconographiques-de-locean-indien-numeriser-pour-mieux-diffuser.
- Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Teddy Furon, and Ondrej Chum. Efficient diffusion on region manifolds: Recovering small objects with compact CNN representations. In *Conference on Computer Vision and Pattern Recognition*, pages 2077–2086, 2017. doi: 10.1109/CVPR.2017.105.
- Vincent Jaillot, Valentin Rigolle, Sylvie Servigne, John Samuel, and Gilles Gesquière. Integrating multimedia documents and time-evolving 3d city models for web visualization and navigation. *Transactions in GIS*, 25(3):1419–1438, 2021. doi: 10.1111/TGIS.12734.

- Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming Embedding and Weak Geometry Consistency for Large Scale Image Search. In *European Conference on Computer Vision*, 2008. doi: 10.1007/978-3-540-88682-2_24.
- Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sanchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes Aggregating local image descriptors into compact codes. *Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704—1716, 2011. doi: 10.1109/TPAMI.2011.235.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *Transactions on Big Data*, 2019. doi: 10.1109/tbdata.2019.2921572.
- Hee Jae Jun, Byung Soo Ko, Youngjoon Kim, Insik Kim, and Jongtaek Kim. Combination of multiple global descriptors for image retrieval. *arXiv*, 2019. URL <http://arxiv.org/abs/1903.10663>.
- Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In *International Conference on Computer Vision*, volume 2015 Inter, pages 2938–2946, 2015. doi: 10.1109/ICCV.2015.336.
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *Transactions on Graphics*, 42(4), 2023. URL <https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/>.
- Margarita Khokhlova, Nathalie Abadie, Valérie Gouet-Brunet, and Liming Chen. Learning embeddings for cross-time geographic areas represented as graphs. In *the 36th ACM/SIGAPP Symposium On Applied Computing - Technical Track Geographic Information Analysis*, pages 559–568, 2021. doi: 10.1145/3412841.3441936.
- Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. Predicting good features for image geo-localization using per-bundle VLAD. In *International Conference on Computer Vision*, pages 1170–1178, 2015. doi: 10.1109/ICCV.2015.139.
- Wonsik Kim, Bhavya Goyal, Kunal Chawla, Jungmin Lee, and Keunjoo Kwon. Attention-based ensemble for deep metric learning. In *European Conference on Computer Vision*, volume 11205, pages 736—751, 2018. doi: 10.1007/978-3-030-01246-5_45.
- Benjamin Klein and Lior Wolf. Learning query expansion over the nearest neighbor graph. In *British Machine Vision Conference*, 2021. URL <https://www.bmvc2021-virtualconference.com/assets/papers/1048.pdf>.
- Laurent Kneip, Davide Scaramuzza, and Roland Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *Conference on Computer Vision and Pattern Recognition*, pages 2969–2976, 2011. doi: 10.1109/CVPR.2011.5995464.

- Tao Kong, Anbang Yao, Yurong Chen, and Fuchun Sun. HyperNet: Towards accurate region proposal generation and joint object detection. *Conference on Computer Vision and Pattern Recognition*, 2016-Decem:845–853, 2016. doi: 10.1109/CVPR.2016.98.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, pages 1106–1114, 2012. URL <https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html>.
- Wolfgang Krüger. Robust and efficient map-to-image registration with line segments. *Machine Vision and Applications*, pages 38–50, 2001. doi: 10.1007/PL00013267.
- Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation Verification for Image Retrieval. In *Conference on Computer Vision and Pattern Recognition*, pages 5374–5384, 2022. doi: 10.1109/CVPR52688.2022.00530.
- Vincent Lepetit, Francesco Moreno-Noguer, and Pascal Fua. EPnP: An Accurate O (n) solution to the PnP problem. *International Journal of Computer Vision*, 81:155–166, 2009. doi: 10.1007/S11263-008-0152-6.
- Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. BRISK Binary Robust Invariant Scalable Keypoints. In *International Conference on Computer Vision*, pages 2548—2555, 2011. doi: 10.1109/ICCV.2011.6126542.
- Jiaxin Li and Gim Hee Lee. Deepi2p: Image-to-point cloud registration via deep classification. In *Conference on Computer Vision and Pattern Recognition*, pages 15960–15969, 2021. doi: 10.1109/CVPR46437.2021.01570.
- Peng Li, Lirong Han, Xuanwen Tao, Student Member, Xiaoyu Zhang, Christos Grecos, Senior Member, and Antonio Plaza. Hashing Nets for Hashing: A Quantized Deep Learning to Hash Framework for Remote Sensing Image Retrieval. *Transactions on Geoscience and Remote Sensing*, 58(10):7331–7345, 2020. doi: 10.1109/TGRS.2020.2981997.
- Yansheng Li, Jiayi Ma, and Yongjun Zhang. Image retrieval from remote sensing big data: A survey. *Information Fusion*, 67(April 2021):94–115, 2021. doi: 10.1016/j.inffus.2020.10.008.
- Ying Li, Xiangwei Kong, Liang Zheng, and Qi Tian. Exploiting hierarchical activations of neural network for image retrieval. In *ACM Conference on Multimedia Conference*, pages 132–136, 2016. doi: 10.1145/2964284.2967197.
- Wei-Chao Lin. Aggregation of Multiple Pseudo Relevance Feedbacks for Image Search Re-Ranking. *IEEE Access*, 7:147553–147559, 2019. doi: 10.1109/ACCESS.2019.2942142.

- Wei-Chao Lin. Block-based pseudo-relevance feedback for image retrieval. *Journal of Experimental and Theoretical Artificial Intelligence*, 34(5):891–903, 2022. doi: 10.1080/0952813X.2021.1938695.
- Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *International Conference on Computer Vision*, 2023. URL <https://arxiv.org/pdf/2306.13643.pdf>.
- Liris Laboratory Vcity Team. Virtual city project, 2023. URL <https://projet.liris.cnrs.fr/vcity/>.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*, pages 9992–10002, 2021. doi: 10.1109/ICCV48922.2021.00986.
- David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. doi: 10.1023/B:VISI.0000029664.99615.94.
- Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 Year , 1000km : The Oxford RobotCar Dataset. *International Journal of Robotics Research*, 36(1):3–15, 2017. doi: 10.1177/0278364916679498.
- Federico Magliani, Kevin McGuinness, Eva Mohedano, and Andrea Prati. An efficient approximate kNN graph method for diffusion on image retrieval. In *International Conference on Image Analysis and Processing*, pages 537–548, 2019. doi: 10.1007/978-3-030-30645-8_49.
- Ferdinand Maiwald, Jonas Bruschke, Christoph Lehmann, and Florian Niebling. A 4D information system for the exploration of multitemporal images and maps using photogrammetry, web technologies and Vr/Ar. *Virtual Archaeology Review*, 10(21):1–13, 2019. doi: 10.4995/var.2019.11867.
- Ferdinand Maiwald, Christoph Lehmann, and Taras Lazariv. Fully automated pose estimation of historical images in the context of 4d geographic information systems utilizing machine learning methods. *ISPRS International Journal of Geo-Information*, 10(11):748, 2021. doi: 10.3390/IJGI10110748.
- Ferdinand Maiwald, Jonas Bruschke, Danilo Schneider, Markus Wacker, and Florian Niebling. Giving historical photographs a new perspective: Introducing camera orientation parameters as new metadata in a large-scale 4d application. *Remote Sensing*, 15(7):1879, 2023. doi: 10.3390/RS15071879.
- Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2020. doi: 10.1109/TPAMI.2018.2889473.

- MAP laboratory, CNRS. Aioli interactive annotation platform, 2017. URL <http://www.aioli.cloud/>.
- Mapillary. Mapillary visualization platform, 2013. URL <https://www.mapillary.com/app/>.
- Carlo Masone and Barbara Caputo. A Survey on Deep Visual Place Recognition. *IEEE Access*, 9:19516—19547, 2021. doi: 10.1109/ACCESS.2021.3054937.
- Yusuke Matsui, Yusuke Uchida, Hervé Jégou, and Shin’ichi Satoh. A survey of product quantization. *Transactions on Media Technology and Applications*, 6(1):2–10, 2018. URL https://www.jstage.jst.go.jp/article/mta/6/1/6_2/_pdf.
- Christofer Meinecke. Labeling of cultural heritage collections on the intersection of visual analytics and digital humanities. In *7th Workshop on Visualization for the Digital Humanities (VIS4DH)*, pages 19–24, 2022. URL <https://arxiv.org/pdf/2208.13512.pdf>.
- Krystian Mikolajczyk and Cordelia Schmid. Scale & Affine Invariant Interest Point Detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004. doi: 10.1023/B:VISI.0000027790.02288.F2.
- Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 2020. doi: 10.1007/978-3-030-58452-8_24.
- Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. LENS: Localization Enhanced by NeRF Synthesis. In *Conference on Robot Learning*, pages 1347–1356, 2022. URL <https://proceedings.mlr.press/v164/moreau22a.html>.
- Arthur Moreau, Thomas Gilles, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. ImPosing: Implicit Pose Encoding for Efficient Visual Localization. In *Winter Conference on Applications of Computer Vision*, pages 2892–2902, 2023. doi: 10.1109/WACV56688.2023.00291.
- Navigae. Navigae platform, 2018. URL <https://www.navigae.fr/>.
- Navilium. Navilium collaborative platform, 2016. URL <https://www.navilium.com/>.
- Tony Ng, Vassileios Balntas, Yurun Tian, and Krystian Mikolajczyk. SOLAR: Second-Order Loss and Attention for Image Retrieval. In *European Conference on Computer Vision*, volume 12370 LNCS, pages 253–270, 2020. doi: 10.1007/978-3-030-58595-2_16.
- Yue Hei Ng, Fan Yang, and Larry S. Davis. Exploiting local features from deep networks for image retrieval. In *Conference on Computer Vision and Pattern Recognition Workshops*, volume 2015-Octob, pages 53–61, 2015. doi: 10.1109/CVPRW.2015.7301272.

- Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-Scale Image Retrieval with Attentive Deep Local Features. In *International Conference on Computer Vision*, volume 2017-Octob, pages 3476–3485, 2017. doi: 10.1109/ICCV.2017.374.
- Office fédéral de topographie swisstopo. Voyages dans le temps, 2020. URL <https://map.geo.admin.ch>.
- Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. doi: 10.1023/A:1011139631724.
- Eng Jon Ong, Sameed Husain, and Miroslaw Bober. Siamese network of deep fisher-vector descriptors for image retrieval. *arXiv*, 2017. URL <https://arxiv.org/pdf/1702.00338.pdf>.
- Jianbo Ouyang, Hui Wu, Min Wang, Wengang Zhou, and Houqiang Li. Contextual similarity aggregation with self-attention for visual re-ranking. volume 34, pages 3135–3148, 2021. URL <https://proceedings.neurips.cc/paper/2021/hash/18d10dc6e666eab6de9215ae5b3d54df-Abstract.html>.
- Kohei Ozaki and Shuhei Yokoo. Large-scale Landmark Retrieval/Recognition under a Noisy and Diverse Dataset. *arXiv*, 2019. URL <https://arxiv.org/pdf/1906.04087>.
- Evelyn Paiz-Reyes, Mathieu Brédif, and Sidonie Christophe. Cluttering Reduction for Interactive Navigation and Visualization of Historical Images. In *International Cartographic Conference*, volume 4, page 81, 2021. doi: 10.5194/ica-proc-4-81-2021. URL <https://hal.science/hal-03465505>.
- Shanmin Pang, Jin Ma, Jianru Xue, Jihua Zhu, and Vicente Ordonez. Deep Feature Aggregation and Image Re-Ranking With Heat Diffusion for Image Retrieval. *Transactions on Multimedia*, 21(6):1513–1523, 2019. doi: 10.1109/TMM.2018.2876833.
- Nicolas Paparoditis, Jean-Pierre Papelard, Bertrand Cannelle, Alexandre Devaux, Bahman Soheilian, Nicolas David, and Erwann Houzay. Stereopolis {II}: {A} multi-purpose and multi-sensor {3D} mobile mapping system for street visualisation and {3D} metrology. *Revue Française de Photogrammétrie et de Télédétection*, (200):69–79, apr 2014. doi: 10.52638/rfpt.2012.63.
- Nikolaos Passalis and Anastasios Tefas. Neural Bag-of-Features Learning. *Pattern Recognition*, 64:277–294, 2017. doi: 10.1016/J.PATCOG.2016.11.014.
- Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-scale image retrieval with compressed fisher vectors. In *Conference on Computer Vision and Pattern Recognition*, pages 3384–3391, 2010. doi: 10.1109/CVPR.2010.5540009.

- James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Conference on Computer Vision and Pattern Recognition*, 2008. doi: 10.1109/CVPR.2008.4587635.
- Nathan Piasco, Désiré Sidibé, Cédric Demonceaux, and Valérie Gouet-Brunet. A survey on Visual-Based Localization: On the benefit of heterogeneous data. *Pattern Recognition*, 74:90–109, 2018. doi: 10.1016/j.patcog.2017.09.013.
- Noé Pion, Martin Humenberger, Gabriela Csurka, Yohann Cabon, and Torsten Sattler. Benchmarking image retrieval for visual localization. In *International Conference on 3D Vision*, pages 483–494, 2020. doi: 10.1109/3DV50981.2020.00058.
- Agnieszka Podpora. Spatial turn in literary research, analysis and reading practices: Perspectives and limitations. *Topos*, 24(1), 2011.
- Timothée Produit, Nicolas Blanc, Sarah Composto, Jens Ingensand, and Loic Furhoff. Crowdsourcing the georeferencing of historical pictures. In *Free and Open Source Software for Geospatial Conference*, volume 18, page 6, 2018. URL <https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1147&context=foss4g>.
- Pup Ventures, LLC. Whatwasthere platform, 2021. URL <https://www.whatwasthere.com/>.
- Filip Radenovic, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Re-visiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking. In *Conference on Computer Vision and Pattern Recognition*, pages 5706–5715, 2018. doi: 10.1109/CVPR.2018.00598.
- Filip Radenovic, Giorgos Tolias, and Ondrej Chum. Fine-Tuning CNN Image Retrieval with No Human Annotation. *Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2019. doi: 10.1109/TPAMI.2018.2846566.
- Remi Ratajczak, Carlos Fernando Crispim-Junior, Elodie Faure, Beatrice Fervers, and Laure Tougne. Automatic Land Cover Reconstruction from Historical Aerial Images: An Evaluation of Features Extraction and Classification Algorithms. *Transactions on Image Processing*, 28(7):3357–3371, 2019. doi: 10.1109/TIP.2019.2896492.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031.
- Siyu Ren, Yiming Zeng, Junhui Hou, and Xiaodong Chen. Corri2p: Deep image-to-point cloud registration via dense correspondence. *Transactions on Circuits and Systems for Video Technology*, 33(3):1198–1208, 2022. doi: 10.1109/TCSVT.2022.3208859.

- Joseph John Rocchio Jr. Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing*, 1971.
- Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *International Conference on Computer Vision*, pages 2564–2571, 2011. doi: 10.1109/ICCV.2011.6126544.
- John Samuel, Vincent Jaillot, Clément Colin, Diego Vinasco Alvarez, Eric Boix, Sylvie Servigne, and Gilles Gesquière. UD-SV: Urban data services and visualization framework for sharing multidisciplinary research. *Transactions in GIS*, 2023. doi: 10.1111/TGIS.13049.
- Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Conference on Computer Vision and Pattern Recognition*, 2019. doi: 10.1109/CVPR.2019.01300.
- Paul Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020. doi: 10.1109/CVPR42600.2020.00499.
- Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. Back to the Feature : Learning Robust Camera Localization from Pixels to Pose. In *Conference on Computer Vision and Pattern Recognition*, pages 3247–3257, 2021a. doi: 10.1109/CVPR46437.2021.00326.
- Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. Back to the feature: Learning robust camera localization from pixels to pose. In *Conference on Computer Vision and Pattern Recognition*, pages 3247–3257, June 2021b. doi: 10.1109/CVPR46437.2021.00326.
- Torsten Sattler, Chris Sweeney, and Marc Pollefeys. On sampling focal length values to solve the absolute pose problem. In *European Conference on Computer Vision*, pages 828–843, 2014. doi: 10.1007/978-3-319-10593-2_54.
- Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1744–1756, 2016. doi: 10.1109/TPAMI.2016.2611662.
- Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8601–8610, 2018. doi: 10.1109/CVPR.2018.00897.

- Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. Understanding the limitations of cnn-based absolute camera pose regression. In *Conference on Computer Vision and Pattern Recognition*, pages 3302–3312, 2019. doi: 10.1109/CVPR.2019.00342.
- Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. doi: 10.1109/CVPR.2016.445.
- Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, volume 9907, pages 501–518, 2016. doi: 10.1007/978-3-319-46487-9_31.
- Yoli Shavit, Ron Ferens, and Yosi Keller. Learning multi-scene absolute pose regression with transformers. In *International Conference on Computer Vision*, pages 2713–2722, 2021. doi: 10.1109/ICCV48922.2021.00273.
- Xi Shen, Alexei A. Efros, and Mathieu Aubry. Discovering Visual Patterns in Art Collections with Spatially-consistent Feature Learning. In *Conference on Computer Vision and Pattern Recognition*, 2019. doi: 10.1109/CVPR.2019.00950.
- Xi Shen, François Darmon, Alexei A. Efros, and Mathieu Aubry. RANSAC-Flow: Generic Two-Stage Image Alignment. *European Conference on Computer Vision*, 12349 LNCS: 618–637, 2020a. doi: 10.1007/978-3-030-58548-8_36.
- Xi Shen, Ilaria Pastrolin, Oumayma Bounou, Spyros Gidaris, Marc Smith, Olivier Poncet, and Mathieu Aubry. Large-scale historical watermark recognition: dataset and a new consistency-based approach. In *International Conference on Pattern Recognition*, pages 6810–6817, 2020b. doi: 10.1109/ICPR48806.2021.9412762.
- Xi Shen, Yang Xiao, Hu Shell Xu, Othman Sbairi, and Mathieu Aubry. Re-ranking for image retrieval and transductive few-shot classification. In *Advances on Neural Information Processing Systems*, pages 25932–25943, 2021. URL <https://proceedings.neurips.cc/paper/2021/hash/d9fc0cdb67638d50f411432d0d41d0ba-Abstract.html>.
- Yuming Shen, Jie Qin, Jiabin Chen, Mengyang Yu, Li Liu, Fan Zhu, Fumin Shen, and Ling Shao. Auto-Encoding Twin-Bottleneck Hashing. In *Conference on Computer Vision and Pattern Recognition*, pages 2815–2824, 2020c. doi: 10.1109/CVPR42600.2020.00289.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, pages 1–14, 2015. URL <https://arxiv.org/pdf/1409.1556.pdf%E3%80%82>.
- Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, volume 2, pages 1470–1477, 2003. doi: 10.1109/iccv.2003.1238663.

- Jifei Song, Qian Yu, Yi Zhe Song, Tao Xiang, and Timothy M. Hospedales. Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval. In *International Conference on Computer Vision*, volume 2017-Octob, pages 5552–5561, 2017. doi: 10.1109/ICCV.2017.592.
- Yafei Song, Xiaowu Chen, Xiaogang Wang, Yu Zhang, and Jia Li. 6-dof image localization from massive geo-tagged reference images. *Transactions on Multimedia*, 18(8):1542–1554, 2016. doi: 10.1109/TMM.2016.2568743.
- Elena Stumm, Christopher Mei, Simon Lacroix, and Margarita Chli. Location graphs for visual place recognition. In *International Conference on Robotics and Automation*, pages 5475–5480, 2015. doi: 10.1109/ICRA.2015.7139964.
- Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. doi: 10.1109/CVPR46437.2021.00881.
- Shuyang Sun, Jiangmiao Pang, Jianping Shi, Shuai Yi, and Wanli Ouyang. FishNet: A versatile backbone for image, region, and pixel level prediction. In *Advances in Neural Information Processing Systems*, volume 2018-Decem, pages 754–764, 2018. URL <https://proceedings.neurips.cc/paper/2018/hash/75fc093c0ee742f6dddaa13fff98f104-Abstract.html>.
- Swiss Art Research Infrastructure. Images of switzerland online, 2022. URL <https://www.timemachine.eu/images-of-switzerland-online/>.
- Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence*, pages 4278–4284, 2017. doi: 10.1609/AAAI.V31I1.11231.
- Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *International Conference on Computer Vision*, pages 12105–12115, 2021. doi: 10.1109/ICCV48922.2021.01189.
- Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, volume 2019-June, pages 10691–10700, 2019. URL <http://proceedings.mlr.press/v97/tan19a.html>.
- Marvin Teichmann, Andre Araujo, Menglong Zhu, and Jack Sim. Detect-to-retrieve: Efficient regional aggregation for image search. In *Conference on Computer Vision and Pattern Recognition*, pages 5109–5118, 2019. doi: 10.1109/CVPR.2019.00525.
- Giorgos Tolias and Hervé Jégou. Visual query expansion with or without geometry: refining local descriptors by feature aggregation. *Pattern Recognition*, 47(10):3466–3476, 2014. doi: 10.1016/J.PATCOG.2014.04.007.

- Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. Image search with selective match kernels: aggregation across single and multiple images. *International Journal of Computer Vision*, 116:247–261, 2016a. doi: 10.1007/S11263-015-0810-4.
- Giorgos Tolias, Ronan Sifre, and Hervé Jégou. Particular object retrieval with integral max-pooling of CNN activations. In *International Conference on Learning Representations*, 2016b. URL <https://arxiv.org/pdf/1511.05879.pdf%Ef%BC%89>.
- Giorgos Tolias, Tomas Jeníček, and Ondřej Chum. Learning and Aggregating Deep Local Descriptors for Instance-Level Recognition. In *European Conference on Computer Vision*, volume 12346 LNCS, pages 460–477, 2020. doi: 10.1007/978-3-030-58452-8_27.
- Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, 2021a. URL <https://proceedings.mlr.press/v139/touvron21a.html>.
- Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *International Conference on Computer Vision*, pages 32–42, 2021b. doi: 10.1109/ICCV48922.2021.00010.
- Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. *European Conference on Computer Vision*, 2022. doi: 10.1007/978-3-031-20053-3_30.
- Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports (Nature)*, 9(1):5233, 2019.
- Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/a42a596fc71e17828440030074d15e74-Abstract.html>.
- Ignacio Úbeda, Jose M. Saavedra, Stéphane Nicolas, Caroline Petitjean, and Laurent Heutte. Pattern spotting in historical documents using convolutional models. In *International Workshop on Historical Document Imaging and Processing*, pages 60–65, 2019. doi: 10.1145/3352631.3352645.
- Ignacio Úbeda, Jose M. Saavedra, Stéphane Nicolas, Caroline Petitjean, and Laurent Heutte. Improving pattern spotting in historical documents using feature pyramid networks. *Pattern Recognition Letters*, 131:398–404, 2020. doi: 10.1016/j.patrec.2020.02.002.
- Nikolai Ufer, Sabine Lang, and Björn Ommer. Object Retrieval and Localization in Large Art Collections Using Deep Multi-style Feature Fusion and Iterative Voting. In *European Conference on Computer Vision Workshops*, pages 159–176, 2020. URL https://hci.iwr.uni-heidelberg.de/sites/default/files/publications/files/918532137/nufer_visart2020_compressed.pdf.

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- Remco C Veltkamp and Mirela Tanase. Content-based image retrieval systems: A survey. In *State-of-the-Art in Content-Based Image and Video Retrieval [Dagstuhl Seminar]*, 1999. doi: 10.1007/978-94-015-9664-0_5.
- Nicolas Verdier, Eric Mermet, and Carmen Brando. Oronce fine platform, 2017. URL <https://psigehess.hypotheses.org/oronce-fine>.
- Gang Wang, Xiaoliang Sun, Yang Shang, Zi Wang, Zhongchen Shi, and Qifeng Yu. Two-view geometry estimation using ransac with locality preserving constraint. *IEEE Access*, 8:7267–7279, 2020a. doi: 10.1109/ACCESS.2020.2964425.
- Jingdong Wang, Ting Zhang, Jingkuan Song, Nicu Sebe, and Heng Tao Shen. A Survey on Learning to Hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):769–790, 2018. doi: 10.1109/TPAMI.2017.2699960.
- Qi Wang, Weidong Min, Daojing He, Song Zou, Tiemei Huang, Yu Zhang, and Ruikang Liu. Discriminative fine-grained network for vehicle re-identification using two-stage re-ranking. *Science China Information Sciences*, 63:1–12, 2020b. doi: 10.1007/S11432-019-2811-8.
- Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions. In *International Conference on Computer Vision*, pages 568–578, 2021. doi: 10.1109/ICCV48922.2021.00061.
- Li Weng, Valérie Gouet-Brunet, and Bahman Soheilian. Semantic Signatures for Large-scale Visual Localization. *Multimedia Tools and Applications*, 2020. doi: 10.1007/s11042-020-08992-6.
- Tobias Weyand, André Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 A large-scale benchmark for instance-level recognition and retrieval. In *Conference on Computer Vision and Pattern Recognition*, pages 2575–2584, 2020. doi: 10.1109/CVPR42600.2020.00265.
- Kelly L. Wiggers, Alceu S. Britto, Laurent Heutte, Alessandro L. Koerich, and Luiz Eduardo S. Oliveira. Document Image Retrieval Using Deep Features. In *International Joint Conference on Neural Networks*, volume 2018-July, 2018. doi: 10.1109/IJCNN.2018.8489722.

- Kelly L. Wiggers, Alceu S. Britto, Laurent Heutte, Alessandro L. Koerich, and Luiz S. Oliveira. Image retrieval and pattern spotting using siamese neural network. In *International Joint Conference on Neural Networks*, pages 1–8, 2019a. doi: 10.1109/IJCNN.2019.8852197.
- Kelly Lais Wiggers, Alceu de Souza Britto Junior, Alessandro Lameiras Koerich, Laurent Heutte, and Luiz Eduardo Soares de Oliveira. Deep learning approaches for image retrieval and pattern spotting in ancient documents. *arXiv*, 2019b. URL <http://arxiv.org/abs/1907.09404>.
- Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Conference on Computer Vision and Pattern Recognition*, volume 2017-Janua, pages 5987–5995, 2017. doi: 10.1109/CVPR.2017.634.
- Jian Xu, Cunzhao Shi, Chengzuo Qi, Chunheng Wang, and Baihua Xiao. Unsupervised part-based weighting aggregation of deep convolutional features for image retrieval. In *AAAI Conference on Artificial Intelligence*, pages 7436–7443, 2018. doi: 10.1609/AAAI.V32I1.12231.
- Fei Xue, Xin Wu, Shaojun Cai, and Junqiu Wang. Learning multi-view camera relocation with graph neural networks. In *Conference on Computer Vision and Pattern Recognition*, pages 11372–11381, 2020. doi: 10.1109/CVPR42600.2020.01139.
- Heekyoung Yang and Kyungha Min. Classification of basic artistic media based on a deep convolutional approach. *The Visual Computer*, 36(3):559–578, 2020. doi: 10.1007/s00371-019-01641-6.
- Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, and Ping Tan. SANet: Scene agnostic network for camera localization. In *International Conference on Computer Vision*, volume 2019-Octob, pages 42–51, 2019. doi: 10.1109/ICCV.2019.00013.
- Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuotong Xue, Fu Li, Errui Ding, and Jizhou Huang. DOLG: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *International Conference on Computer Vision*, pages 11772–11781, 2021. doi: 10.1109/ICCV48922.2021.01156.
- Guangnan Ye, Dong Liu, I-Hong Jhuo, and Shih-Fu Chang. Robust late fusion with rank minimization. In *Computer Vision and Pattern Recognition*, pages 3021–3028, 2012. doi: 10.1109/CVPR.2012.6248032.
- Deng Yu, Yujie Liu, Yunping Pang, Zongmin Li, and Hua Li. A multi-layer deep fusion convolutional neural network for sketch based image retrieval. *Neurocomputing*, 296: 23–32, 2018. doi: 10.1016/J.NEUCOM.2018.03.031.

- Wei Yu, Kuiyuan Yang, Hongxun Yao, Xiaoshuai Sun, and Pengfei Xu. Exploiting the complementary strengths of multi-layer CNN features for image retrieval. *Neurocomputing*, 237:235–241, 2017. doi: 10.1016/J.NEUCOM.2016.12.002.
- Eyasu Zemene, Yonatan Tariku Tesfaye, Haroon Idrees, Andrea Prati, Marcello Pelillo, and Mubarak Shah. Large-Scale Image Geo-Localization Using Dominant Sets. *Transactions on Pattern Analysis and Machine Intelligence*, 41(1):148–161, 2019. doi: 10.1109/TPAMI.2017.2787132.
- 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. ResNeSt: Split-Attention Networks. In *Conference on Computer Vision and Pattern Recognition Workshops*, pages 2736–2746, 2020a. doi: 10.1109/CVPRW56347.2022.00309.
- Hao Zhang, Xin Chen, Heming Jing, Yingbin Zheng, Yuan Wu, and Cheng Jin. ETR: An Efficient Transformer for Re-Ranking in Visual Place Recognition. In *Winter Conference on Applications of Computer Vision*, pages 5665–5674, January 2023a. doi: 10.1109/WACV56688.2023.00562.
- Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning Two-View Correspondences and Geometry Using Order-Aware Network. pages 5844–5853, 2019. doi: 10.1109/ICCV.2019.00594.
- Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *International Conference on Computer Vision*, pages 2998–3008, 2021. doi: 10.1109/ICCV48922.2021.00299.
- Shaoting Zhang, Ming Yang, Timothee Cour, Kai Yu, and Dimitris N Metaxas. Query specific fusion for image retrieval. In *European Conference on Computer Vision*, pages 660–673, 2012. doi: 10.1007/978-3-642-33709-3_47.
- Xuanmeng Zhang, Minyue Jiang, Zhedong Zheng, Xiao Tan, Errui Ding, and Yi Yang. Understanding Image Retrieval Re-Ranking: A Graph Neural Network Perspective. 2020b. URL <https://arxiv.org/abs/2012.07620>.
- Xulu Zhang, Zhenqun Yang, Hao Tian, Qing Li, and Xiaoyong Wei. Indicative Image Retrieval: Turning Blackbox Learning into Grey. *arXiv preprint*, 2022. URL <https://arxiv.org/abs/2201.11898>.
- Zhongyan Zhang, Lei Wang, Luping Zhou, and Piotr Koniusz. Learning spatial-context-aware global visual feature representation for instance image retrieval. In *International Conference on Computer Vision*, pages 11250–11259, 2023b. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Zhang_Learning_Spatial-context-aware_Global_Visual_Feature_Representation_for_Instance_Image_Retrieval_ICCV_2023_paper.pdf.

- Liang Zheng, Shengjin Wang, Lu Tian, Fei He, Ziqiong Liu, and Qi Tian. Query-adaptive late fusion for image search and person re-identification. In *Conference on Computer Vision and Pattern Recognition*, pages 1741–1750, 2015. doi: 10.1109/CVPR.2015.7298783.
- Liang Zheng, Yali Zhao, Shengjin Wang, Jingdong Wang, and Qi Tian. Good Practice in CNN Feature Transfer. *arXiv*, 2016. URL <http://arxiv.org/abs/1604.00133>.
- Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017. doi: 10.1109/CVPR.2017.389.
- Weixun Zhou, Zhenfeng Shao, Chunyuan Diao, and Qimin Cheng. High-resolution remote-sensing imagery retrieval using sparse features by auto-encoder. *Remote Sensing Letters*, 6(10):775–783, 2015. doi: 10.1080/2150704X.2015.1074756.
- Wengang Zhou, Houqiang Li, and Qi Tian. Recent advance in content-based image retrieval: A literature survey. *arXiv*, 2017. URL <http://arxiv.org/abs/1706.06064>.
- Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 19370–19380, 2023. doi: 10.48550/ARXIV.2304.03410.
- Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning Transferable Architectures for Scalable Image Recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 8697–8710, 2018. doi: 10.1109/CVPR.2018.00907.

Résumé détaillé de la thèse en français

Structuration des fonds iconographiques patrimoniaux : de l'interconnexion automatique à une validation visuelle semi-automatique

3 Chapitre 1 : Introduction

Cette thèse se déroule en parallèle d'une grande entreprise globale de numérisation du patrimoine iconographique contenu dans les archives, musées et bibliothèques. Cette numérisation intervient à la fois pour des soucis de conservation mais aussi dans un contexte d'amélioration de l'accessibilité des données publiques. Cette démarche de numérisation des documents les rend ainsi potentiellement accessibles à de nombreux publics et révèle aussi leur grande diversité. Or, cette diversité, combinée à un manque de structure, notamment entre collections, freine leur mise à disposition et donc leur utilisation à grande échelle.

Pour aider à résoudre ce manque de structure entre collections, de plus en plus de nouvelles méthodes automatiques d'indexation d'images à grande échelle pourraient être envisagées. Or, ces dernières sont entraînées sur des contenus récents et ne s'adaptent donc pas très bien aux contenus patrimoniaux.

Cette thèse a pour objectif tout d'abord d'étudier l'adéquation des méthodes automatiques aux contenus patrimoniaux. L'objectif est ensuite de proposer des solutions pour aider à la structuration des collections patrimoniales, à la fois selon une approche grande échelle et le plus automatiquement possible.

Notre travail va donc se concentrer sur la structuration des collections patrimoniales selon trois objectifs :

- évaluer la pertinence des méthodes automatiques état de l'art ;
- proposer de nouvelles approches pour pallier les difficultés des méthodes existantes ;
- exploiter la structure propre à chaque collection étudiée pour améliorer la structura-

tion globale entre collections.

Par ailleurs, la multiplicité des contenus patrimoniaux étant très grande, nous nous concentrons sur le patrimoine architectural parisien du 20ème siècle.

Notre thèse et nos contributions sont présentées en huit chapitres :

Chapitre 1, introduisant le contexte de notre travail ;

Chapitre 2, qui présente en détail les spécificités inhérentes aux contenus patrimoniaux puis les images que nous avons sélectionnées pour nos expériences ;

Chapitre 3, évaluant les méthodes automatiques de recherche d’images et de ré-ordonnancement, en les appliquant à l’iconographie patrimoniale pour déterminer les approches les plus performantes ;

Chapitre 4, qui expose nos propositions de méthodes de ré-ordonnancement. Ces dernières exploitent une structure plus globale, géométrique ou spatiale pour améliorer la recherche d’images similaires. La diversité visuelle des contenus considérés est telle qu’exploiter plus d’information est bénéfique pour la structuration globale ;

Chapitre 5, présentant tout d’abord l’évaluation des méthodes proposées, puis celle des combinaisons d’approches de ré-ordonnancement. Certaines combinaisons d’approches sont plus performantes. Ce chapitre détaille et explique les raisons de cette variation en termes de performances ;

Chapitre 6, qui passe en revue les approches de spatialisation, de structuration et de visualisation des contenus iconographiques patrimoniaux ;

Chapitre 7, qui présente notre proposition de processus semi-automatique de structuration des collections. Combiné aux approches automatiques précédentes, ce processus s’avère pertinent pour résoudre les cas les plus complexes, qui échappent aux méthodes automatiques. Il se place dans un paradigme de représentation en graphe de la structure des collections. Cela permet à la fois la visualisation et la modification manuelle de cette structure par un expert ;

Chapitre 8, conclusion de cette thèse, résume le travail accompli et propose de futures pistes de travail.

4 Chapitre 2 : Contenus iconographiques patrimoniaux, focus sur Paris

Ce chapitre décrit tout d’abord la diversité des contenus iconographiques patrimoniaux en termes de représentation telles que la couleur, le niveau de détail, la variété de points

de vue, les changements dans le temps, etc. Outre ces représentations multiples, exploiter ces contenus est ardu du fait de leur distribution en silo où chaque collection a son schéma d'organisation particulier. De fait, cela limite l'interconnexion entre les contenus et la structuration globale entre collections, limitant leur accessibilité et donc leur utilisation à grande échelle et par le plus grand nombre. Pallier ces difficultés en termes de structuration est la principale motivation de notre thèse.

Ce chapitre présente ensuite l'ensemble des contenus représentant la ville de Paris tout au long du XXe siècle sur lesquels notre étude porte, que nous détaillons quelque peu ici. Le jeu de données englobe une variété d'images se concentrant sur l'architecture parisienne, à la fois des monuments célèbres et des façades classiques. Il combine des collections issues de diverses bibliothèques, archives et musées, chacune possédant des caractéristiques spécifiques comme l'illustre la Figure 8.1.



Figure 8.1: Exemple d'images issues de notre jeu de données¹

Le jeu de données utilisé dans cette étude comprend deux collections clés : l'une de "La Parisienne de la Photographie", qui a numérisé les collections de l'agence photographique Roger-Viollet, et l'autre de "Stereopolis", un système de cartographie mobile complet ayant capturé en images et en nuage de points 3D les rues de Paris en 2015.

Tout d'abord, le jeu de données de "La Parisienne de la Photographie" s'étend de 1910 à 1979 et couvre un large éventail de l'architecture parisienne. Cette collection est intéressante pour notre thèse du fait de la faible densité en termes de couverture à la fois spatiale et temporelle. Cela rend en effet la recherche d'images similaires plus complexe.

D'autre part, le jeu de données "Stereopolis" fournit une acquisition complète et systématique des rues de Paris en 2015, offrant un jeu de données de référence avec des informations de localisation précises. Ce jeu de données vise à donner de la structure et de la certitude à l'information, en contraste avec les données patrimoniales plus éparses et incertaines.

De plus, le jeu d'images contient des images de six autres fournisseurs, à savoir le jeu de données Paris 6K, la Médiathèque du Patrimoine et de la Photographie, le Musée Albert Kahn, la Cité de l'Architecture et du Patrimoine, la Commission du Vieux Paris et la Conservation des Œuvres d'Art Religieuses et Civiles. Ces collections contribuent à l'aspect multi-fournisseurs du jeu de données, essentiel pour évaluer l'efficacité de la recherche d'images basée sur le contenu face à des images provenant de sources variées.

¹Pour les droits des images, voir la Figure 2.3

Le chapitre fournit un aperçu complet du jeu de données, qui se compose de 1637 images réparties en 31 classes. Les classes représentent des monuments célèbres, des églises parisiennes et des bâtiments classiques de Paris. A cela s'ajoutent des distracteurs, pour un total de 9834 images. Le jeu de données se caractérise par sa diversité visuelle due aux différentes techniques d'acquisition, résolutions et périodes. Il pose de nombreux défis pour les méthodes de recherche d'images automatiques évaluées dans la partie suivante, en faisant ainsi un objet d'étude adapté à nos objectifs de recherche.

Partie I : Recherche d'image par contenu visuel et ré-ordonnancement *a posteriori*

Cette partie se concentre sur les méthodes automatiques de recherche d'images par contenu visuel, qui apparaissent comme une solution au manque d'interconnexion entre les différents contenus. Elle est composée de trois chapitres. Le premier passe en revue et évalue l'état de l'art sur notre jeu de données. Le second propose de nouvelles approches de ré-ordonnancement plus adaptées aux données considérées. Finalement, le dernier chapitre évalue les nouvelles approches proposées et analyse la pertinence des méthodes automatiques pour interconnecter des contenus iconographiques patrimoniaux.

5 Chapitre 3 : Etat de l'art et évaluation

Ce premier chapitre passe en revue l'état de l'art des méthodes de recherche d'image par contenu visuel. Deux grands types de méthodes sont présentés ici. Tout d'abord, celles reposant sur des descripteurs d'image pour la recherche d'images similaires par contenu visuel. Ensuite, celles se concentrant sur le ré-ordonnancement des résultats de la première recherche en ré-ordonnant les images retrouvées selon un autre critère de similarité. Après les avoir passées en revue, ce chapitre évalue leurs performances sur notre jeu de données complexe.

Les descripteurs d'images peuvent être produits selon plusieurs méthodes. Débutant avec les méthodes "fait main", l'état de l'art porte maintenant quasiment exclusivement sur des descripteurs entraînés. Les réseaux utilisés pour extraire les descripteurs sont multiples et nous passons en revue les architectures des réseaux les plus utilisés (cf Tableau 3.1 dans le corps de la thèse). Ces réseaux nécessitent d'être entraînés sur de larges bases d'images, ce qui rend leur adaptation à la donnée patrimoniale plus complexe. Cependant, quelques approches thématiques se concentrant sur des cas très particuliers commencent à se développer.

Une fois que les descripteurs d'images ont été produits, on peut retrouver les images similaires entre-elles en comparant ces descripteurs. Pour affiner encore plus les résultats, une seconde étape, dite de ré-ordonnancement, est très souvent effectuée. Le principe est de ré-ordonner la liste d'images similaires selon un autre critère de similarité (autre

descripteur, cohérence géométrique, etc.). De nombreuses méthodes existent, comme indiqué dans le schéma de la Figure 8.2.

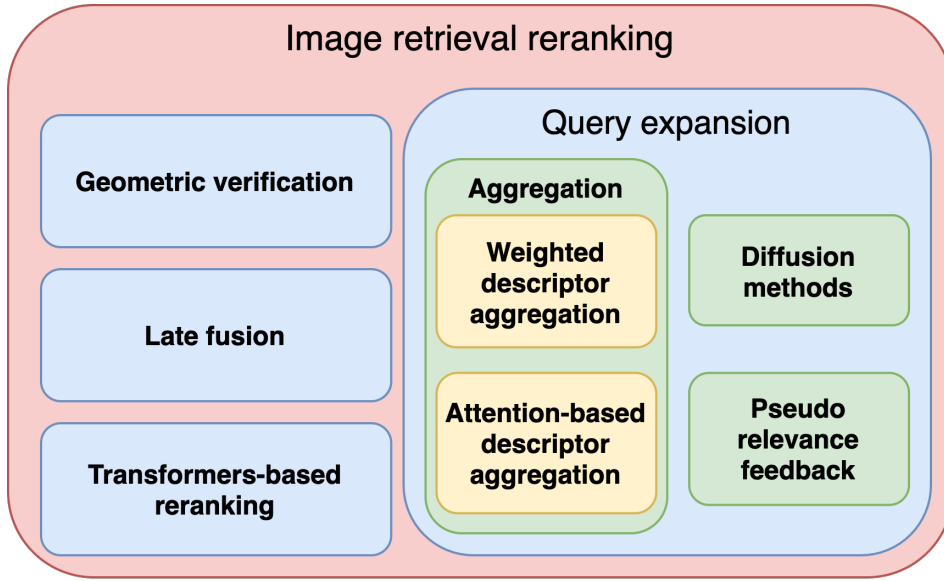


Figure 8.2: Différentes méthodes de ré-ordonnement *a posteriori* des images

Une fois les approches possibles (description d'images et ré-ordonnement) passées en revue, les plus adaptées *a priori* sont évaluées sur notre jeu de données pour déterminer leur performances respectives.

Tout d'abord, certains descripteurs d'images sont évalués. Tous sont des descripteurs entraînés et tous sont des descripteurs globaux sauf How + ASMK. Le Tableau 8.1 présente les scores selon différents cadres d'évaluation (taille du jeu de données, DB_{small} et DB_{large} et présence d'images distracteurs, $DB_{large+dist}$).

Table 8.1: Scores de mAP des descripteurs évalués

	DB_{small}	DB_{large}	$DB_{large+dist}$
DELG (Cao et al., 2020)	53.2	-	-
R101 - GeM (He et al., 2016; Radenovic et al., 2019)	57.9	53.3	38.5
How + ASMK (Tolias et al., 2016a, 2020)	53.8	55.1	41.0
CV-Net global (Lee et al., 2022)	-	67.3	37.1

Outre cette évaluation globale, une évaluation plus fine a montré que How+ASMK est le descripteur d'image le plus adapté. Il est performant à la fois face à l'hétérogénéité visuelle des contenus mais aussi pour répondre à nos objectifs d'interconnexion entre collections. C'est le descripteur d'image que l'on conserve pour la suite de notre étude, en l'appelant How-A.

Les méthodes de ré-ordonnement *a posteriori* sont ensuite évaluées, les résultats sont dans le Tableau 8.2.

Table 8.2: Modification du score de mAP par ré-ordonnancement

Approche	Ordre de grandeur
Weighted descriptor aggregation (Chum et al., 2007; Radenovic et al., 2019)	+ 0.1
Pseudo relevance feedback (Lin, 2019)	< + 0.5
Transformers-based : CSA (Ouyang et al., 2021), RRT (Tan et al., 2021)	- 10
Geometric Verification : RANSAC (DeTone et al., 2018; Sarlin et al., 2020) CV-Net Rerank (Lee et al., 2022)	+ 0.5-1.5 - 2
Diffusion (Shen et al., 2021; Zhang et al., 2020b)	+ 16

Les différentes méthodes testées n’apportent pas le même bénéfice une fois appliquées à nos données. En effet, l’hétérogénéité visuelle des contenus empêche les méthodes de ré-ordonnancement sus-mentionnées de performer correctement. Ainsi, certaines ont un très faible impact et d’autres sont même préjudiciables à la recherche d’images. Il faut noter toutefois la performance importante des approches de diffusion qui exploite les résultats de recherche d’images de manière globale, à l’échelle du jeu de données entier. Cette étude nous permet donc d’analyser ce qui fonctionne et ce qui manque à ces méthodes et permet de proposer de nouvelles approches dans le chapitre suivant.

6 Chapitre 4 : Contributions pour le ré-ordonnement des images

S’appuyant sur l’analyse des différentes méthodes évaluées précédemment, ce chapitre propose trois nouvelles approches de ré-ordonnement *a posteriori* des images. Pour chacune d’entre elles, l’idée est d’exploiter des informations à une échelle plus globale que l’image requête elle-même. Dans notre thèse nous exploitons des informations spatiales dans deux contextes différents.

Tout d’abord, deux méthodes se placent dans un cadre d’expansion de requête géométrique, étendant le paradigme classique de la vérification géométrique. L’idée de nos deux premières propositions est d’exploiter les informations géométriques collectées dans les premières images retrouvées par similarité visuelle avec l’image requête. La vérification géométrique s’appuie ensuite sur une géométrie de la scène plus globale, agrégeant l’information de la requête et celle des images similaires.

La première méthode proposée, R3D, reconstruit une scène 3D (voir Figure 8.3) à

partir de l'image requête et des images les plus similaires visuellement. La cohérence géométrique des images est ensuite évaluée par rapport à cette scène 3D globale plutôt que par rapport à l'information 2D uniquement contenue par l'image requête.

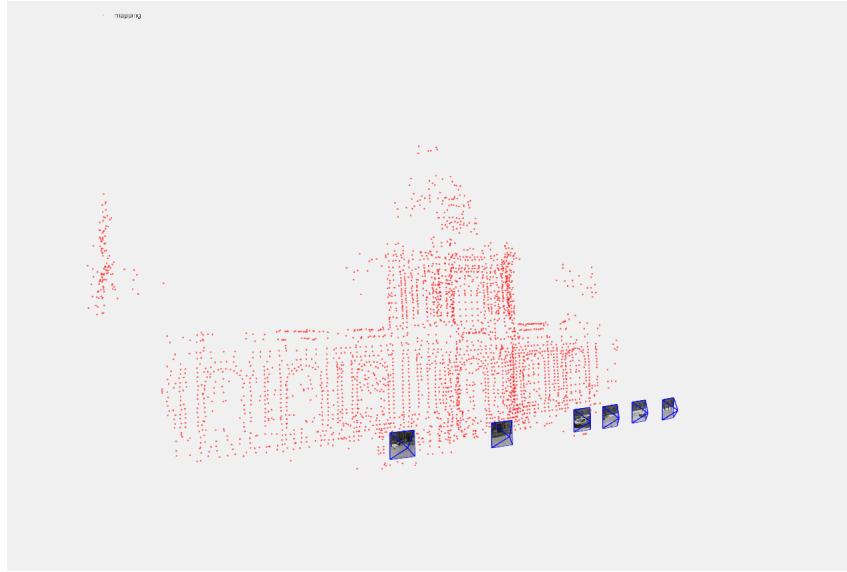


Figure 8.3: Une scène 3D reconstruite

Pour simplifier cette approche par reconstruction 3D qui s'avère couteuse en temps de calcul, nous proposons une version 2D (R2D) qui agrège l'information géométrique 2D des premières images retrouvées dans l'espace de la requête (Figure 8.4). Cette aggrégation étend l'information géométrique à une scène plus globale, à la manière de la reconstruction 3D mais sans le coût en temps important de la reconstruction 3D.

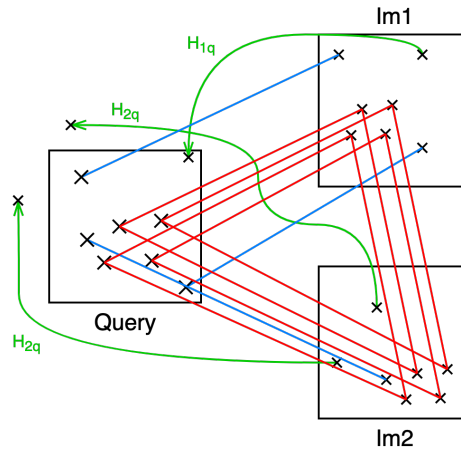


Figure 8.4: Exemple d'extension de la géométrie 2D de l'image requête à partir des images similaires, les points rouges sont des points importants dans la scène, les bleus dépendent de l'image requête et les points verts sont reprojetés depuis les images similaires

Pour finir, notre troisième approche exploite également une information spatiale globale. Cette fois-ci, on utilise les informations de localisation partiellement disponibles avec les images. L'idée est d'évaluer la cohérence de la similarité visuelle calculée automatiquement (et sensible à l'hétérogénéité visuelle des contenus). Pour cela, elle est pondérée par

un score de proximité géographique entre les images. L'idée est que les images éloignées sont moins susceptibles de représenter la même scène que des images proches, comme l'illustre la Figure 8.5. Si cette méthode exploite des informations spatiales dans notre cas, elle pourrait utiliser n'importe quelle autre information de structuration associée aux images pour pondérer la similarité visuelle.

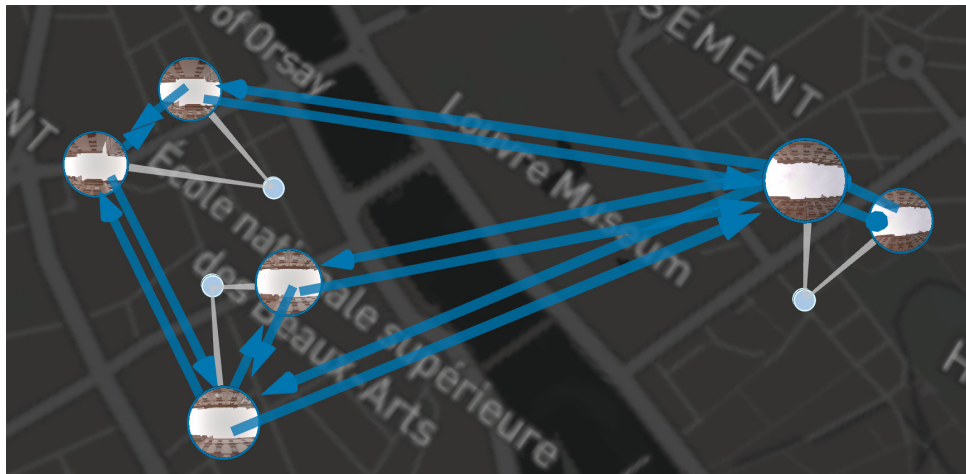


Figure 8.5: Contradiction entre les liens de similarité visuelle (flèches bleues) entre images et la localisation des images sur la carte (points bleus ciels)

7 Chapitre 5 : Evaluation des stratégies de ré-ordonnement

Ce chapitre présente d'abord l'évaluation des méthodes proposées dans le chapitre précédent. Ces résultats sont résumés dans le Tableau 8.3.

Table 8.3: Résumé des méthodes de ré-ordonnement proposées

Descripteur + Méthode de ré-ordonnement	mAP
How-A	41.0
How-A + RANSAC-SG	41.5
How-A + RANSAC-LG	41.9
How-A + R3D-SG	44.4
How-A + R3D-LG	44.2
How-A + R2D-SG	36.2
How-A + R2D-LG	41.9
How-A + location weighting (Sp)	42.0
How-A + location weighting (No dist)	40.5
How-A + location weighting (All)	42.5

Toutes les méthodes proposées améliorent le résultat de mAP sauf R2D-SG. En effet, pour cette dernière, le descripteur de points d'intérêts locaux utilisé (SuperGlue) ne performe pas suffisamment bien pour que la scène étendue géométriquement soit suffisamment précise pour permettre une vérification géométrique efficace.

Si les approches avec ré-ordonnancement sont un peu plus performantes que la simple recherche d’images, ce chapitre montre ensuite que le plus efficace pour améliorer les résultats *a posteriori* est de combiner plusieurs méthodes de ré-ordonnancement de manière successive. Ainsi, nous combinons différentes méthodes de ré-ordonnancement s’appuyant sur de l’information spatiale avec une méthode de diffusion (GNN-R) qui s’avérait très performante dans le chapitre 3. Les résultats de ces combinaisons sont présentés dans le Tableau 8.4.

Table 8.4: Scores de mAP pour différentes combinaisons de méthodes de ré-ordonnancement. En couleurs sont indiqués les **premiers**, **deuxièmes** et **troisièmes** meilleurs résultats de chaque colonne. Le meilleur score est indiqué en **gras**.

Descriptor + Méthode de ré-ordonnancement	Diffusion après ré-ordonnancement préalable				Computation time
	Sans GNN-R	GNN-R \times 1	GNN-R \times 2	GNN-R \times 3	
How-A	41.0	57.2	59.3	57.0	
How-A + RANSAC-SG	41.5	57.2	59.3	57.0	+120s
How-A + RANSAC-LG	41.9	61.2	65.5	63.3	+100s
How-A + R3D-SG	44.4	61.9	64.2	61.9	+220s
How-A + R3D-LG	43.2	61.1	63.2	60.7	+210s
How-A + R2D-SG	36.2	59.6	62.9	60.5	+150s
How-A + R2D-LG	41.9	61.0	64.4	62.1	+140s
How-A + location weighting (Sp)	42.0	58.9	61.8	59.5	+1/30s
How-A + location weighting (No dist)	40.5	57.8	61.1	59.0	+1/30s
How-A + location weighting (All)	42.5	60.2	63.1	61.8	+1/30s
How-A + RANSAC-SG + R3D-SG	44.9	62.9	65.8	63.3	+340s
How-A + RANSAC-LG + R3D-LG	43.0	61.8	64.1	61.9	+300s
How-A + RANSAC-SG + R2D-SG	36.9	60.1	63.0	60.5	+270s
How-A + RANSAC-LG + R2D-LG	41.7	61.2	64.3	62.2	+240s
How-A + location weighting (Sp) + R3D-SG	44.7	62.4	64.9	62.4	+220s
How-A + location weighting (Sp) + R2D-LG	41.9	61.1	64.7	62.1	+140s

L’enchaînement de multiples étapes de ré-ordonnancement s’avère efficace pour améliorer les résultats de la recherche d’image avec une amélioration maximale de près de 26% de mAP. Cependant, cela s’accompagne d’un temps de calcul très long.

Par ailleurs, une étude plus poussée des performances montre que l’étape de diffusion finale performe idéalement si deux critères sont conjointement maximisés. D’une part, un **résultat de mAP initial correct** (indiquant avoir correctement retrouvé des images similaires) et d’autre part dans notre cadre d’interconnexion des collections, une **entropie maximale des fournisseurs** présents dans les premiers résultats. En effet, la diffusion exploite le graphe des images les plus similaires. Dès lors, comme la recherche d’image a tendance naturellement à performer intra-fournisseur (où l’hétérogénéité visuelle est réduite), si tous les fournisseurs sont présents dans les premiers résultats, la diffusion va pouvoir ramener des résultats issus de tous les fournisseurs. Ainsi, la recherche d’images tant inter-fournisseurs que globale est améliorée.

Cela explique aussi pourquoi nos méthodes d’expansion de requête géométrique proposées performant mieux que la vérification géométrique classique. Elles proposent une scène globale qui englobe les multiples points de vues, permettant aux images de fournisseurs différents (et donc aux caractéristiques visuelles différentes) de s’y rattacher.

Cependant, cet enchainement de méthodes de ré-ordonnancement permet de réaliser que les cas les plus compliqués semblent toujours hors d'atteinte pour les méthodes intégralement automatiques qui semblent plafonner. C'est ce qui nous pousse à proposer dans la seconde partie de la thèse un processus semi-automatique combinant méthodes automatiques et interventions manuelles dans une plateforme de visualisation. Ce processus permet de résoudre les cas les plus complexes de similarité entre images puis de diffuser ces informations dans le graphe de similarité pour améliorer globalement la recherche d'images.

Partie II : Structuration semi-automatique, approche basée sur le graphe

La première partie de la thèse s'est intéressée à la performance des méthodes automatiques de recherche d'image (et de ré-ordonnancement *a posteriori*) et propose des méthodes plus adaptées aux contenus iconographiques patrimoniaux. Cependant, cette partie révèle qu'adapter les méthodes automatiques et exploiter la meilleure combinaison possible atteint un plafond en termes de performance, ne parvenant pas à résoudre les cas les plus compliqués que seul un expert peut résoudre manuellement. Cette seconde partie propose donc un processus semi-automatique de structuration des collections iconographiques. Considérant la collection d'image et les liens de similarité entre images comme un graphe, un expert peut intervenir sur la structure du graphe dans une plateforme visuelle. Ses modifications sont ensuite diffusées automatiquement pour améliorer plus globalement la structure du graphe de similarités. Cette partie se compose de deux chapitres, le premier passe en revue tout d'abord les méthodes de spatialisation des contenus iconographiques et ensuite les plateformes de structuration et visualisation des contenus. Le second chapitre détaille le processus semi-automatique proposé et évalue sa pertinence pour la structuration des collections iconographiques.

8 Chapitre 6 : Etat de l'art en structuration, spatialisation et visualisation du patrimoine iconographique

Ce chapitre s'intéresse aux potentielles méthodes de structuration des collections d'images. Nous nous concentrons tout d'abord sur les méthodes de spatialisation des contenus. Elles sont devenues très importantes pour organiser, requêter et visualiser les contenus, améliorant ainsi leur accessibilité. La seconde partie présente les approches de structuration des contenus qui sont ensuite exploitées par des plateformes de visualisation.

La première partie du chapitre passe en revue la question de la spatialisation des contenus iconographiques. Cette spatialisation peut se faire de diverses manières, selon différents référentiels et peut aboutir à différentes informations de localisation.

Tout d’abord, l’information de localisation obtenue peut être en 2D (position sur une carte) ou en 3D (position sur un globe). Mais aussi en 6D (information de pose), à savoir la position et l’orientation de la caméra (ou pseudo-caméra) qui a pris la photographie.

Cette information peut être obtenue en s’appuyant sur différents types de référentiels. D’une simple base d’adresses à un nuage de point 3D très précis, la localisation peut aussi être dérivée des localisations des images similaires.

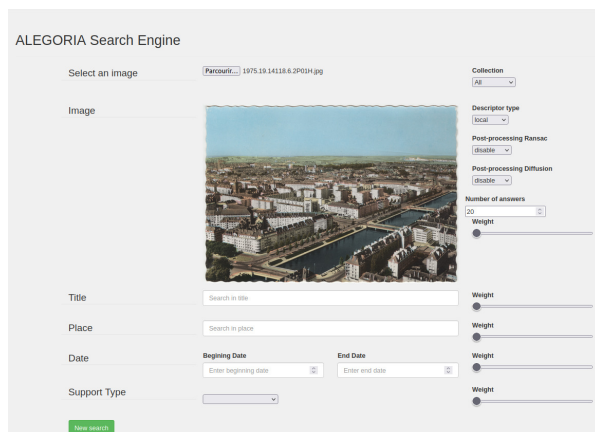
La localisation peut également être obtenue selon une multitude de méthodes plus ou moins automatiques. Tout d’abord, manuellement, en pointant la localisation d’une image sur une carte. De manière semi-automatique ensuite, en sélectionnant des points correspondants en 2D dans l’image et en 3D dans une scène puis en utilisant des algorithmes de résection spatiale pour obtenir une pose. Pour finir, de manière totalement automatique, en géocodant les adresses associées pour obtenir une position 2D ou 3D ou bien en exploitant des réseaux entraînés sur une scène 3D qui peuvent produire directement la pose de l’image.

Cette information est importante car elle permet de structurer de manière très simple les collections d’images et offre un paradigme naturel de visualisation qui peut être poussé très loin, comme la seconde partie du chapitre le montre.

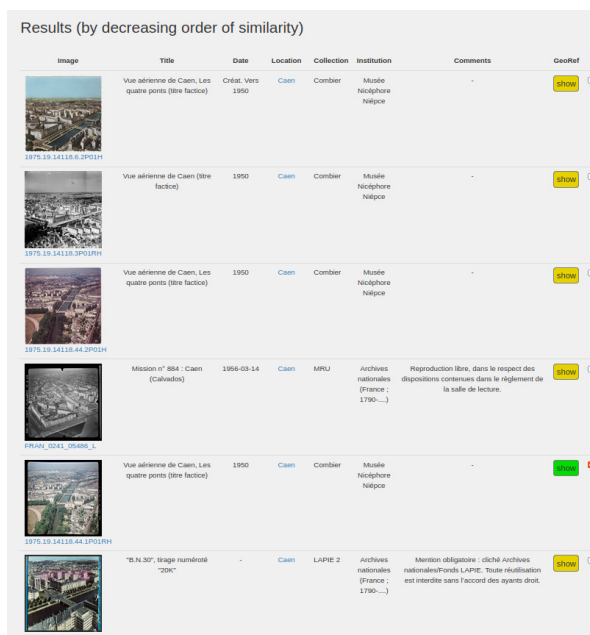
La deuxième partie du chapitre se concentre sur les modes de structuration et de visualisation de l’iconographie patrimoniale. Différentes modalités sont possibles et combinables pour obtenir des représentations visuelles plus ou moins complètes et adaptables à d’autres collections.

Tout d’abord, certaines plateformes n’exploitent qu’un seul paradigme de structuration qui impacte leur visualisation. Cela peut être une organisation par métadonnées. Créant soit des plateformes basées sur des mots-clés, soit des représentations en graphes de métadonnées similaires pour visualiser l’organisation globale de la collection. La similarité visuelle peut aussi être exploitée pour refléter la structure des collections. Les images sont alors organisées visuellement selon leur proximité dans l’espace des descripteurs d’image utilisés. Pour finir, le troisième paradigme de structuration et de visualisation exploite la similarité spatiale. En effet, le fait de simplement positionner les contenus sur une carte donne de l’information sur la structure de la collection et permet de sélectionner les images potentiellement similaires (*a fortiori* dans le cas de collections représentant des paysages ou des objets géographiques inamovibles).

La structuration et la visualisation peuvent ensuite être enrichies et complexifiées en combinant les modalités d’organisation. Certaines plateformes vont donc combiner plusieurs des modalités sus-mentionnées, notamment les métadonnées et l’information spatiale. Par ailleurs, la multiplicité des informations de localisation (présentées plus tôt) permet différents paradigmes de visualisation. La visualisation peut être plus ou moins immersive et plus ou moins adaptable à de nouveaux contenus. Un exemple de plateforme combinant les trois modalités précédentes est la plateforme du projet ALEGORIA (ALEGORIA project, 2018) illustrée dans les Figures 8.6 et 8.7. Les images similaires à



(a) Recherche d'images similaires



(b) Liste d'images similaires

Figure 8.6: Plateforme de recherche d'images du projet ALEGORIA, images de (Geniet et al., 2022)

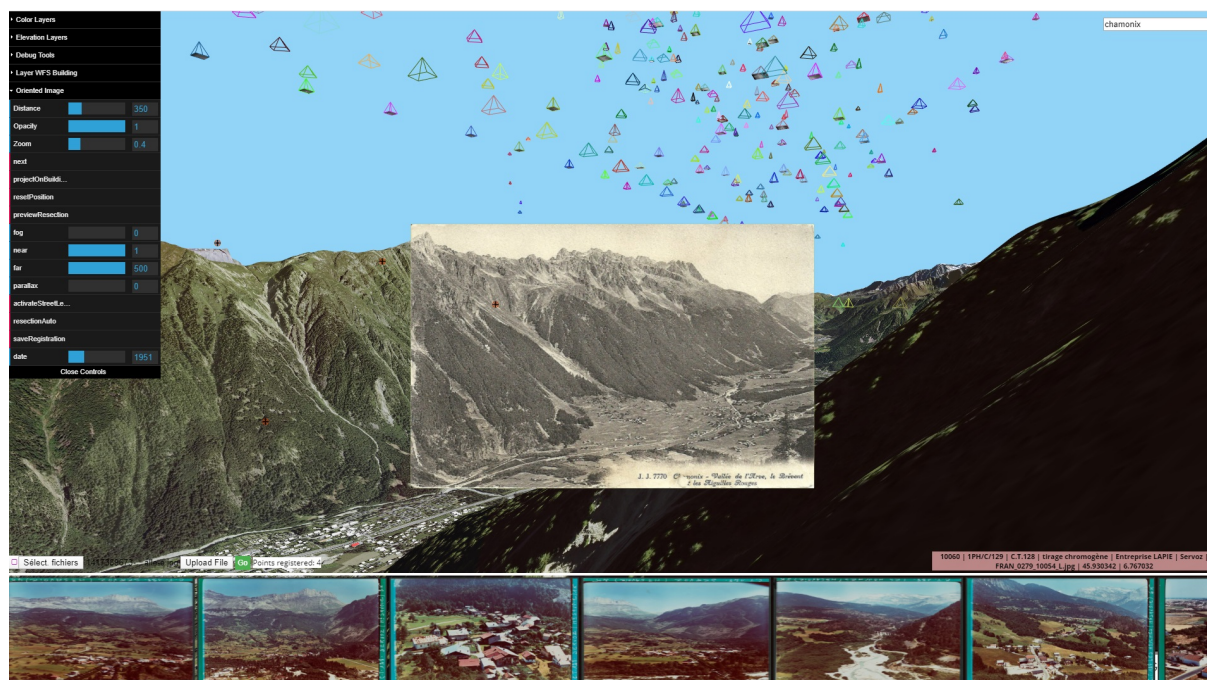


Figure 8.7: Plateforme de visualisation du projet ALEGORIA, de nombreuses images sont visualisées au même moment (toutes les "pyramides" correspondent à une image), image de (Blettery et al., 2020)

l'image requête peuvent d'abord être retrouvées selon une similarité visuelle ou selon des métadonnées semblables. Les images peuvent ensuite être visualisées et localisées dans une plateforme de visualisation 3D immersive couvrant potentiellement le monde entier.

Ces différentes plateformes peuvent permettre une certaine structuration des collections considérées (notamment par la spatialisation) mais ne permettent pas toujours de combiner visualisation et structuration globale selon différentes similarités. Les plateformes se limitent souvent à la visualisation, ou bien limitent la structuration à des traitements hors-ligne distincts de la plateforme de visualisation. C'est ce qui justifie notre proposition d'un processus de visualisation et de structuration semi-automatique global dans le chapitre suivant.

9 Chapitre 7 : Ré-ordonnancement semi-automatique par une approche basée sur le graphe

Ce chapitre présente notre dernière contribution : un processus semi-automatique de structuration des collections d'images qui traite l'ensemble des images et les différentes similarités entre elles comme un graphe. Le processus s'appuie sur des méthodes automatiques performantes -notamment la diffusion-, et une plateforme de visualisation de graphe dans laquelle un expert peut intervenir sur la structure de la collection.

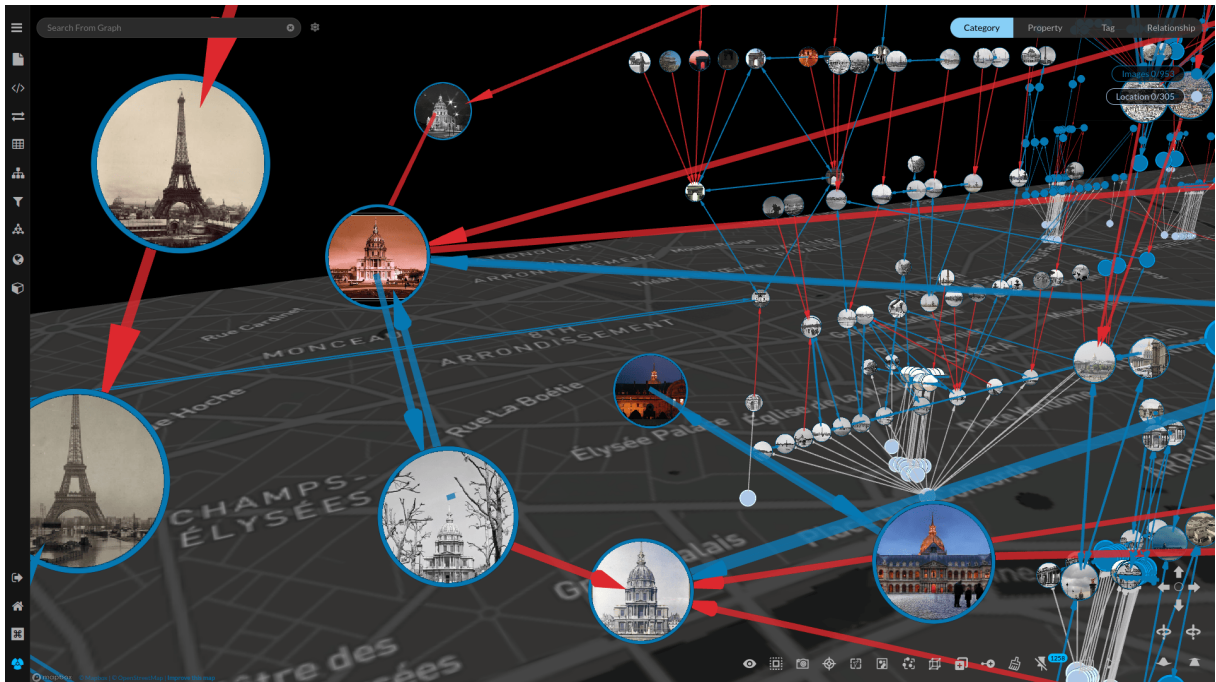


Figure 8.8: La plateforme de visualisation de graphe en 3D

Ce processus exploite différentes similarités. Tout d'abord, la similarité visuelle extraite automatiquement avec les descripteurs présentés dans la première partie. Ensuite une éventuelle similarité spatiale, extraite à partir des informations de localisation associées aux images. Et pour finir, une similarité experte, ajoutée manuellement par l'expert,

qui encode une similarité entre contenus plus ou moins définie, telle qu'évaluée par un connaisseur de la collection. L'ensemble de ces similarités permet de créer une structure de graphe que l'on visualise de manière spatialisée en 3D dans une plateforme dédiée, comme l'illustre la Figure 8.8.

Une fois la structure visualisée dans la plateforme, l'expert peut la modifier en supprimant ou ajoutant des similarités. Ces modifications sont ensuite exploitées dans le processus de diffusion de manière itérative comme présenté dans le schéma de la Figure 8.9.

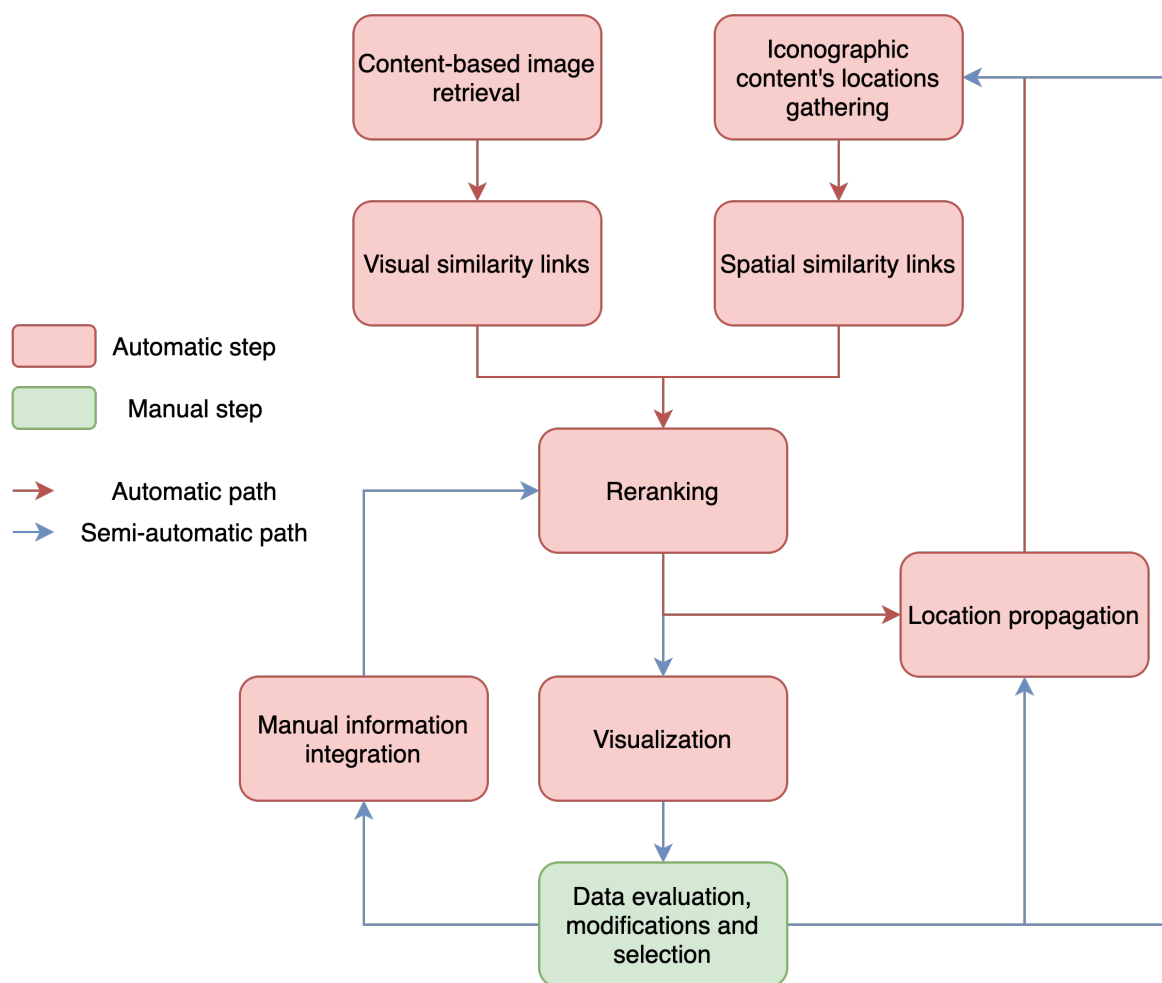


Figure 8.9: Schéma du processus de structuration semi-automatique

Suite à l'action de l'expert, la performance de la diffusion est meilleure, ce qui améliore globalement la structuration de la collection. Différents indices visuels sont proposés dans la plateforme pour guider l'expert vers des zones incertaines de la structure. Exploitant les algorithmes de graphe ou différents paradigmes de représentation, ces indices visuels accélèrent le travail de l'expert et le rendent plus impactant. L'objectif est que le processus combine le meilleur des deux mondes, à savoir la certitude des interventions expertes et l'impact à grande échelle des méthodes automatiques.

Ce processus est prometteur pour améliorer la structuration globale des collections considérées, comme le montre le Tableau 8.5. En effet, alors que le nombre d'interventions manuelles reste modéré, avec un temps d'intervention total d'environ 2 heures, l'améliora-

tion de la structuration est très importante. Par ailleurs, les similarités ajoutées ou supprimées le sont avec un fort degré de confiance. Et de plus, l’expert peut concentrer ses interventions pour assurer une structure de départ idéale pour la diffusion (notamment l’entropie des différents fournisseurs pour l’interconnexion des collections).

Ces résultats apparaissent prometteurs pour confirmer que l’intervention ciblée d’un expert permet de corriger des erreurs des processus automatiques de liage des données. De plus, le processus itératif passant par la visualisation permet également de mieux appréhender les collections et d’identifier des dynamiques qui leur sont propres.

Table 8.5: Evolution du score de mAP après plusieurs itérations de diverses modifications de la structure

Action #	Intervention type	Automation level	Number of added information	mAP before diffusion	mAP after diffusion
1	Image retrieval	Automatic	-	41.97	61.77
2	+ Location propagation	Automatic	85	42.32	62.20
Interventions on the first 5 links					
3	+ Deletions (visual)	Manual	70	42.36	62.32
4	+ Creations (expert)	Manual	30	42.40	62.46
5	+ Creations (spatial)	Manual	33	42.43	62.58
Interventions on the 5th to 10th links					
6	+ Deletions (visual)	Manual	78	42.44	62.59
7	+ Creations (expert)	Manual	26	42.48	63.83
8	+ Creations (spatial)	Manual	27	42.51	64.21

10 Chapitre 8 : Conclusion

Cette thèse s’est donc concentrée sur la structuration des collections d’images par interconnexion de contenus, au sein des collections et entre collections. Elle se concentre sur un objet d’études complexe : les contenus iconographiques patrimoniaux. Les conclusions de cette thèse sont doubles.

Nos **premières contributions** se concentrent sur les approches automatiques de recherche et de ré-ordonnancement d’images similaires. Ces approches peuvent en effet être utilisées pour structurer et relier des contenus iconographiques provenant de plusieurs collections. Tout d’abord, une évaluation des méthodes existantes est effectuée pour identifier ce qui fonctionne ou non pour un cas d’étude plus complexe. Exploitant ces conclusions, plusieurs nouvelles approches de ré-ordonnancement sont proposées, exploitant une information de structure plus globale. Deux approches explorent la géométrie de la scène, tandis qu’une troisième se base sur des informations de localisation pour pondérer la similarité visuelle entre les images. Par ailleurs, nous avons montré l’importance de combiner différentes approches de ré-ordonnancement pour améliorer au maximum la performance

de la recherche d'images.

Ces approches automatiques ont des applications pratiques pour les gestionnaires de collections. Cela peut être la détection de doublons, la vérification de la cohérence des métadonnées, l'enrichissement des métadonnées ou bien l'attribution de nouvelles informations à des images sans métadonnées.

Notre **dernière contribution** porte sur la structuration semi-automatique des collections d'images. En effet, nous avons montré que malgré des approches automatiques performantes, certains cas complexes nécessitent une intervention manuelle. Nous proposons un processus de structuration semi-automatique utilisant une représentation en graphe de la structure des collections d'images, liées par différentes similarités. Ce processus intègre des méthodes automatiques de recherche d'images pour créer une structure initiale. Celle-ci peut ensuite être affinée grâce à des corrections manuelles par un expert dans une plateforme de visualisation adaptée. Cette approche permet donc une visualisation structurée de la collection et permet de résoudre les problèmes de structuration les plus complexes que les méthodes automatiques ne peuvent pas traiter efficacement.

Concernant les perspectives futures de la première partie de la thèse, les pistes suivantes pourraient être explorées :

- réentraîner les réseaux des méthodes automatiques sur des données iconographiques patrimoniales pour améliorer les méthodes de recherche automatique ;
- explorer l'utilisation d'autres types d'information de structuration présents au sein des collections, en plus des informations spatiales que nous exploitons ;
- tester ces approches sur des problèmes de liage réels rencontrés par les gestionnaires de collections, ce qui permettrait d'affiner encore plus les besoins et d'adapter les méthodes.

En ce qui concerne la seconde partie de la thèse, de nouvelles pistes d'améliorations sont envisagées :

- exploiter d'autres informations de structuration, potentiellement basées sur des annotations au sein de la plateforme ;
- améliorer la plateforme avec de nouvelles fonctionnalités et une meilleure ergonomie.

L'objectif final est de faciliter son utilisation par des non-experts, notamment en permettant à plusieurs personnes de travailler de manière concurrente. Par exemple, que chaque amélioration de la structure se propage chez tous les contributeurs en même temps.