Generalizable features and image search for multi-source interconnection and analysis

Dimitri Gominski

Université Gustave Eiffel

Generalizable features and image search for multi-source interconnection and analysis

Description et recherche d'image généralisables pour l'interconnexion et l'analyse multi-source

Thèse de doctorat de l'Université Gustave Eiffel

École doctorale n° 532, Mathématiques et sciences et technologies de l'information et de la communication (MSTIC)
Spécialité de doctorat: Informatique
Unité de recherche : LaSTIG (IGN), LIRIS (Centrale Lyon)

Thèse présentée et soutenue à l'École Centrale de Lyon, le 09/11/2021, par

Dimitri GOMINSKI

Composition du Jury

Jantien STOTER Professeure, Delft University of Technology	Présidente
Peter BELL Professeur, Friedrich-Alexander Universität Erlangen-Nürnberg	Rapporteur
Philippe JOLY Professeur, Université Paul Sabatier Toulouse	Rapporteur
Dimitris SAMARAS Professeur, Stony Brook University	Examinateur
Valérie GOUET-BRUNET Directrice de recherche, Université Gustave Eiffel	Directrice de thèse
Liming CHEN Professeur, École Centrale Lyon	Directeur de thèse

Abstract

With an ever increasing volume of digitally accessible images, establishing connections to organize and analyse data is all the more important. A typical formulation for connecting images without using metadata is content-based image retrieval (CBIR). Similarly to other applications in computer vision, CBIR has benefited from the expressivity of convolutional neural networks (CNN) and obtained unprecedented results on usual benchmarks. However, it is hard to say whether this performance is explained by the proposal of more and more sophisticated architectures and models, or simply by the presence of a training dataset that matches the use case, i.e. that has similar visual and semantic characteristics. Indeed, the usual paradigm of the model-training dataset couple shows its limits as soon as one leaves the case characterized by the training data: the performance drops when the model is tested on different data, or data with too high variability.

This thesis addresses this issue with a critical look at deep learning methods and their real application potential. In a context of multi-source geographical imagery, a benchmark is proposed to characterize a new research problem: heterogeneous image retrieval, "low-data" (without training data), with a use case where defining a training dataset and a baseline method is not easy: the interconnection of iconographic collections from different heritage institutions. With this benchmark, new measures are proposed to qualify the generalization ability of the model in a CBIR context, then technical solutions that allow to get rid of the hazardous definition of similar visual and semantic characteristics. The discussion around the results highlights a probably too great importance given to the architecture of neural networks, and promising ideas in CBIR which provide model agnostic tools, and allowing to exploit the comparative advantages of different models trained on different data sets. Finally, the interest of this generalist approach is confirmed by a second application to land-use classification with high-resolution satellite imagery, a case where despite the abundance of methods and data, they are encapsulated in a set of small datasets and therefore with a limited application potential.

Résumé

Avec un volume toujours plus grand d'images accessibles numériquement, établir des connexions pour structurer et analyser les données devient d'autant plus important. Une formulation typique pour connecter entre elles des images sans utiliser de métadonnées est la recherche d'image basée contenu (RIBC). Similairement aux autres applications en vision par ordinateur, la RIBC a bénéficié du pouvoir expressif des réseaux de neurones convolutifs (CNN) et obtenu des résultats inédits sur les benchmarks usuels. Cependant, il est difficile de dire si cette performance est due à la proposition d'architectures et de modèles toujours plus évolués, ou simplement à la présence d'un jeu de données d'entraînement qui correspond bien au cas d'usage, c'est-à-dire qui a des caractéristiques visuelles et sémantiques similaires. En effet, le paradigme habituel du couple modèle-jeu d'entraînement montre ses limites dès lors qu'on sort du cas caractérisé par les données d'entraînement: la performance chute si on teste sur des données différentes ou avec une variabilité trop grande.

Cette thèse s'intéresse à cette question avec un regard critique sur les méthodes d'apprentissage profond et leur potentiel réel d'application. Dans un contexte d'imagerie géographique (vue aériennes obliques ou verticales) multi-source, un benchmark est proposé pour caractériser un nouveau problème de recherche: la recherche d'image hétérogène, "low-data" (sans données d'entraînement), avec un cas d'utilisation où définir un jeu de données d'entraînement et une méthode adéquate n'est pas facile: l'interconnexion de collections iconographiques provenant de différentes institutions patrimoniales. Avec ce benchmark, de nouvelles mesures sont proposées pour qualifier la capacité à généraliser du modèle dans un contexte RIBC, puis des solutions techniques qui permettent de s'affranchir de la définition hasardeuse des caractéristiques visuelles et sémantiques similaires. La discussion autour des résultats permet de mettre en valeur une importance probablement trop grande donnée à l'architecture des réseaux de neurones, et des pistes prometteuses dans la RIBC qui fournit des outils agnostiques du modèle utilisé, et permettant d'exploiter les avantages comparatifs de différents modèles entraînés sur différents jeux de données. Enfin, l'intérêt de cette approche généraliste est confirmé par une application à un deuxième cas, où malgré l'abondance de méthodes et de données, elles sont encapsulées dans un ensemble de petits datasets et donc peu généralisables: la classification d'occupation au sol en imagerie satellite.

Acknowledgments

This thesis and my introduction to the world of research would not have been possible without the trust my advisors Valérie & Liming have placed in me, and their valuable advice throughout these three years (and a half) to guide my work.

I am deeply grateful to my colleagues at LIRIS and LaSTIG, Martyna, Hugues, Margarita, and others with whom I exchanged ideas, their views were oases of certainty in the desert of questioning that is the thesis. I also want to thank members of the computer vision lab in Stony Brook, Dimitris, Hieu, for their invaluable advices, and for helping me build a necessary critical eye on the intimidating flow of computer vision research output.

As artists do, I think there is an important support work to maintain during the thesis in order to keep an open and positive mind, and the energy to continue the exploration. My family and friends have been in that regard essential. Most importantly, my now partner Esperanza has always supported and encouraged me, and never failed to lift my spirit when needed. This thesis is dedicated to her, and to the giants on whose shoulders I stand, to, hopefully, see further.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Despite a global pandemic and a related temporary decline of interest for photography, humanity has produced an estimated 1.12 *trillion* photos in 2020 [1]. With the democratization since the 19th century first of cameras, then of mobile phones which are both sensors and displays, saying that we have grown an interest for images would be an understatement. Imagery allow us to capture and share scenes or moments that have caught our attention, with purposes ranging from identifying a cancerous tumor to expressing a feeling.

Behind each image, there is an intention. Besides the intention of capturing a certain set of objects (or object parts) at a certain time, photographs are also the expression of a point of view on reality. Depending on the intention, the photographer makes this point of view personal (artistic photography) or on the contrary follows a standard to maximize relevant information (medical imagery). Onto the image, this context is translated as an ensemble of choices for organizing and presenting visual information in a way that will hopefully carry the desired message.

Now, considering the volume of images captured since the invention of photography in the 19th century, and the 7.9 billion potential photographers on Earth (not counting those who are machines and do it automatically), there will inevitably be multiple images depicting the same realities. And each of these images is a potential source of new, relevant information, a different perspective. This is particularly true for photos of places and persons, or more broadly any object that is experienced by or interacts with multiple persons. The variety of viewpoints constitutes a rich support for undestanding both the depicted reality, and how different individuals perceive it.

But with the richness of variety comes the challenge of bringing structure to the content, connecting images together. Different viewpoints imply different visual characteristics, and there are limits to what a visual system can recognize and match, be it the human visual system or a computer-based system. Hence the major interest in studying the one that can be easily modified and that can quickly process enormous volumes of images: computer vision. Computers typically process images sequentially, passing pixel values through series of operations with the goal of extracting meaningful

information: *features*. A core challenge is thus to make these features insensible to visual variations in the input image.

Consider figure 1.1. I personally know the place, so it is easy for me to say "these three photographs are connected, they represent the same place". But an uninformed person would probably end up saying "they are visually similar, it might be the same place but I am not sure". With a little bit more of context, like noticing the common word between the top and the bottom images, or the common shape of the lake between the top and middle images, they would gain confidence and now be able to identify new images of the lake. Now imagine these three images drowned in a volume of thousands, if not millions, other images, of which a certain fraction will be visually similar since the Lac de Nantua is probably not the only lake stuck between two mountains in the world (a geological particularity named "cluse" in French): it becomes even more difficult to identify images of the same place. Computers have the advantage of not being particularly bothered by high volumes of data, but they lack the modularity and adaptability of our visual reasonning, which allow us to quickly grasp an understanding of any object and apply the gained knowledge to a new task (here, the task of deciding if this new picture shows the Lac de Nantua).

In this thesis, I will characterize and explore the problem of connecting and analysing images coming from various sources and put together in a mixed database. Approaching the problem from a content-based image retrieval perspective, I will identify and propose solutions for enhancing the abillity of computer vision models to *generalize*, *i.e.* have satisfying performance in a broad range of situations, and show how this approach has applications beyond the task of retrieving images similar to a query. The experiments presented here are conducted with a focus on territorial imagery because it is a good example of data where generalization is crucial, but the approaches are agnostic of the type of images used.

The main contributions of this thesis are:

- Two datasets: the ALEGORIA benchmark for evaluating image retrieval methods in the context of territorial imagery, against difficult variations such as extreme viewpoint or scale changes ; and the SF300 large-scale training dataset, the first to propose multiple oblique aerial images of the same location with more than 300,000 samples.

- A methodological framework for evaluating computer vision on the task of interconnecting images from different sources in a database, with associated metrics.

- A post-processing method for image retrieval allowing the combination of multiple image features to exploit their comparative advantages. Coined Multi-descriptor Diffusion (MD), we obtain competitive results on the ALEGORIA benchmark with it and show that its constrained variant, cMD, is able to solve the compromise between absolute performance and cross-collection retrieval.

- An exploratory work on few-shot retrieval, inspired from recent propositions in meta-learning.

Figure 1.1: An example with old photographs of the lake near my hometown: do you recognize that it is the same location on the three photographs ? First picture is a scan of a photography bought by my father on a flea market, second is from the French Mapping Agency, third is from a collection hosted by the French National Archives.

Photo credits - copyright, top to bottom: None, IGN Photothèque Nationale (1950) - See ign.fr, Collection Lapie - See Alegoria rights

We introduce the rCNAPS method, and experimentally prove that it is possible to exploit external information to gain some performance in image retrieval.

- A new baseline for classifying satellite images with land-use information. Borrowing ideas from image retrieval, we obtain state-of-the-art performance on multiple datasets and question previous conclusions in the litterature on the scientific obstacles to the land-use classification task.

I will begin by precising the historical, scientific and application context in chapter 2 motivating this thesis. Chapter 3 starts with a look at the data already available in territorial imagery. Introducing the ALEGORIA benchmark, I will continue by reviewing tasks and methods, to outline a toolkit of ideas answering to the problem of interconnecting images. In chapter 4, our methodological framework is presented along with experiments on ALEGORIA to evaluate the state of the art. The preliminary results are discussed, and provide a basis for our proposed MD, cMD and rCNAPS methods which we present and compare with other approaches. We will switch to satellite imagery in chapter 5, and show that while the data changes, our image retrieval approach stays relevant and competitive. The proposed baseline for land-use classification is shown to be both simple and accurate, without requiring extensive pre-existing knowledge of the target task. Finally, we conclude by having a look at what did not work and what could be the future of the methods proposed here.

Chapter 2

Context

The title of this thesis, "Generalizable features and image search for multi-source interconnection and analysis", is voluntarily quite vague in its formulation. Computer vision is a wide field of research, including many tasks and applications, which often have strong links in their formulation or in the methods used. This chapter will define the scientific context, *i.e.* the application interests and family of methods that will be covered, as well as related subjects that have provided inspiration.

2.1 Vocabulary

I will detail here the vocabulary terms used in the subject, as well as a few important terms used in the following. I invite the reader to go to the glossary should they need another definition. Note that the definitions you will find here result from my own understanding and from my exchanges with the research community, and may (as I have observed in my readings) slightly vary in the litterature.

Image search is the act of searching for images in a database. Naturally, a criterion is needed to decide which images are relevant, this criterion is usually a **similarity** measure: a score stating how similar two images are, based on a visual (shapes, colors, textures) or semantic (objects, functions, locations, more abstract concepts) comparison. **Image retrieval**, more specifically, is a particular task in a computer vision where a system is given a query image and must produce a list of results. To quantify the relevance of the resulting list, images are generally associated with labels (classes), which allow the definition of **positive** images as images sharing the same label as the query (we want their similarity score to be high), and **negative** images as images with a different label (we want their similarity score to be low). The term "image search" is used in the title because this thesis intends to go beyond image retrieval but builds on its core processing step: comparing images in databases.

Interconnection is the process of building links in the data, here using only image data (pixels). A plain approach for interconnecting images is image search.

Analysis is the process of inferring information on a new piece of data, here images, based on previous evidence.

Multi-source is not a common term in computer vision, but is the central motivation and challenge in this thesis. The devices and conditions with which images are acquired have an influence on how they visually represent the objects of interest. Accordingly, research in computer vision studies and measures this influence, and how it can be a problem for systems. Generally, the term **domain** is used to identify a group of images with common characteristics (*e.g.* paintings, photos) and **cross-** is used to identify applications where images vary along a single representation axis (*e.g.* cross-view for images with varying viewpoints). Considering the multiple variations presented in section 3.2, I propose to rather speak of multi-source or **heterogeneous** data, to account for all the visual variations that can be encountered with different image sources, and to not be restricted to a single type of visual characteristic.

Features are image representations, computed to carry a task. They have several expected properties that are discussed in Section 2.3.1.

Generalizable is an umbrella term referring to the ability of going beyond the original or principal purpose. This is particularly important in computer vision, where we would like systems to reach or exceed the ability of humans to adapt to new problems. The change of visual characteristics implied by multi-source data is a good example of a situation where generalization is a key property to enforce on features. The notion of robustness is important to describe how to obtain generalization but needs more context before being introduced in Section 2.3.1.

2.2 Application context: the ALEGORIA project

This work, as part of a project funded with grant ANR-17-CE38-0014-01 from the French national research agency (Agence Nationale de la Recherche, ANR), has direct motivations and applications that will be detailed here.

Coined Advanced Linking and Exploitation of diGitized ge0gRaphic Iconographic heritAge (ALEGORIA), this research project stems from the observation that institutions housing image collections could benefit from automated processing. In particular, the project focuses on iconographic funds describing the French territory from the 1920s (when analog photography became accessible) to today. With digitization, a rich heritage of collections containing this type of content is available, but said collections are scattered among various institutions and lack metadata that could give them a role beyond consultation. The project proposes to develop tools to analyse, index and connect these collections, in a coordinated dialogue with end users. To demonstrate use cases, two online platforms are planned: an immersive display engine aligning ancient pictures with modern 3D models, and a multimodal image search engine coupling search by metadata and by image content (which is the purpose of this thesis).

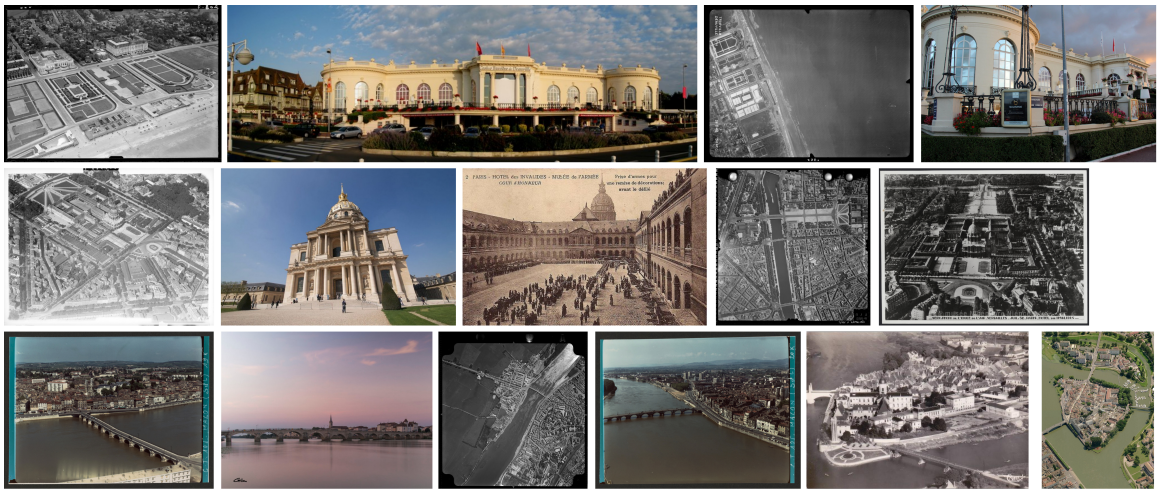

Figure 2.1: Examples from the ALEGORIA dataset. Images in the same row depict the same location.

Figure 2.1 shows some example from the ALEGORIA dataset. Note the wide range of visual characteristics, be it through color, vertical orientation, scale, image quality... These characteristics are usually shared for images coming from the same collection, *e.g.* images 1 and 4 on the third row.

The ALEGORIA project is a consortium of 5 actors:

- Two research laboratories in information and communication technology.
  - Laboratoire des Sciences et Technologies de l'Information Géographique (LaSTIG) from the French Mapping Agency (Institut National de l'Information Géographique et Forestière, IGN) and University Gustave Eiffel
  - Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS) from École Centrale de Lyon (ECL)
- Three content providers.
  - The French National Archives
  - The Nicéphore Niépce Museum
  - Data services of IGN
- One research laboratory in Social Sciences and Humanities.
  - Laboratoire Architecture Ville Urbanisme Environnement (LAVUE) of University Paris Ouest Nanterre

The project milestones, meetings and digitization processes have naturally influenced this thesis, but will not be detailed here. One point which should be mentioned however, is that the project being

application-driven, this thesis follows the same logic, with an emphasis on empirical findings rather than on modelization. Note that there is no (and there should not be any) value judgment here, as science has always been a back and forth between theorical fundations and real-world experimenting, both being equally necessary to progress towards a documented understanding of our world. Studying the case of ALEGORIA in this thesis serves to confront the state of the art to reality, to highlight the underlying problematics and to propose potential solutions.

2.3 Scientific context: the quest for generalization in computer vision

From a scientific point of view, this work finds itself at a particular moment in the history of computer science. Figure 2.2 shows some important dates for computer vision and machine learning. These two fields emerged in the 1950s, when the idea of building a system replicating or even exceeding human abilities began to take shape. Even if they had this common finality, there remained important technical milestones to pass before connections could be made between a digital vision system and an "intelligent machine": the first commercial digital camera was sold in 1989, years after theoretical works settled the basis for 3D reconstruction from 2D images for example ; and neural networks have not gained significant interest before 2012 when Graphics Processing Units (GPUs) became widespread, whereas the principle of backpropagation, essential to train a neural network, was formulated in the 1960s and 1970s.

It is precisely this 2012-breakthrough of GPU parallelized computation that intimately tied machine learning (now often rebranded as deep learning) and computer vision. It is not a coincidence if the milestone paper of Krizhevsky et al. used a neural network to conduct the computer vision task of classification: because images are by essence very expressive to us (we highly value our visual system [27]), easy to digitally represent and manage, and because machine learning relies on large amounts of annotated data, it was naturally through vision that deep learning displayed its potential. Since then, the frontier between deep learning and computer vision has been blurry and will likely stay so forever.

But the end goal shall not be forgotten: building models reaching or surpassing the human visual system. And a major characteristic of our eyes and brain is their ability to quickly capture, treat and analyse visual information, in a broad range of conditions. While our sensors (eyes and optic nerve) have limitations in the light wavelenghts, resolution, and information quantity they can handle, our recognition system is remarkably good at processing patterns through various conditions. We can easily infer the parts of an object that are not visible, efficiently treat a high volume of visual information by focusing on what is important, effortlessly deduce depth and size information on familiar objects, recognize objects even if they have been visually altered by low lightning, distance, shape changes... This ability to handle a lot of situations is called *generalization* in computer vision,

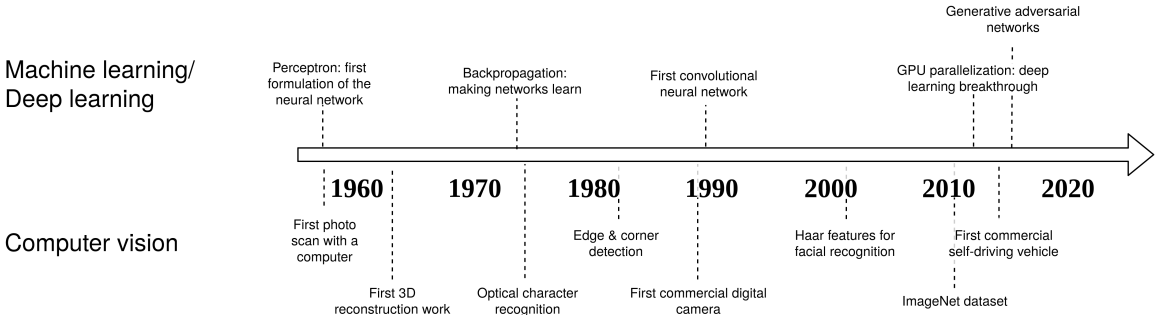

Figure 2.2: A brief history of computer vision and deep learning.

and remains a driving challenge, even with the recent advent of deep learning.

This thesis is by its application to images mainly a computer vision work, but heavily influenced by recent advances in deep learning. Convolutional neural networks and related methods are considered as tools to answer to the main challenge of this work: generalizable multi-source image interconnection and analysis. In this section, I will detail the fundamentals of computer vision used to this purpose in Section 2.3.1, how they are now expressed through deep learning in Section 2.3.2, and why generalization remains an open challenge even with the unprecedented performance of convolutional neural networks in Section 2.3.3.

2.3.1 Image features

A common paradigm in computer vision is to process images through two main steps. The first step consists in applying series of transformation to the input image to extract *features*: a compact, condensed set of information, used in the second step to conduct the considered task. The motivation behind this conceptual separation is twofold:

- Images are very high dimensional pieces of data: for example a modern 512x512 pixels image, with 3 color channels (red, green, blue) corresponds to 786,432 floating point values. A commonly adopted view on images is that they are mainly visual noise, and the primary goal of the algorithm is to filter out everything that is not relevant to the task. Taking the simplistic example of recognizing a cat on a 512x512x3 pixel image, one can expect that only a small fraction of the input information actually serves to produce the decision "it is a cat".

- Features have more potential for generalization. Transforming the input image to a set of data with a pre-defined and compact size, with only relevant information, makes it controllable and potentially reusable for other tasks. Again with the cat recognition example, suppose now that we use another image of another cat: we can expect their features to have some sort of similarity that allow us to say: "these two images show the same thing", without the need to reach the level of abstraction necessary to say "they are cats".

Building informative and compact features is thus a core underlying task of computer vision, all the more when we want to build links in the data. In the case considered here, *i.e.* connecting and analyzing images independently of how the content is represented, the following properties are particularly essential:

1. Discriminativeness: the features must include unambiguous information about the content, allowing the retrieval of true positives (images depicting the same reality, with similar visual characteristics) while ignoring false positives (images depicting a different reality, with similar visual characteristics).

2. Robustness (and Expressiveness): the features must capture enough available information to get a wide enough conceptual knowledge of the object, and more importantly be able to capture it regardless of different visual conditions. This allows the retrieval of images depicting the same reality despite very dissimilar visual characteristics. From one image to the other, the object can radically change through 3D transformations (rotation, illumination change, scaling, perspective change..).

3. Compactness: the features must stay relatively compact so they can be applied to large databases of images without requiring unrealistic computational times or material needs.

Note that there is intuitively a compromise to solve here: finding the optimal "spot" combining these three properties is not trivial, and enhancing one property while maintaining one constant (typically compactness) can lead to the remaining being worsened.

2.3.2 Deep learning

The main contribution of deep learning for computer vision is (arguably) the possibility of precisely tuning features to a task. Figure 2.3 shows a simplified overview of the usual deep learning framework. The essential components are:

- The Convolutional Neural Network (CNN). This is the core of the framework, it is the model (*i.e.* simply a long chain of parametrized operations) responsible for outputting relevant deep features. The network design sets the potential for how discriminative and expressive the features can be, and also how robust they can be to visual transformations in the input.

- The Loss function, a function expressing our satisfaction (or rather, our dissatisfaction) with the result. It relies on two inputs: one corresponding to the ground truth (here, any sort of information related to the content of the input image, a cat), the other corresponds to what was produced by the CNN. Here it is shown as directly taking deep features as input, but in reality it needs them to be processed so they can be compared to the ground truth that here will most likely be in the form of a textual label indicating "this is a cat". In our case, the loss function would take high values if the CNN doesn't predict "this is a cat", and vice-versa.

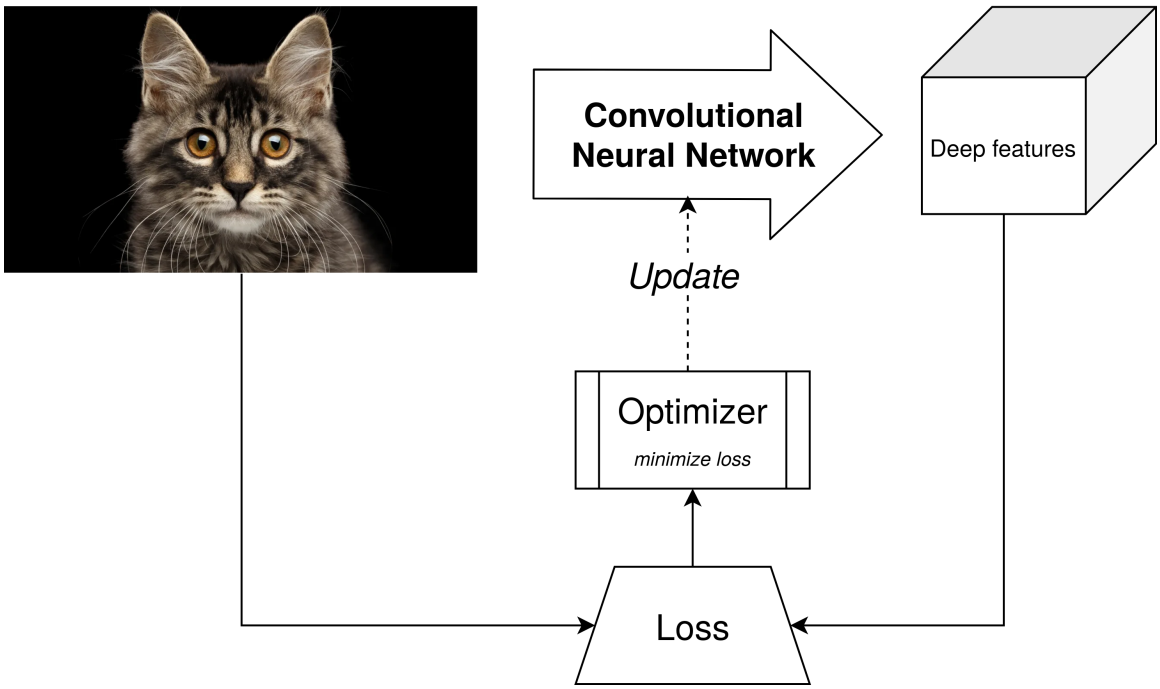

Figure 2.3: Simplified principle of deep learning.

- The Optimizer. The role of the optimizer is to provide feedback to the CNN so that it produces the desired qualities in features. This is done by minimizing the loss function. Multiple loss functions can be jointly optimized.
- The input data. Here only one image is represented, but in reality, if we aim for generalization, it is crucial to repeat the process of optimization through a very high number of examples. Indeed, it would be trivial to force the CNN to output "this is the cat" in this precise example, but it is not so trivial to make it output the same result for all possible images of cats.

2.3.3 Generalization

Summing it up, features are an essential building block for a computer vision framework, and deep learning has provided tools to impose constraints on them to get the wanted properties. The CNN defines the potential properties of features, and through optimization we express this potential and aim for the optimal compromise between these properties with the condition that enough data is available. This thesis focuses on how to learn and use features in a multi-source setup, *i.e.* in a setup where the visual characteristics of images can vary widely.

With the integration of deep learning into computer vision, some questions merit consideration:

Has deep learning enhanced the generalization of features ? If we look at the impressive

results on large-scale classification [44], retrieval [106] or detection [75] benchmarks, yes, progress has undoubtedly been made in the quality of features. But note that these benchmarks are all characterized by a limited range of visual conditions: objects are generally unambiguous, central, not occluded. The purpose of these benchmarks is to evaluate methods on a given task and semantic content (classify object categories in ImageNet, retrieve landmark pictures in GoogleLandmarks), not to exhaustively represent all the ways in which reality can possibly be visualized. So a remaining and more precise question would be:

Has deep learning enhanced the robustness of features, *i.e.* their ability to be used in a wide variety of visual conditions ? This is more a matter of debate. Concurrently with general-purpose computer vision, works have tackled the idea of handling images with different visual characteristics. The keyword "domain" is used to refer to a set of visual characteristics that can be grouped with a common denominator, and that separates corresponding images from others with different characteristics (for example pictures/drawings/paintings, or vertical/oblique/ground-level). Research interests such as cross-domain retrieval [10], domain adaptation [66] or domain generalization [26] have proposed multiple solutions, applied to formulations where "domain" means representation technique (e.g. PACS [47] with painting/cartoon/sketch/photo) or aspect variation (Geo-YFCC [26] with objects across countries). A common point of all of these methods is that they do not significantly modify the architecture of the feature-extracting CNN, instead they introduce some post-processing steps on features and additional models to conduct the task. This is not surprising: even basic image transformations such as rotation [108] can significantly degrade the performance. Building invariant architectures is in itself an active field of research [81] which has yet to produce easily reusable feature extractors.

In this thesis, I approach the problem of interconnecting and analyzing multi-source images from an image retrieval perspective, but the underlying computer vision issue is undoubtedly the ability to generalize, through domains and data. "Generalization" is not a keyword commonly found in image retrieval litterature: it is implicitely an expected property, because the goal is to retrieve any positive image, regardless of its visual characteristics. In Chapter 3, I will try to show why it can be helpful to have a closer look at how these variations influence performance, and in Chapter 4 propose some measures and solutions to this problem.

2.4 Data

As presented in section 2.3, deep learning and more specifically deep features have strong links with data. A short definition seems important before continuing:

> *dataset* - a collection of separate sets of information that is treated as a single unit by a computer
>
> (Cambridge dictionary)

In computer vision, datasets are fundamental to represent real-world applications by providing a support to train, validate and evaluate methods. In this last case, the term *benchmark* is also used. A benchmark is a dataset expressly made for testing and comparing methods, with the following expected properties:

- Benchmarks are generally smaller than training datasets to limit computation time for evaluation, but large enough to give significant measures.
- An evaluation protocol must be provided (standardized evaluation metrics, standardized pre- and postprocessing steps), to allow comparison of methods in the same conditions.
- The variety of images, annotations and test setups should be as representative as possible of real-world use cases and allow precise evaluation of how methods behave in these setups.

On the contrary, when a dataset is specifically designed for training models, it is usually referred to as a *training dataset*. Such datasets usually have the following characteristics, especially in the deep learning era:

- High volume, with orders of magnitude ranging from 10^4 to 10^7 samples for image datasets.
- A carefully tuned balance between variety and repetition of visual patterns. Due to the architecture and learning process of Convolutional Neural Networks, training images must display the same pattern enough time for it to be learned, while displaying enough variety for the network to be able to generalize. A caricatural example in say, animal species recognition, would be a dataset only composed of paw pictures: the model would reach high accuracy in this specific case but fail to recognize a cat from its face. On the contrary, a dataset where each species is represented by only two samples (*e.g.* one face image, one picture from above) would not successfully learn any relevant pattern. Datasets that were empirically proved to have achieved this balance can be referred to as *clean* datasets.

As we will see later in Chapter 3, deep learning has blurred the line between method and data, and it is still not well understood, for many state-of-the-art methods, if their performance is due to the data or to the algorithm itself. To try to reduce the ambiguity, I will try as much as possible to dissociate methods from their training datasets in this thesis. Also, it seems important to adopt a critical look regarding how performance is evaluated. On many benchmarks, methods are compared with a relatively small set of indicators, averaging performance over a large number of images. While there is obviously a need for concise and comparable measures, they must be put in perspective with what they represent: real-world performance.

Chapter 3

Characterizing the cross-source retrieval problem

Taking into account the scientific and application context presented in Chapter 2, this chapter will present the elements that led me to consider that the task of connecting images accross sources with a high variance in visual characteristics constitutes a new problem for computer vision, both from a data and a method perspective, the two being intricately linked. Contrarily to what is usually done in computer vision papers (1. Pick a problem 2. Propose a method 3. Train and test on relevant datasets), I will begin by presenting the data, first related datasets and benchmarks in the literature in section 3.1, then the proposed benchmark in the context of ALEGORIA in section 3.2. I chose this order because even if many methods are implicitely presented as applicable to any image data, they are in fact only proven effective on the relatively small number of benchmarks they were tested on. Presenting the data first helps to understand what exactly makes the ALEGORIA case new and challenging. I will then propose a formulation of the cross-source content-based image retrieval problem in section 3.3, and end with a broad literature review to identify potential technical answers in sectio 3.4.

3.1 Datasets and benchmarks in the literature

In this section, I will give an overview of available training datasets and benchmarks with semantics or motivations close to what constitutes the ALEGORIA project. Recall from section 2.2 the application context: we want to connect images of the French territory, with a broad range of variations due to the different acquisition conditions (old digitized photographs, modern satellite images...).

Looking at the literature, there are three research subjects which share some similarity with our case in terms of data and motivations: landmark retrieval (retrieving historical or architectural

notable monuments and buildings), remote sensing image analysis (analysing images captured from planes and satellites) and place recognition (matching multiple images of the same place).

Landmark retrieval is a driving application of instance retrieval, because it presents interesting challenges: monuments are large objects, with a high level of details and possible points of view (from the ground, from the air, from the inside, on the monument..), placed in complex and changing environments (changing weather, time of day or surroundings). Common benchmarks are presented in Table 3.1. The two most influential benchmarks were Oxford5k and Paris6k, which proposed a separate ground-truth for each query differentiating positive, negative and *junk* images, *i.e.* positive images that are judged too difficult to retrieve (sharing less than 25% of landmark surface with the query). Radenovic et al. later enhanced these two benchmarks in 2018 with corrected, more detailed annotations and three evaluation setups: *easy* (only positive images sharing more than 25% of the landmark are considered, other positive images ignored), *medium* (all positive images are considered) and *hard* (only positive images sharing less than 25% of the landmark are considered, other positive images ignored). These benchmarks are associated with a set of one million relevant distractors $\mathcal{R}$1M (images sharing visual and semantic similarities with the core benchmark), that serves to simulate real-world conditions where images might be mixed in large scale databases.

See Table 3.1 for an overview of the three most used training datasets in landmark retrieval. Training datasets mostly differ with their process of annotation. Considering the demand for high volumes, it is virtually impossible to manually annotate each image individually, it therefore becomes essential to design semi-automated solutions: Babenko et al. first use the Yandex image browser for scraping images of the most famous landmarks, then roughly filters out poorly built classes with human supervision to build the Landmarks dataset ; Radenović et al. use 3D models to automatically build n-tuplets with one challenging positive image (sharing multiple points with the query but limited in scale variation) and multiple challenging negative images (with similar descriptors) to build the SfM-120k dataset ; and Noh et al. exploit image metadata (notably GPS coordinates) and local feature filtering to build the GoogleLandmarks dataset.

The main caveats that limit the knowledge that can be transfered from landmark retrieval to the ALEGORIA case are 1. Extreme changes in viewpoints are considered too difficult to build a single class in benchmarks, whereas the relatively low volume of images of the same place in ALEGORIA collections, with sometimes extreme scale or vertical orientation changes, does not allow such simplifications. 2. While datasets claim that they include enough variety and difficult cases to represent real-world conditions, their data is biased towards their source: internet images (from Flickr, Google, Yandex image databases), which is not representative of the variety of aerial and digitized photography found in ALEGORIA. 3. Landmarks are by definition distinctive, whereas images in ALEGORIA sometimes depict "ordinary" objects and locations that might not show distinctive visual features. Landmark retrieval however remains the closest application compared to ALEGORIA, as we will empirically demonstrate later in section 4.1.

Table 3.1: Overview of landmark retrieval datasets.

Name	Year	Nb. of classes	Nb. of queries	Total size
Benchmarks				
Holidays [39]	2008	500	500	1491
Oxford5k [70]	2007	11	55	5062
Paris6k [71]	2008	11	55	6392
$\mathcal{R}$Oxford [73]	2018	11	70	4993
$\mathcal{R}$Paris [73]	2018	11	70	6322
Training datasets				
GoogleLandmarks [63]	2017	81k	–	1.4M
GoogleLandmarksv2 [106]	2020	200k	–	4.1M
SfM-120k [74]	2019	551	–	120k
Landmarks [7]	2014	672	–	214k

Remote sensing image analysis is a growing research subject with applications in hazard assessment, natural ressource management, urban and rural planning... Note that the literature sometimes confounds remote sensing image analysis with satellite imagery analysis, because of the growing volume of available satellite data from large-scale acquisitions campaigns (*e.g.* Sentinel missions in the European Union or Landsat program in the United States), but remote sensing also includes data sourced from aircrafts. Remote sensing tools offer a wide range of acquisition techniques beyond simple RGB sensors, including but not limited to multispectral imagery (using a few discrete spectral bands), hyperspectral imagery (aiming for a full spectrum for each pixel), LIDAR (laser) data... Here we are only interested in data that has been processed to build standard RGB images. There is a great variety of datasets in remote sensing [78], most of them oriented towards categories of objects, *e.g.* roads, agricultural fields, buildings... ; with tasks such as scene classification (see Chapter 5), object segmentation, or object detection. The main caveats that limit the knowledge that can be transfered from remote sensing imagery to the ALEGORIA case are 1. The significant semantic gap between object categories (how to differentiate a road from a building ?) and object instances (how to differentiate this road from other roads ?). To my knowledge, there is no dataset mixing multiple sources of remote sensing images towards instance analysis, *i.e.* recognizing a particular object (for example a particular building) rather than a broader semantic class. This is particularly problematic regarding ALEGORIA, considering that we need to recognize specific places that are sometimes not distinctive. In Section 5.1 I will present a new dataset with this objective in mind. 2. The variety of ALEGORIA representation characteristics is still not covered by this type of data, with for example oblique imagery (aerial imagery with an angle from the vertical) absent (to my knowledge) from remote sensing datasets but well represented in ALEGORIA. Yet, considering

the significative volume of vertical satellite and aerial imagery found in ALEGORIA, remote sensing imagery will be kept as a potential source of valuable data in the following.

Place recognition is a wide task closely related to image retrieval, lead mostly by the application of geolocating embedded systems, such as autonomous cars or mobile phones, using their camera. Table 3.2 presents commonly used datasets. All of these datasets have been built with a certain variation in mind, linked either to temporal (weather, seasons, day-night) or point-of-view variations (ground/oblique/vertical). For temporal datasets, the goal is to be able to recognize a place by retrieving other images of the same place that differ with illumination, visual clutter (cars, persons) or visual modifications (snow, seasonal changes on vegetation). For cross-view datasets, the typical setup is to match street-level views (panorama or oriented) with aerial or satellite images . University1652 in particular is a dataset closely related to ALEGORIA, the only one to my knowledge mixing ground, oblique and vertical views. There however still remains caveats limiting the knowledge that can be transfered from place recognition datasets to the ALEGORIA case: 1. Each of these datasets considers a very specific case of variation. University1652 for example includes all orientations found in ALEGORIA, but ignores illumination changes. On the contrary, ground-level datasets (first 5 in Table 3.2) offer challenging illumination and scene variations but ignore cross-view problems. 2. The place recognition problem, even when formulated with image retrieval, is a simplification in the sense that only one positive image at the top of the list of results is needed to correctly localize the query, with the exception of Zheng et al. (University1652) including the AP (Average Precision) measure in their formulation. In ALEGORIA, we want to potentially connect any image of the same location. 3. Datasets are either biased towards urban environments, roads or distinctive buildings. In ALEGORIA, locations can be anything ranging from a natural scene to a whole city, including indistinctive buildings or semi-urban scenes. The University1652 dataset however presents relevant characteristics and will be included in the following.

Table 3.2: Overview of visual place recognition datasets.

Name	Year	Changes	Nb of queries	Nb. of locations	Total size
Pittsburgh [99]	2013	Illumination	24,000	1,000	254,064
Tokyo 24/7 [98]	2015	Day-Night	1,125	125	1,125
Extended CMU Seasons [8, 83, 93]	2020	Seasons + Weather	56,613	6,000	60,937
Aachen Day-Night [93]	2020	Day-Night	922	824	4,328
RobotCar Seasons [93]	2020	Seasons + Weather + Day-Night	5,616	3,978	26,580
CVUSA [107, 116]	2017	Ground-Vertical	71,064	35,532	71,064
University1652 [121]	2020	Ground-Oblique-Vertical	50,218	701	50,218
Lin et al. [53]	2015	Ground-Oblique	75,000	37,500	75,000

Reflecting on the study of these three topics and how they are formulated in the literature (*i.e.* what are their objectives and what challenges does the corresponding data present), a conclusion emerges:

Premise. *The* ALEGORIA *problem is unprecedented in the literature in terms of data, not because it includes new semantics or challenges, but because it is a mix of multiple challenging variations, not well represented in popular datasets, in a setup that doesn't allow difficult cases to be ignored.*

3.2 The ALEGORIA benchmark

To characterize the ALEGORIA problem and allow empirical comparison of approaches, a benchmark is needed. It should ideally represent all possible scenarios encountered in a real-world application, *i.e.* have statistics similar to those that a dataset with all the target data would have, it should contain enough samples to produce stable and representative measures, and it should, if possible, give detailed information on failure and success cases.

The realization of these objectives is expressed through the establishment of a detailed ground-truth. Put into practice, the ground-truth in image retrieval is a file stating for each query image, the images in the database that are considered positive and the images that are considered negative, but it can also include additional information about the images for extensive evaluation.

In this section, I will describe how we built the ALEGORIA benchmark with these constraints in mind, highlight the factors making the problem of cross-source interconnection challenging, and present some statistics about the class and attribute distribution.

3.2.1 Presentation

The ALEGORIA benchmark is a new research dataset for evaluating computer vision algorithms in the application context of interconnecting images coming from collections hosted in various heritage institutions. Consisting of multi-date, multi-sources images, with extensive annotations of classes and characteristics, its main theme is cultural, historical and geographical imagery. Classes are defined around an object or a location in urban or natural scenery. It is designed mainly for content-based image search, but can also be used in related tasks:

- Image-based geolocalization
- Cross-view image matching
- Invariant representation learning
- Few-shot landmark recognition
- Multi-temporal image matching

3.2.2 Dataset building

The dataset was built using images sources from partners of the ALEGORIA project, and completed with images from the internet with permissive rights.

Figure 3.1: Examples from ALEGORIA with varying illumination, shadow distribution.
Photo credits - copyright, left to right: Hunza - None, failing_angel - CC-BY-NC-SA 2.0, jean-nicolaslehec - CC-BY-NC-SA 2.0

Recall from section 2.2 the two use-cases in ALEGORIA: retrieving images depicting the same location, and retrieving similar local patterns or objects such as architectural elements. There is a significant gap in annotation costs between those two cases. The former requires annotation at image-level (assigning one or multiple locations to one image), while the latter requires annotation at object level. Moreover, image-level annotations for locations are compatible with straightforward automation based on keyword or GPS metadata, while object-level annotations need precise feedback from experts (ALEGORIA end user institutions). The ALEGORIA benchmark is thus, for now, annotated at image level with a class label.

Concretely, this was done with the following process:

1. A visual inspection of the collections allowed us to identify some objects or locations of interest, *i.e.* candidate classes that are represented with multiple images in a given collection.

2. Candidate classes were filtered to only keep those with images in at least two different collections.

3. The IGN "Photothèque" collection has the particularity of being particularly voluminous, and georeferenced. This allowed us to easily fill small classes for which we have some annotations giving hints about the location.

4. Classes with particularly well-known objects or locations were completed with images from the internet with permissive rights. Mostly, they correspond to mobile phone or amateur camera tourist photography.

3.2.3 A look at ALEGORIA challenges

A picture might not be worth a thousand words, but I think using visual examples to attest what makes ALEGORIA challenging is essential before jumping into statistics.

Figure 3.2: Examples from ALEGORIA with varying occlusion level/visible surface.
Photo credits - copyright, left to right: Nicéphore Niepce - See Alegoria rights, Alain Delavie - None, Henrard - See Alegoria rights

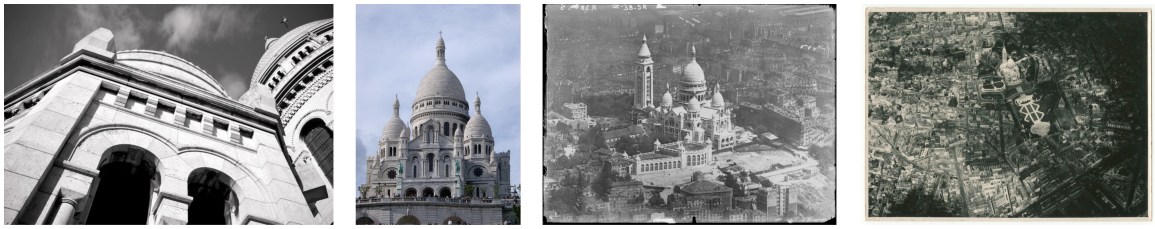

Figure 3.3: Examples from ALEGORIA with varying scale.
Photo credits - copyright, left to right: sottolestelle - CC-BY-SA 2.0, Photo Everywhere - CC-BY 2.0, Nicéphore Niepce - See Alegoria rights, Henrard - See Alegoria rights

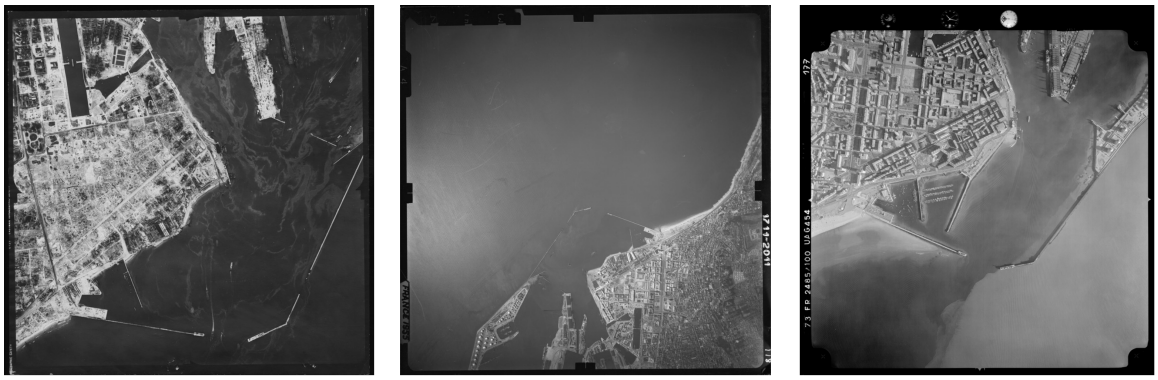

Figure 3.4: Examples from ALEGORIA with varying time of acquisition.
Photo credits - copyright, left to right: Photothèque IGN - See Alegoria rights, Photothèque IGN - See Alegoria rights, Photothèque IGN - See Alegoria rights

Figure 3.5: Examples from ALEGORIA with varying vertical orientation.
Photo credits - copyright, left to right: Tobias von der Haar - CC-BY 2.0, Photothèque IGN - See Alegoria rights, Photothèque IGN - See Alegoria rights

Figure 3.6: Example of an undistinctive ALEGORIA class defined with a location.
Photo credits - copyright, left to right: Photothèque IGN - See Alegoria rights, Photothèque IGN - See Alegoria rights, Photothèque IGN - See Alegoria rights

Figures 3.1 through 3.5 show some examples of images from the same class, varying along a single (or a predominant) type of image variation. Additionally, Figure 3.6 show a semantic difficulty: many classes in the ALEGORIA are not defined with a specific, easily recognizable object but rather with a location.

The fact that there are many different visual variations in the ALEGORIA benchmark, and that they are often mixed together, makes it hard to define precise "domains" as it is often done in the literature. Rather, it seems more relevant to identify the main sources of visual variation, and separate images depending with a degree of visual variation. This is the goal of **attributes**, the second type of annotation in the benchmark. Each image was annotated along 7 attributes that will be detailed in the next section. These attributes will, hopefully, help us understand with a finer level of detail what visual variations are problematic.

3.2.4 Statistics

Three data providing institutions collaborated with us in the ALEGORIA project. Classes were built by finding common images in collections, and then completed with content from Google Images and Flickr (according to their licenses). Table 3.4 shows statistics about the data sources.

Table 3.3: General statistics of the ALEGORIA benchmark.

Item	**Value**
Number of images	13174
of which annotated	1858
of which distractors (non annotated)	11316
Number of classes	58
Min number of images per class	10
Max number of images per class	119
Mean number of images per class	32
Median number of images per class	25
Image file format	.jpg
Image dimension (width*height)	800px*variable

Table 3.4: Sources of images in the ALEGORIA benchmark

Item	**Value**
Number of annotated images	1858
of which from Institut National de l'information Géographique et forestière	711
of which from Archives nationales	339
of which from Nicéphore Niepce museum	161
of which from the internet	647
Number of distractors	11316
of which from Institut National de l'information Géographique et forestière	1260
of which from Archives nationales	6701
of which from Nicéphore Niepce museum	2904
of which from the internet	451

The data providers possess multiple collections. Table 3.5 shows statistics about the collections in the ALEGORIA benchmark.

Table 3.5: Collections of images in the ALEGORIA benchmark

Item	**Value**	Corresponding institutions
Number of annotated images	1858	
of which from Henrard	99	Nicéphore Niepce
of which from Lapie	40	Archives nationales
of which from Combier	7	Nicéphore Niepce
of which from MRU	299	Archives nationales
of which from photothèque IGN	711	IGN
of which from the internet/unkown	702	Internet, Nicéphore Niepce
Number of distractors	11316	
of which from Henrard	935	Nicéphore Niepce
of which from Lapie	4508	Archives nationales
of which from Combier	1969	Nicéphore Niepce
of which from MRU	2193	Archives nationales
of which from photothèque IGN	1260	IGN
of which from the internet	451	Internet

Each class is defined by taking as a reference an object or a zone that is partially or completely visible in all images of the class. There is a total of 58 classes. Classes include, independently of the target object, urban, semi-urban and natural sceneries. Table 3.6 shows all class names with information about the class semantics, and table 3.7 sums up this information with statistics.

While there is naturally more images available for urban situations, the benchmark was created with the goal of representing the variety of territorial imagery, and thus also include semi-urban and natural landscapes. Statistics about the landscapes are found in Table 3.8.

3.2.5 Attributes

Each image has been manually annotated along 7 different attributes quantized with values ranging from 0 to 2 or 3.

Table 3.6: Class definition in the ALEGORIA benchmark

Class name	urban/natural	class definition	class type
amiens	urban	tower	object
annecy	semi urban	lake mouth	zone
arc de triomphe	urban	monument	object
basilique sacre coeur	urban	church	object
biarritz	semi urban	hotel, beach	zone
amiral bruix boulevard	urban	crossroad	zone
bourg en bresse	semi urban	factory	object
brest	urban	port	zone
fourviere cathedral	urban	church	object
reims cathedral	urban	church	object
saint etienne de toul cathedral	urban	church	object
deauville international center	urban	hotel	object
charlevilles mezieres	urban	square	zone
chantilly castle	semi urban	castle	object
palace of versailles	urban	castle	object
choux creteil	urban	tower	object
cite internationale lyon	urban	neighborhood	zone
foix	semi urban	castle	object
gare du nord paris	urban	train station	object
gare est paris	urban	train station	object
gare perrache lyon	urban	train station	object
grenoble	urban	river	zone
guethary	natural	hotel	object
saint laurent hospital chalon	urban	hotel	object
nantes island	urban	neighborhood	zone
invalides	urban	hotel	object
issy moulineaux	urban	bridge	object
la madeleine paris	urban	monument	object
le havre	urban	tower	object
lery seyne sur mer	semi urban	church	object
macon	urban	bridge	object
mairie lille	urban	tower	object
chasseneuil memorial	natural	monument	object
mont blanc	natural	mountain	object
mont saint michel	natural	nieghborhood	zone
neuilly sur seine	urban	neighborhood	zone
notre dame de lorette	natural	church	object
notre dame garde	urban	church	object
notre dame paris	urban	church	object
pantheon paris	urban	monument	object
picpus	urban	neighborhood	zone
place bourse bordeaux	square	square	zone
place marche clichy	urban	square	zone
bouc harbour	semi urban	harbor	zone
porte pantin	urban	neighborhood	zone
porte saint denis	urban	monument	object
aubepins neighborhood	urban	neighborhood	zone
reims racetrack	urban	neighborhood	zone
riom	urban	neihgborhood (town)	zone
saint claude	semi urban	church	object
gerland stadium	urban	monument	object
st tropez	semi urban	neighborhood (town)	zone
toulon	urban	neighborhood	zone
eiffel tower	urban	tower	object
tours	urban	neighborhood	zone
aillaud towers nanterre	urban	tower	object
vannes	urban	neighborhood	zone
villa monceau	urban	neighborhood	zone

Table 3.7: Object type for class definition in the ALEGORIA benchmark

Class definition	**Number of classes**	**Number of images**
Object	35	1290
of which churches/cathedrals	9	405
of which castles	3	167
of which towers	6	234
Area	23	671
of which neighborhood/whole city	14	452

Table 3.8: Landscape type for class definition in the ALEGORIA benchmark

Landscape	**Number of classes**	**Number of images**
Natural	5	133
Semi-urban	9	224
Urban	44	1501

Table 3.9: Attribute definition in the ALEGORIA benchmark

Attribute	Value	Meaning
Scale	0	Very close - object covers majority of the picture
	1	Close - object covers around half of the picture
	2	Midrange - object is small but distinguishable
	3	Far - object is hardly distinguishable
Illumination	0	Underilluminated - object is too dark, details are difficult to distinguish
	1	Well illuminated - object is well illuminated, details are visible, contrast is good
	2	Over illuminated - object is too bright, details are difficult to distinguish
Vertical orientation	0	Vertical - object is seen from the top, viewing direction is perpendicular to the ground
	1	Oblique - object is seen from a plane or high viewpoint, top surface and side(s) are visible
	2	Street view - object is seen from ground level, top surface is not visible
Representation domain	0	Picture
	1	Drawing
	2	Schematic
	3	Digital
	4	Other
Occlusion	0	No occlusion - object is not hidden behind other objects
	1	Partially hidden
	2	Occluded - only a small portion of the object is visible through multiple and/or large occluding objects
Alterations	0	No alteration - picture is in perfect quality, object is in focus
	1	Minor alterations - some blur or mild degradation of the medium is visible, but doesn't impact understanding of the object
	2	Major alterations - picture is heavily degraded and/or blurry
Color	0	Color picture
	1	Grayscale picture
	2	Monochrome picture - e.g. sepia
	3	Infrared picture

The attribute values have been as much as possible uniformized, priorizing first vertical orientation then scale. There is however classes with skewed distributions, generally towards an over-represented value. Table 3.10 below shows for each value of each attribute the maximum and minimum proportions that can be found in the same class.

Table 3.10: Attribute statistics in the ALEGORIA benchmark

Attribute	Value	Global proportion	Max proportion	Min proportion
Scale	0 - Very close	13%	35%	3%
	1 - Close	20%	60%	4%
	2 - Midrange	29%	77%	4%
	3 - Far	38%	100%	12%
Illumination	0 - Underilluminated	13%	36%	3%
	1 - Well illuminated	80%	100%	49%
	2 - Over illuminated	7%	51%	2%
Vertical orientation	0 - Vertical	33%	81%	8%
	1 - Oblique	38%	78%	3%
	2 - Street view	29%	52%	10%
Representation domain	0 - Picture	99%	100%	85%
	1 - Drawing	1%	15%	<1%
	2 - Schematic	<1%	<1%	<1%
	3 - Digital	<1%	<1%	<1%
	4 - Other	<1%	<1%	<1%
Occlusion	0 - No occlusion	76%	100%	50%
	1 - Partially hidden	20%	50%	3%
	2 - Occluded	5%	21%	3%
Alterations	0 - No alteration	69%	100%	22%
	1 - Minor	23%	59%	4%
	2 - Major	8%	36%	1 %
Color	0 - Color	39%	84%	5%
	1 - Grayscale	48%	92%	16%
	2 - Monochrome	13%	62%	3%
	3 - Infrared	<1%	6%	<1%

3.3 Proposed formulation: interconnection through image retrieval

In short, the objective is to connect images through various sources with challenging visual variations, and with the context of deep learning, a lack of training data.

In this section, I will present tasks in computer vision with links to these motivations, with the goal of arriving at a formulation of the cross-source interconnection problem through a well-defined task which will provide a basis of methods and ideas. In the literature, methods are often task-specific: they answer to a particular problem on a particular benchmark. There are however interesting connections between these different problems that are worth considering.

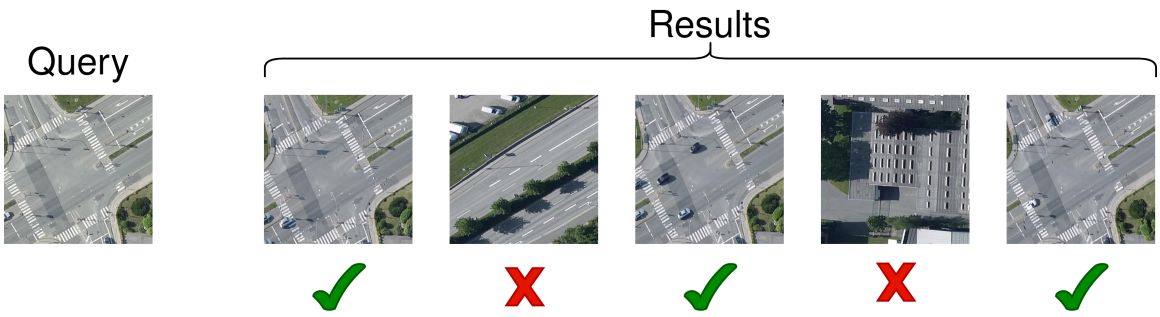

Figure 3.7: Content-based image retrieval principle: the goal is to retrieve images considered positive in relation to the query, using only visual information.

Image retrieval is one of many tasks in computer vision, and includes different definitions that may vary depending on the field of application. It is the task of finding images similar to a given image, the query, in a database (a commonly known example is Google reverse image search). A first important notion to precise is what we want to retrieve. This will directly define the notion of class, essential in computer vision to measure performance accordingly. Image retrieval, in its most generalistic definition, defines class according to the considered dataset, *e.g.* with a dataset of animal species, we want to distinguish (*classify*) species between them, we will thus define N classes for the N different species. In some cases however, specific definitions are used and will influence the approach:

- **Fine-grained retrieval** defines classes with details smaller than objects, such as birds with certain tail features that distinguishes them from close, but different species. In this case, we want to retrieve only the exact same species.
- **Instance** or **particular object retrieval**, defines classes with an unique object, *e.g.* a person or a landmark.
- **Pattern retrieval** defines classes with an unique visual pattern, independent from the notion of object. For example, in paintings or achitecture, repeated patterns can be found across different works, with little variation.
- Retrieval-based **visual localization**, a task where the final goal is to compute the pose (exact position and orientation of the camera used) of a given query using a database of reference images, defines classes with an unique geolocation (and adds the orientation as an additional goal), which might evolve through time with moving objects.

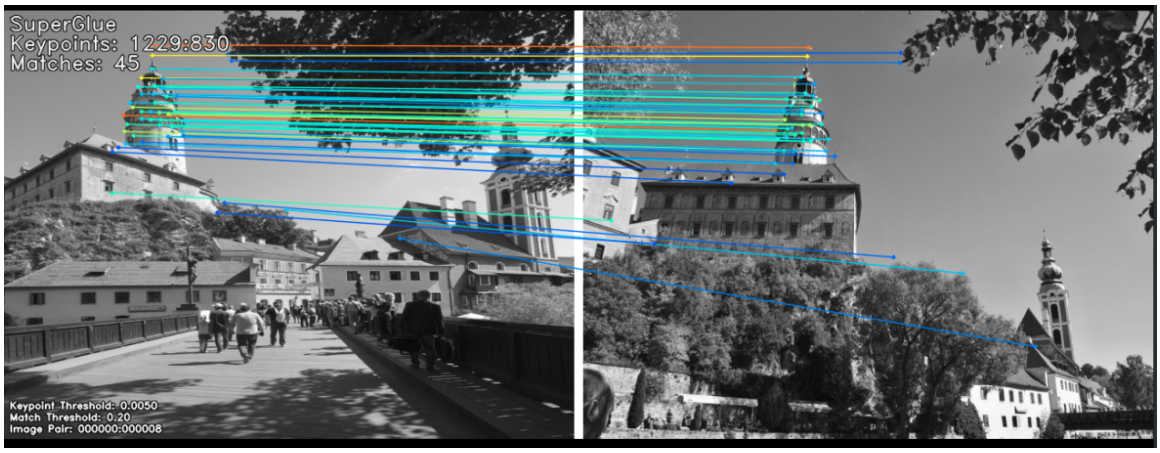

Figure 3.8: Image matching principle: local points on the image on the left are matched to local points on the image on the right.

There is a link with **image matching** worth mentioning. While image retrieval corresponds to a 1-to-N similarity measure (comparing a query to the database), image matching is an "easier" version of comparing images 1-to-1. The difference lies in how these tasks are motivated and consequently evaluated, image matching often involves precise point-to-point evaluation on local features to verify that each association is correct (see Figure 3.8) and build stable 3D models (Structure-from-Motion task), while image retrieval usually only considers the image-level label as an indicator of positive or negative. Accordingly, image matching descriptors have special properties maximizing repeatability (detecting the same real-world keypoint on different images) without necessarily maintaining the properties expected in image retrieval (discriminativeness, compactness). This can explain why they are generally not included in the evaluation of image retrieval or visual-based localization methods [65, 73].

Note that if we repeat image matching N times, it is the same as image retrieval. While this would be in practice computationaly expensive, it is in fact not so different than what is done in image retrieval, where database descriptors are computed and indexed in an offline step without particular time or resources limitations. Going-further, the idea of exhaustively comparing images in a database in a N-to-N manner can be referred to as **image clustering**. Some works exist within this formulation [14], but this is not a well-known task in computer vision.

Image retrieval aims at associating a given query with all other images belonging to the same class. This association is expressed through a list of results, that ranks images based on a decreasing similarity criteria. It is only for measuring performance that we need to precisely define separate classes with one (or multiple) labels. Computer vision, especially recently (see Chapter 2.3.2), is particularly influenced by the task of **classification**, which consists in associating labels to images, generally relying only on knowledge gained from example images seen during training. This idea of

assigning labels to data is also expressed through object localization (classification + a bounding box), object detection (object localization on multiple objects), segmentation (classification on pixels)... ; approaches that can be grouped with the keyword of *recognition*.
The main difference between classification and retrieval lies in how labels are used:

- Recognition starts from a predefined set of labels (or classes) and assigns them to data points, i.e. there is a known and finite set of labels when designing the model.
- Retrieval associates data points together in a list, then uses labels to measure the quality of this list, i.e. there is no known set of labels when designing the model.

With this point of view, note how recognition (particularly classification) can be seen as a special case of retrieval. The fundamental difference here is that recognition depends on a predetermined context, whereas retrieval uses a more generalistic approach that will produce different results depending on the data that is considered, and resorts to the predetermined context only for evaluation. It thus seems promising to invert the usual view that retrieval borrows from recognition, considering the "bigger picture" which is establishing computer vision systems that conduct tasks with as little human supervision as possible. Moreover, as the volume of data continues rising, it is worth questioning if said predetermined context will always be available. A data-driven approach seems a better fit to handle the challenges brought by the "Big Data" paradigm.
Some works have already reached a similar conclusion [51] and begun paving the way in this direction [111, 100].

In ALEGORIA, following discussions with "user" institutions, two use cases were defined:

1. Connecting images depicting the same location, with a broad definition of location, and real-time execution to treat new queries. This corresponds to a mix between instance retrieval and visual localization, where an exact building or GPS position is not expected but rather any image containing at least partially the same objects or zones. If the real-time constraint is relaxed, the problem can be reduced to image matching or image clustering.
2. Connecting architectural works containing the same patterns. This corresponds to pattern retrieval.

This thesis focuses on definition 1. It is the driving application in ALEGORIA, i.e. connecting images through databases to propagate metadata, and it can include definition 2. by cropping query images to the pattern of interest.

Considering these related tasks, and especially their dependence on a well-defined context (and training data), image retrieval seems the most appropriate approach to the cross-source connection problem.

The typical framework for image retrieval is presented in Figure 3.9. It is usually separated in two parts:

- In the Offline stage, features are computed for the whole image database. They are stored in an index, *i.e.* an optimized data structure linking features with their source images in a way that allows fast retrieval.
- In the Online stage, features are computed for the query image using the same algorithm as in the offline stage. The query features are then compared with a **similarity measure** to features stored in the index, which gives a list of the N most similar images (the nearest neighbours in the space of similarity), the list of results. Optionally, it is possible to add a diffusion step, where this list of results is improved through additional nearest neighbour searches.

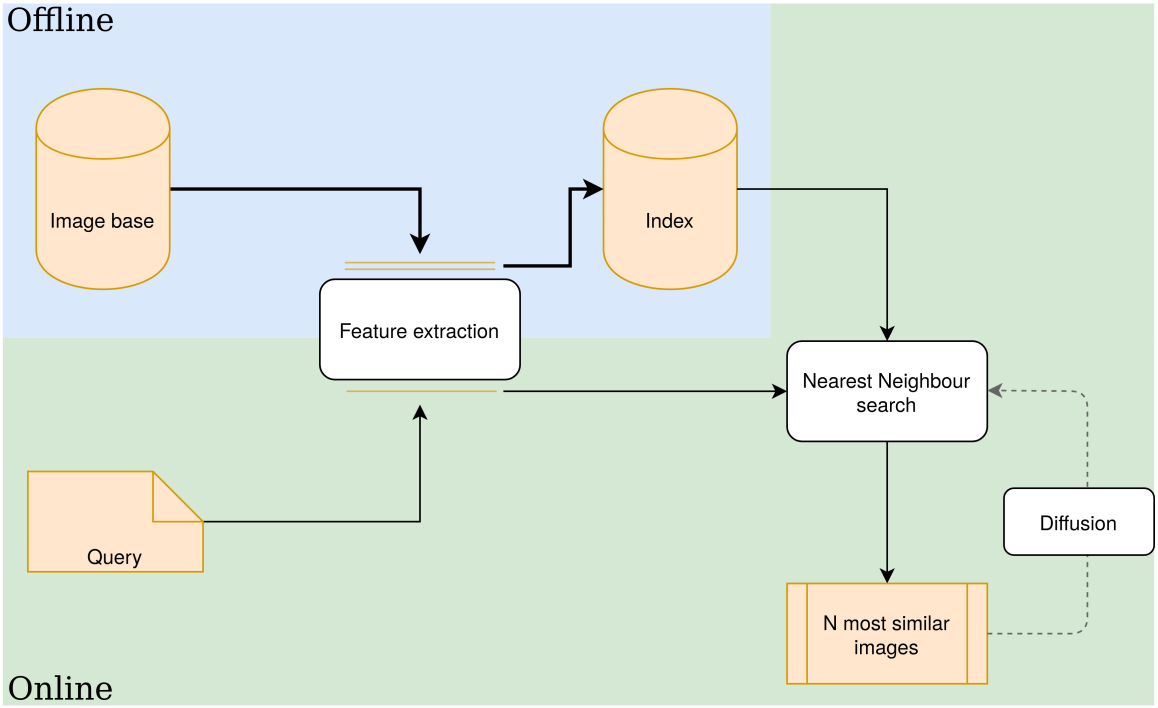

Figure 3.9: Standard image retrieval framework. Data is depicted in orange, computational steps in white.

The literature review in the following section will present some research problematics around these elements.

3.4 Problems and methods: related works

In this section I will give an overview of methods available in the literature around image retrieval, in the context of deep learning, along with some ideas borrowed from parallel problems. To conceptually organize this review, I propose to browse three research axis, and each axis will be presented along its two directions to get a grasp of the fundamental ideas and contributions of each side.

The attentive reader will notice that, while the term "domain" is occasionally used in this thesis, I do not include a systematic review of cross-domain approaches. To my knowledge, there hasn't been any work broadening the notion of domain, *i.e.* not restricting visual variations to a single axis, even in topics such as domain generalization. Rather, there is on one side domain-related propositions working on well-defined problems, and on the other side generalist approaches that abstract away the notion of domain to concentrate on absolute performance, regardless of the characteristics of input images. Image retrieval belongs to this second family of methods.

3.4.1 Global-local axis

The first dichotomy that might come to the mind of a person knowledgeable in content-based image retrieval is between local and global features. Recall the various ways in which image retrieval can be defined presented in section 3.3, these definitions imply some sort of variation on the *locality* of information on images. Accordingly, it seems logical to produce features that describe only the relevant zones. This idea has influenced image retrieval since its apparition, when handcrafted features (calculated with a previously defined formula) were the norm. With deep learning, the separation was kept but as we will see in this section the consequences on performance are still not well understood.

Let us begin with some notations: methods using deep features rely on a backbone CNN, applying a function f_i to the input image I (depending on what layer i is considered), and producing a tensor of activations T_i : $T_i = f_i(I)$, with width W and height H (see Fig. 3.10) depending on the dimensions of the input image, and depth D depending on i. The following methods will be presented along this guideline of how T_i is handled, giving either a local or a global descriptor.

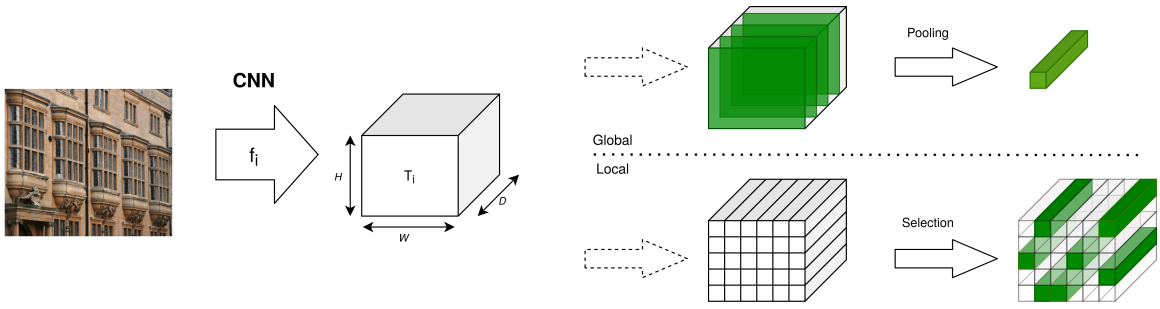

Figure 3.10: Principle of deep local and global descriptors.

The backbone network architecture, characterizing the function f, can be any CNN. Radenovic et al. compared global descriptors with varying backbones (AlexNet, VGG, ResNet) on $\mathcal{R}$Paris and $\mathcal{R}$Oxford , all other hyperparameters equal, and noted a significant disadvantage of AlexNet but varying results with VGG and ResNet. Since then, it seems that ResNet has became the "go-to" backbone, undoubtedly because it is a smaller network than VGG (thus easier to train) [97, 13]. I will follow the same standard in the following, but note that VGG must still be considered as a candidate backbone on real-world applications since there hasn't been (to my knowledge) extensive evaluation proving its inferiority.

Global: compact and semantic descriptors

Global methods describe an image as a whole, embedding all important information in a single vector. This is conceptually closer to the classification task for which common architectures like Resnet [33] or VGG [87] were designed. Babenko et al. [7] indeed showed that simply taking intermediate features from a classification network and using them for retrieval yields good performance.

To handle varying sizes in images and allegedly get more invariance, a standard seems to have emerged in deep global descriptors : extracting information at one of the last layers with a pooling process giving one value per channel. Following our notation, here the tensor T_i is seen as a set of D activation maps that contain each a different type of highly semantic information. To get a global descriptor, a straightforward approach is thus to get the most meaningful value per channel. We can then compare two images simply with dot-product similarity of their descriptors. See figure 3.11 for an example with the MAC descriptor detailed below. Babenko and Lempitsky [6] propose to sum the activations per channel, establishing the SPoC descriptor. This is equivalent to an average pooling operation. Differently, Kalantidis et al. [41] tweak the SPoC descriptor with a spatial and channel weighting, while Tolias et al. [96] get better results by using the maximum value per channel (MAC descriptor). They also propose the regional version RMAC, by sampling windows at different scales and describing them separately. Radenović et al. [74] generalize the preceding approaches with a generalized mean pooling (GeM) including a learnable parameter.

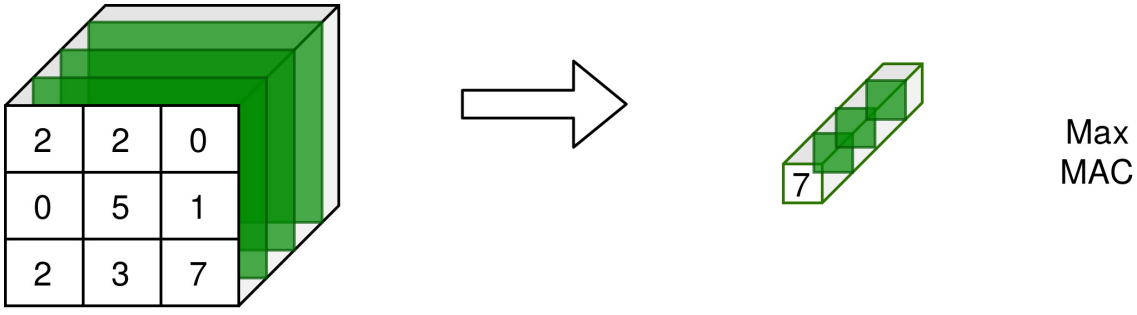

Figure 3.11: Principle of MAC global descriptor.

All of the previously presented methods have the advantage of being fully differentiable. This allows efficient fine-tuning on relevant data, with an objective function that encourage discriminability.

The dimension of global deep descriptors is fixed by the backbone output channel number (2048 for ResNet50), but can be reduced with a whitening layer [74], PCA, or Product Quantization [38]. Tolias et al.[97] compares the total memory usage of some methods, showing roughly similar memory usage for the top global and local descriptors but with high disparities, *e.g.* a GeM global descriptor reduced with Product Quantization yielding 97% of its original performance while using only 2% of its original size (0.2 Gb, to compare with the 15.3Gbs and 14.3Gbs of respectively the best global and local descriptor).

Local: geometry and efficiency

Local methods use a set of carefully selected points on an image. This involves identifying points that maximize invariance, and describing small patches around these points to extract information.

The SIFT local descriptor [57] is a good example of a computational pipeline for handcrafted local features:

1. The input image is passed through a series of gaussian filters, at different scales. The gaussian filters are parameterized with a σ parameter (standard deviation) serving as a scale factor.

2. The resulting gaussian-filtered images are substracted pairwise to build a pyramid of gaussians (see Figure 3.12). The difference-of-gaussian (DoG) images obtained this way capture visual details at different scales parameterized with the different scale factors.

3. Local minima and maxima values of the DoG images are identified to produce candidate keypoints.

4. Candidate keypoints are processed and filtered to refine their position, discard low-contrast keypoints, discard sets of keypoints on edges, and assign orientations. The resulting keypoints are considered valid for description.

5. In a local neighborhood around each keypoint, information is extracted as a histogram of gradients, each bin giving the magnitude of gradients in a direction. To ensure orientation invariance, orientations of gradients are normalized using the assigned keypoint orientation. The final feature is a concatenation of the histograms of the 4x4 region around the keypoint, with 8 bins each, leading to a 128-dimensional descriptor for each keypoint.

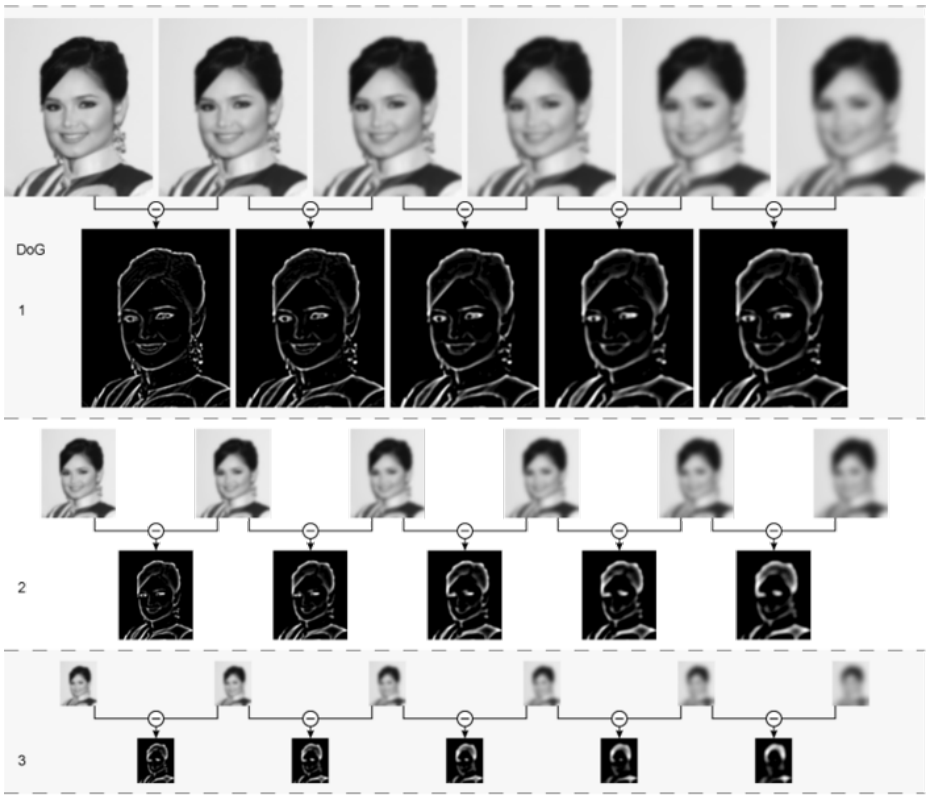

Figure 3.12: Pyramid of gradients for SIFT keypoint detection
Photo credits: amrfum, *Derivative image credits*: Indif, CC BY 2.0

Early deep methods replaced parts of the historical handcrafted pipeline by trainable pieces. Verdie et al. [102] designed a learnable keypoint detector maximizing repeatability under drastic illumination changes. Yi et al. [115] extended the architecture for full detection and description. However these two methods rely on multiple computational steps that are separately optimized, which leads to sub-optimal results.

Noh et al. [63] solved the issue with a pipeline producing a set of local descriptors in a single forward pass, and that can be trained directly on any dataset with image-level labels. In their method, T_i is seen as a dense grid of local descriptors, where each position in the activation map is a D-dimensional local feature, whose receptive field in the input image is known. Additionally, an on-top function assigns a relevance score to each feature, and a threshold is set to only select most meaningful features. The output is a set of N DELF descriptors per image. See Figure 3.13 for a visual interpretation. This departs from the traditional detect-then-describe process by selecting points *after* describing them, but it is simple to train and has shown good results on standard benchmarks [73]. Note that D typically ranges from 512 to 2048 in the last layers of the CNN, hence the PCA reduction to $D = 40$ proposed by the authors. Recently, DELF was revisited and studied

in the context of a state-of-the-art matching technique (ASMK, described later in this section), leading to the similar HOW local feature. For these methods, optimization is done with aggregated representation and image-level classification labels, which is still not optimal because it limits the control that can be applied on the properties of local descriptors.

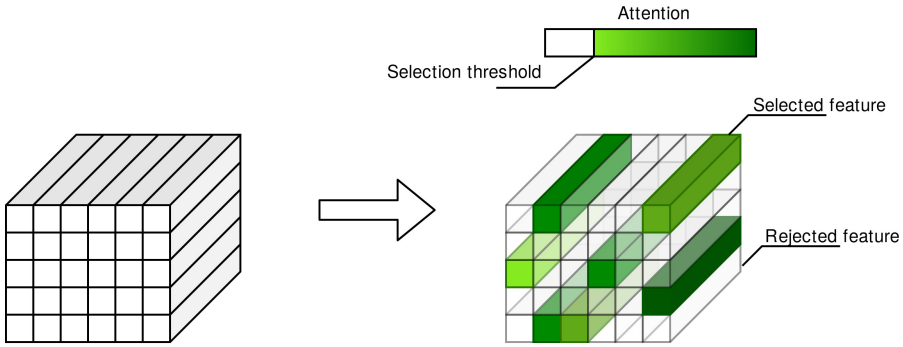

Figure 3.13: Principle principle of the DELF descriptor

The similarity measure on two images described with local features is a non-trivial N-to-M matching problem (see Figure 3.8 for a visual example of what is desired). Some points in image A might not have correspondences in image B, or multiple correspondences, and vice-versa. A simple solution is to first build correspondences with the euclidean distance, then use a criterion to only keep at most one correspondence per local feature, for example by comparing the distance with the first candidate and with the second candidate. If the nth feature in image A is very close in the euclidean space to the mth feature in image B, but far from any other feature, we can confidently assign a correspondence between those two keypoints. On the contrary, if the nth feature in image A has a similar distance to multiple features in image B, we can consider that it cannot be matched.

In image retrieval, such a pairwise matching approach would be too computationnally expensive for large databases. With hundreds of keypoints per image, and thousands of images, the space of local features becomes cluttered, making it delicate to rapidly establish a similarity measure between two images. The bag-of-words paradigm has brought a comprehensive solution to this problem [89]. Borrowing from text analysis, the idea is to identify "visual words" by clustering local features: groups of local features close in the description space are merged into single features (visual words). For example, with the SIFT descriptor a typical visual word would consist of a shape commonly found on images, such as corners. From image to image, corners are detected and assigned to the same visual word, even if they have slight visual differences. Images are represented in the form of a sparse vector, the bag of words, indicating which visual words were found in the image(see Figure 3.14 for a representation). The bag-of-words paradigm also has the advantage of allowing efficient indexing with an inverted index structure, *i.e.* a file linking each visual word to all images containing it. The measure of similarity can either be computed by directly comparing sparse vectors

(with the example of Figure 3.14, we would have $f(A) = (2, 1, 1)$ and $f(B) = (1, 2, 0)$, with $f(A)$ and $f(B)$ the vectors representing the image on the left and on the right respectively. Without any sort of normalization, that would lead to a similarity score of $(2, 1, 1) \cdot (1, 2, 0) = 4$), or by iteratively increasing the similarity score of images assigned to common visual words.

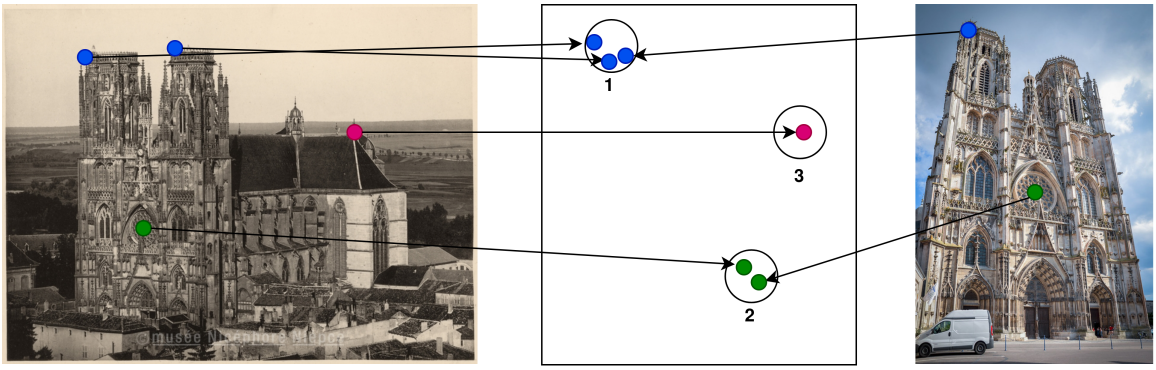

Figure 3.14: Simplified example of bag-of-words matching. The image on the left is represented as a vector stating that it has 2 occurences of visual word 1, 1 occurence of visual word 2, and 1 occurence of visual word 3. The image on the left is represented as a vector stating that it has 1 occurence of visual word 1, and 1 occurence of visual word 2.
Photo credits, left: Nicéphore Niepce, see Alegoria rights, *Photo credits, right*: Zéphyrios, public domain.

Starting from the bag of visual words, multiple ideas have been proposed to enhance it[1]:

- In the original paper [89], authors additionally propose to use the TF-IDF (Term Frequency, Inverse Document Frequency) weighing scheme used in text retrieval to balance the influence of frequently occurring visual words in an image (TF) and the influence of frequently occurring visual words in the database (IDF).

- Jegou et al. [39] explored the trade-off between using a small vocabulary (more robustness to visual changes) and a bigger vocabulary (better separability, or discriminativeness), and introduced an additional distance measure for descriptors assigned to the same visual word. The proposed Hamming Embedding (HE), combined with the bag-of-words approach, combines the advantages of a coarse codebook (visual vocabulary) and a finer descriptor matching method.

- Instead of browsing visual words one by one, which is necessary for large vocabularies, it is possible to build a compact representation encoding more information than the simple sparse vector mentioned above. This is the idea of VLAD [40] (Vector of Locally Aggregated Descriptors), which represents an image by a concatenation of k vectors, each vector being the sum of residuals between a feature and its assigned visual word. This leads to a $k * d$ representation, with d the dimension of local features. Arandjelović et al. [4] mimic VLAD

[1] Mostly by Hervé Jégou and Matthijs Douze, to whom I am grateful for their contributions in image retrieval in the last 15 years and their always relevant and clear publications

with a learnable pooling layer, giving NetVLAD. By replacing the hard assignment step with soft assignment to multiple clusters, they can train this layer. In this work, T_i is considered, as in DELF, as a block containing $W * H$ D-dimensional descriptors.

- Tolias et al. [95] proposed a generic framework to describe how matching methods perform the similarity measure, and introduced a method synthesizing the ideas proposed in previous works, the Aggregated Selective Match Kernel (ASMK). This method seems to have been kept as a standard until now, used for example as a basis for HOW [97] or by Teichmann et al. [92] who propose a regional version of ASMK. See Appendix B for a mathematical formulation.

Finally, the also recent work of Siméoni et al. [88] proposes a new point of view on activations, following the observation that shapes of objects of interest in the input image can be found in some channels of T_i. They perform detection and description of interest points in T_i using a handcrafted detector (MSER [61]), and then match images based on spatial verification. However, this method is not suitable for large-scale retrieval since it works on pairs of images.

3.4.2 Recall-precision axis

Starting from our formulation of the cross-source problem as an image retrieval problem, let us look at the output of the framework: the list of results. The list is typically of arbitrary length (there is normally no need to include all images in the database in the results), and sorted in a decreasing order of similarity. There are two ways in which information is expressed through this list: the absolute positioning of a given image (*e.g.* it is the 3rd result) and the relative positioning of a given image compared with other images (*e.g.* this image is 3 ranks higher than this other image).

Consider Figure 3.15 which gives some examples of possible lists of results, with length 5. First row shows the best scenario, second row the worst. Now, there remains 30 ways in which these 5 first results can be arranged. Third and fourth rows show two possible ways that would be satisfying in our retrieval problem: we retrieve multiple positive images. These two rows however highlight a different way of positioning these results. Third row shows a behaviour typically expressed by a system with high **precision** (or selectivity or specificity): images at the top of the list are confidently positive (probably mostly easy positives), while other positive images can be mixed in the rest of the list. Fourth row shows a behaviour typically expressed by a system with high **recall**: positive images (easy and hard positives) are, in average, comparatively higher on the list than negative images (recall that there are 5 positive images and only the top 5 results are showed here), even if some negative images can be mixed in the first results.

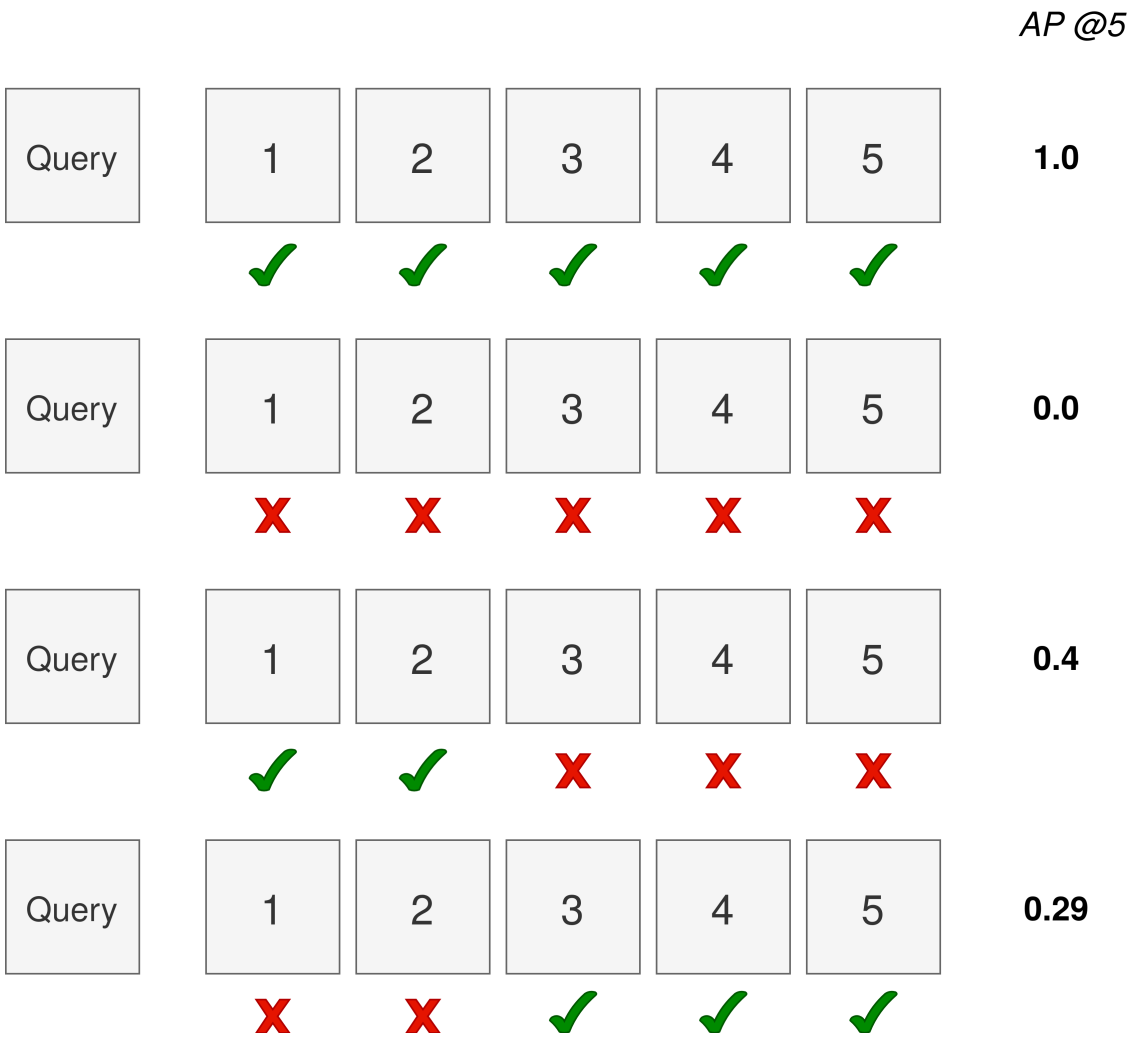

Figure 3.15: Example of results layout for a search in image retrieval. AP@5 scores (ignoring the rest of the list) are indicated on the right, assuming there are 5 positive images for the considered query.
Computation of AP@5 scores:
Row 1: $AP = (1/1 + 2/2 + 3/3 + 4/4 + 5/5)/5 = 1.0$
Row 2: $AP = (0 + 0 + 0 + 0 + 0)/5 = 1.0$
Row 3: $AP = (1/1 + 2/2 + 0 + 0 + 0)/5 = 0.4$
Row 4: $AP = (0 + 0 + 1/3 + 2/4 + 3/5)/5 = 0.29$

The mean Average Precision score (mAP), used to synthesize the quality of the list of results in a single value, is computed by averaging the AP score over a set of queries. The AP score is computed as follows for a list of N results:

$$AP = \frac{\sum_{k=1}^{N} \text{precision}(k) * \text{relevance}(k)}{|P|} \tag{3.1}$$

precision(k) corresponds to the fraction of positive images in the list from position 1 to k[2], and relevance(k) is an indicator being 1 if the image at position k is positive, 0 otherwise. P is the set of all positive images.

Precision and recall are two different ways of enhancing the list of results, *i.e.* the mAP score and corresponding approaches differ in their formulation. In this section, I will present methods available in the literature, using this dichotomy.

Precision: reading the list from the top

Raising precision is a quest for discriminativeness: the ability to separate positive and negative images with high confidence. This is a goal similar to what is expected in object recognition: in single label classification, we need the positive class to be the one with the highest probability. It is thus not surprising that deep feature-based image retrieval has heavily borrowed from classification in its formulation, and was initially performed simply by taking features extracted from CNNs trained to classify images [7, 6].

The loss function in particular is essential to enforce the discriminability of descriptors. Recall from Section 2.3.1 the idea of feature learning: we need features depicting the same reality to be close in the feature space, and features corresponding to different realities to be far. The loss function is responsible for expressing this objective. Some examples of loss functions that have been proven to lead to good descriptors are:

- Standard cross-entropy loss used in classification (see Figure 3.16. This loss function only enforces a high probability for the positive class during training, and is thus by definition sub-optimal when there is potentially inter-class similarity, but nonetheless has been used with success to train local descriptors [63]. A drawback of classification losses for feature learning is that they need an additional classification network to output labels.

- ArcFace loss [20], an enhanced version of the cross-entropy loss learning class centroids while training and using an angular margin. It is supposed to enforce better inter-class separability, and has indeed lead to better global descriptors [13].

- Contrastive learning (or tuplet losses), *i.e.* building n-tuplets of images and enforcing the convergence of positive pairs and the divergence of negative pairs. The easiest version corresponds to the triplet loss (see Figure 3.17: one anchor image, one positive image (an image from the same class), and one negative image (an image from another class). With the contrastive loss, an arbitrary number of positive and negative images can be used [74]. Note that this is less

[2] A clarification is needed regarding the term "precision". It can be understood as a quality of the model, which is the definition I use to guide this literature review. In its mathematical definition, it is a proportion of positive images in a given list, which is a bit different. The difference lies in k, the ranking at which it is computed. By "high precision", I mean "a high mathematical precision value for small values of k", *i.e.* at the beginning of the list. It would not make sense to compute precision for high values of k, since there are far more negative than positive images.

computationaly efficient than the formulation with classification losses, and is often associated with a form of epoch mining: before each epoch, hard negatives (negative images with high similarity) and hard positives (positive images with low similarity) are identified and provide harder training batches leading to better discriminability.

- Listwise losses. Recent works have derived differentiable approximations of the AP metric, allowing direct optimization of the evaluation measure [12, 77]. It can be seen as a generalization of the contrastive loss to multiple classes: if we have k examples distributed in N different classes, we can construct a training batch where each image has $k-1$ positives and $(N-1)*k$ negatives. This is arguably more consistent with the testing setup of image retrieval, but in reality, results do not show a clear difference with classification or tuplet losses [106].

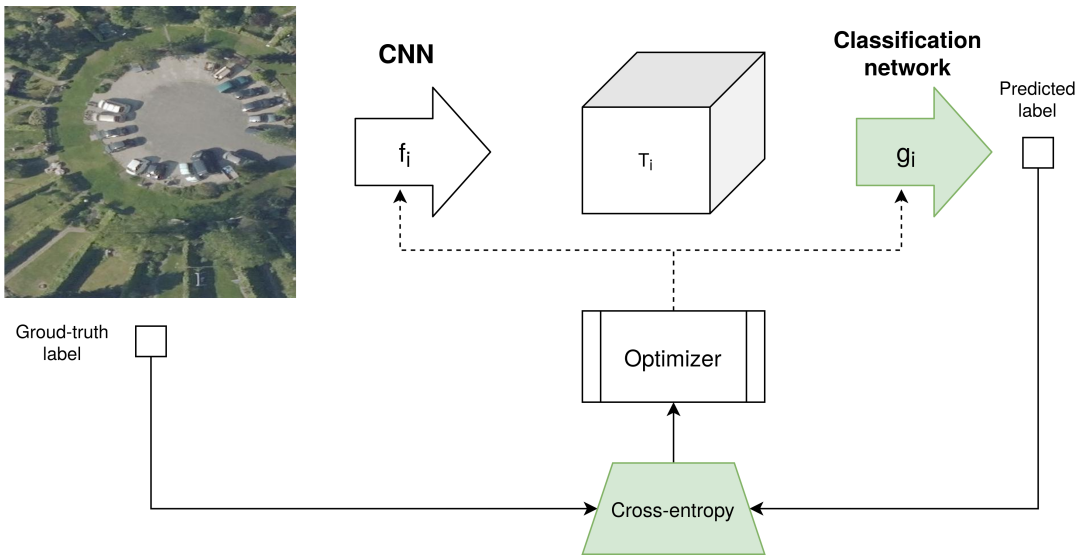

Figure 3.16: Principle diagram of learning deep features with the classification loss. The ground-truth label is compared with the predicted label.

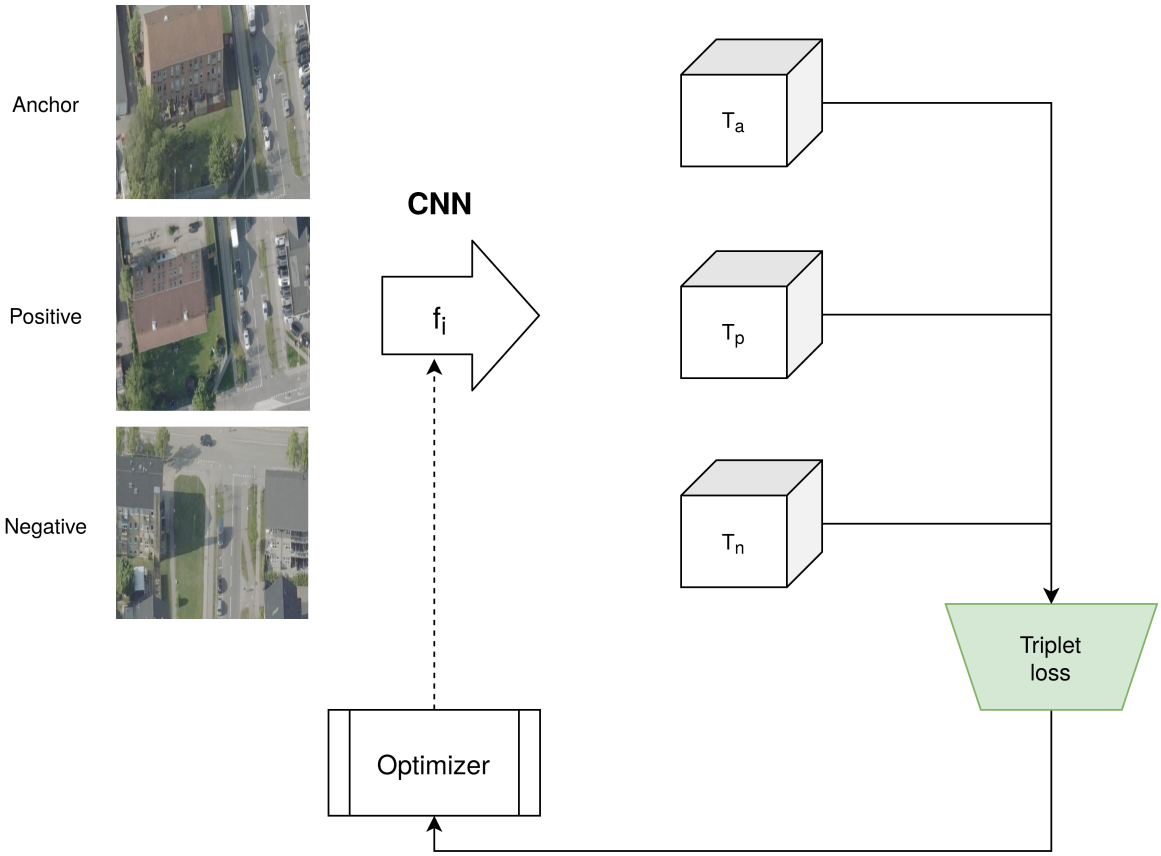

Figure 3.17: Principle diagram of learning deep features with the triplet loss. Image features are compared pairwise, pulling the anchor and positive together while pushing the anchor and negative apart.

Assuming that good descriptors are available, it is also possible to raise precision with additional computational steps. Ideas include:

- Simply using higher dimensional descriptors (works both for global and local features). In deep learning, this corresponds to using bigger backbone networks, or decrease the dimension reduction typically used for large-scale setups. Obviously, the computational capacity will quickly be the limiting factor for this kind of approach, but distributed setups can alleviate this [45].

- Concatenating different descriptors (works only for global descriptors). It is an approach similar to using higher dimensional descriptors but not limited to the output dimension of the backbone feature extractor. This idea has shown good results on large-scale retrieval setups [67] but is of limited interest considering the high computational cost, it is a "brute-force" approach with low margin for improvement.

- Geometric verification. Local descriptors have the advantage of recording their original position on the corresponding image, which allows for a post-processing step where the query image and the candidate images at the top of the list can be compared through a set of points, to check there is a consistent geometric transformation leading the object of interest from the query to the candidate. This is usually conducted with the RANSAC algorithm [29].

Recall: reading the list from the end

Looking back at Figure 3.15, note how, as shown by the last two rows, the mAP measure commonly used to evaluate image retrieval has a tendency to favor precision rather than recall (even if, despite its name, it is a measure including both precision and recall in a single value). Especially for classes with few positive images, the absolute positioning of these few images has a big influence on the AP score. See Appendix A for a more detailed example of this phenomenon.

This behaviour is not necessarily what we aim for in the ALEGORIA case: retrieving multiple positive images in say, the top 20 results, can be in certain cases more interesting than having a positive image in the top 3. This formulation of the problem is actually closer to what is found in visual localization, where a commonly used metric is the recall at k ($R@k$), *i.e.* the percentage of queries with at least one relevant image among the top k results.

A key property associated with higher recall is **robustness**, *i.e.* how resilient the descriptor is to the visual variations. For handcrafted descriptors, robustness was directly enforced through their calculation, for example SIFT [57] is scale- & rotation-invariant by design. For deep features, there are two ways of enforcing invariance:

- Explicitely, by choosing differentiable operations that are invariant to variations. The current standard for feature extraction, a common CNN such as ResNet or VGG without fully connected layers, only has a translational equivariance property (due to the convolution operation). There is no guaranteed robustness to illumination, rotation & viewpoint change, scale. Some works have proposed to modify common architectures to enforce such properties [108, 37], but they involve a computational overhead and an algorithmic complexity that make them troublesome to adapt to other tasks.

- Implicitly, through training. While the architecture does not guarantee invariance to many transformations, some operations offer potential for learned invariance through training, notably pooling and selection operations. The max pooling (or the similar GeM pooling) operation commonly used for global descriptors, can bring invariance to all of the above-mentioned variations: by selecting only one value per channel or per region, any modification done on input pixels can be ignored, under the hypothesis that the training process has lead to convolution parameters that will select and promote meaningful information. Similarly, the selection operation used for local descriptors, performed by keeping the local features with the top k

scores of attention, can discard unwanted variations if the attention scores are relevant enough.

Taking a step back and looking at the literature around feature learning, no proper conclusion emerges regarding the choice of a robust architecture or rules to follow to build invariant descriptors. State-of-the-art methods with top performance on $\mathcal{R}$Paris and $\mathcal{R}$Oxford do not mention this issue, even if doubts can be raised about the hypothesis that descriptors can learn invariance through training [80, 11].

Recall depends by definition on the number of positive images for a given query. It can be seen as an indicator of how well the model has built an understanding of the considered object (the *class*), and how it uses it to retrieve all images corresponding to this class. This understanding of the object is expressed, in machine language, as descriptors that are placed in a N-dimensional space. Taking global descriptors for simplicity, the standard single search consists in comparing the query descriptor with all other descriptors of the database, one by one, to get a similarity score for each database image. This is efficient, but does not quite correspond to building an understanding of the object: it is merely a comparison of a vector with other vectors. A more complete approach is to consider already available information, *i.e.* database descriptors, to propose a similarity measure that takes into account the relations between different descriptors. The result is a better list of results, potentially able to retrieve images that are very different from the query image. This process is called **diffusion**, but can also be referred to as query expansion or re-ranking in the literature.

Figure 3.18 shows a case where diffusion is promising. Suppose we query the system with an image on the first or on the third row. Due to their higher visual similarity, images in the same row will likely be higher on the list of results than other images. Considering that there is no common visual information between the first and the third row (the top of the Arc de Triomphe is not visible on the first row, and is the only visible part on the third row), there is a low chance that the system will successfully pair images on these two rows. But if we take into account the visual and semantic information available with images on the second row (oblique images, *i.e.* we see both the top and the sides of the Arc de Triomphe), the task becomes easier: we can match an image on the first row with an oblique view on the second row, which will have higher similarity with vertical images on the third row.

Diffusion is thus a way to bridge the gap between single-step image retrieval (quickly retrieve most similar results) and database-wide image clustering (build links in the data, organize images in class groups). Concretely, this can done by using the first N results of the initial search to re-compute descriptors (a process also referred to as Query Expansion - QE). Early methods simply combined the query descriptor with the descriptors of the first results by averaging them [18, 6, 41, 96], and have been improved later with α-weighted query expansion where similarities are used as weights when averaging [74]. These approaches have the benefit of staying relatively economic in computational overhead, while bringing significative improvements in many setups. More advanced methods go

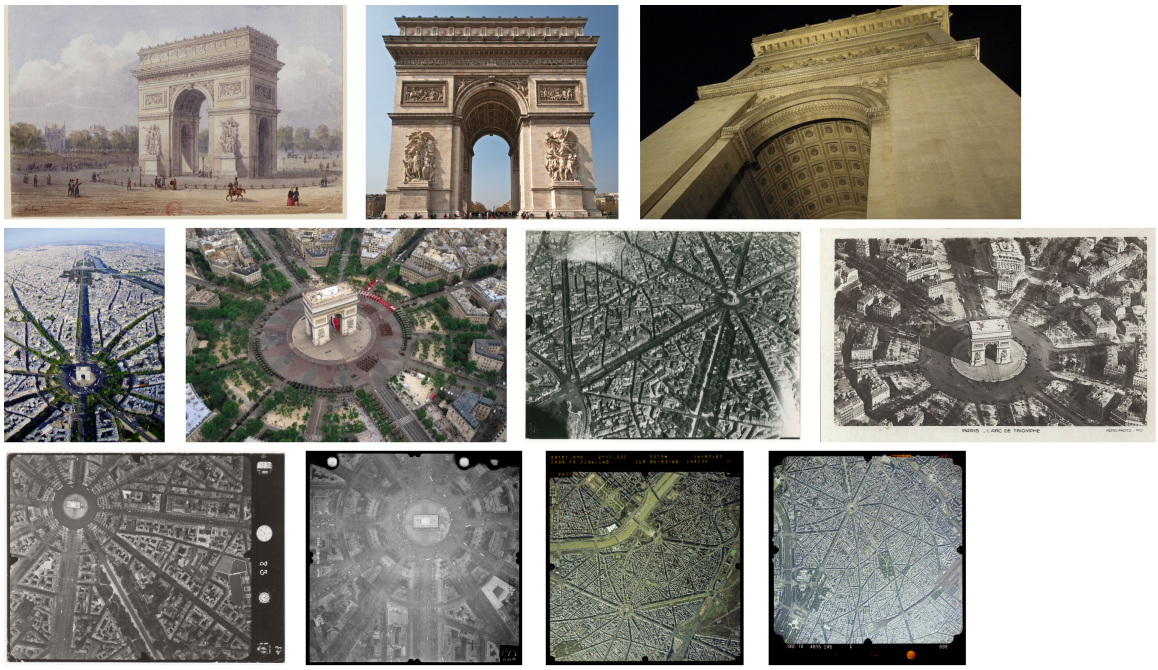

Figure 3.18: A situation from the ALEGORIA benchmark where diffusion might improve recall. Examples are from the same class, with a different viewpoint category on each row (first row: ground-level, second row: oblique, third row: vertical).

further by exploiting the affinity matrix, defined as the set of pairwise similarities between a given image and all other images. For a database Ω, a similarity matrix S is calculated with pairwise similarities :

$$s_{i,j} = sim(d(x_i), d(x_j)), \forall (x_i, x_j) \in \Omega \tag{3.2}$$

sim can be any measure of similarity, depending on the considered descriptor d (cosine similarity, euclidean distance on global descriptors, matching kernels [95] on local descriptors...). This matrix can be interpreted as a graph [119], where each image is a node, and edges linking nodes to each other are weighted using the corresponding similarity S_{ij}. In its simplest version, the adjacency matrix is all ones since each image is connected to each other image:

$$a_{i,j} = 1, \forall (i, j) \in [1..|\Omega|]^2 \tag{3.3}$$

Diffusion is conducted by updating node features or directly S using an update rule. [24] presents a set of strategies regarding this update rule. This modular setup allows offline computation [113], region-based variations [36]. Note however that using all neighbors (A from eq. 3.3) is computationally expensive considering the $|\Omega|$ node updates depending on the $|\Omega|$ adjacent nodes. A simple but effective way to avoid this systematic computation is to filter edges to only keep the k nearest neighbors:

$$a_{i,j} = \begin{cases} 1 & \text{if } x_j \in NN_k(x_i), \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

with $NN_k(x_i)$ the k nearest neighbors of x_i according to S. [122, 119] show good results with a similar setup going further using reciprocal nearest neighbours.

3.4.3 Supervision axis

Without surprise, the deep features now widely used in image retrieval have inherited all the hypotheses coming with the use of CNNs, notably 1. a large training dataset is available, 2. the testing dataset has semantics close to the training dataset. The ALEGORIA case is an example where these assumptions cannot be met, it is a **low-data** scenario. Here I will present some methods to handle the gap between currently available training data and ALEGORIA images, or more generally any collection of images for which there is no corresponding, well-defined training dataset. The goal of uncoupling test performance from training data has given rise to a variety of research topics, that can be sorted as shown in Figure 3.19 along the supervision axis.

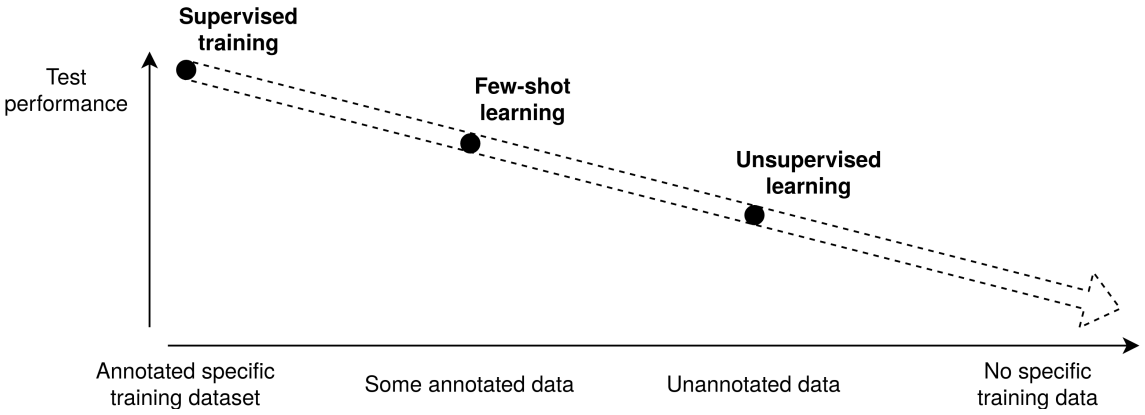

Figure 3.19: Overview of research topics along availability of training data. Typical learning setups differ according to their data-dependency (x-axis), with expected test performance (y-axis) decreasing with less training data.

Supervised or transfer learning

Supervised learning corresponds to optimizing the backbone on a setup (characterized with a task and an associated dataset) as close as possible to the target setup, using large volumes of annotated data.

Annotations typically take the form of a single class label associated with an image as in the ImageNet dataset [19], but can also be more detailed as the bounding boxes in [106]. Annotating images is a very demanding task, requiring human expert supervision. To alleviate this constraint, a first idea is to propose semi-automated annotation tools. For example, Radenović et al. mines class samples from 3D models [74], with a fine control on viewpoint variations. The SF300 dataset presented in section 5.1 is also an example of semi-automated labelling, where only one manual annotation was needed to produce a batch of 10 to 20 classes with multiple examples.

The criteria that must be used to pick a training dataset are not defined, and to my knowledge are not discussed in the literature. The common approach is to gather a large volume of annotated data, define the associated problem, and then split between training and testing data. In our case, this is not possible, so how can a training dataset be chosen ? The commonly accepted answer seems to be a qualitative *a priori* evaluation of visual characteristics, or a quantitative *a posteriori* comparison of descriptors trained on candidate datasets. This relatively unsatisfying answer indicates that there is still room for improvement on understanding the links between training data and test performance, and constitutes a driving motivation for experiments conducted in section 4.1.

Neither the work of Radenovic et al.[73] nor the work of Noé Pion et al.[65], both influential benchmarks, have compared methods through various training datasets, all other hyperparameters equal. Yet, it has been noted that the training dataset significantly impacts the quality of descriptors,

to the point that a descriptor can performe worse than the worst compared method and better than the best compared method only by changing the training dataset [106].

Few-shot learning, meta-learning: a promising path?

A recent branch of research is investigating how to learn using a limited number of annotated data, or more broadly how to learn more efficiently (learn to learn). [105] presents a taxonomy of few-shot learning methods separating approaches exploiting prior knowledge in the data, in the model or in the algorithm. Few-shot learning is usually formulated with **episodic training**, shown in Figure 3.20. An episode is formed with a query set, on which we want to conduct a task (*e.g.* classifying cats and dogs), and a small support set that contains valuable information regarding the query set (e.g. examples of cats and dogs). The model is optimized through episodes, with the goal of making it learn how to rapidly gain accuracy on the query set using a limited support set. A baseline is for example to compute class prototypes (a single point in feature space representing a class), and assign queries to the closest prototype to guess the class [90, 103]. Note how similar this formulation is to image retrieval, and especially contrastive learning presented in Section 3.4.2: 1. well-defined class information is not available during testing 2. during training, the testing setup is mimicked by ignoring class labels 3. without class labels, we instead use a measure of similarity between images. But despite the similarity of formulation and approaches, the panel of propositions in image retrieval has not to my knowledge been tested in the context of few-shot learning (with one exception [100]). In Chapter 5 I propose a framework building on this idea.

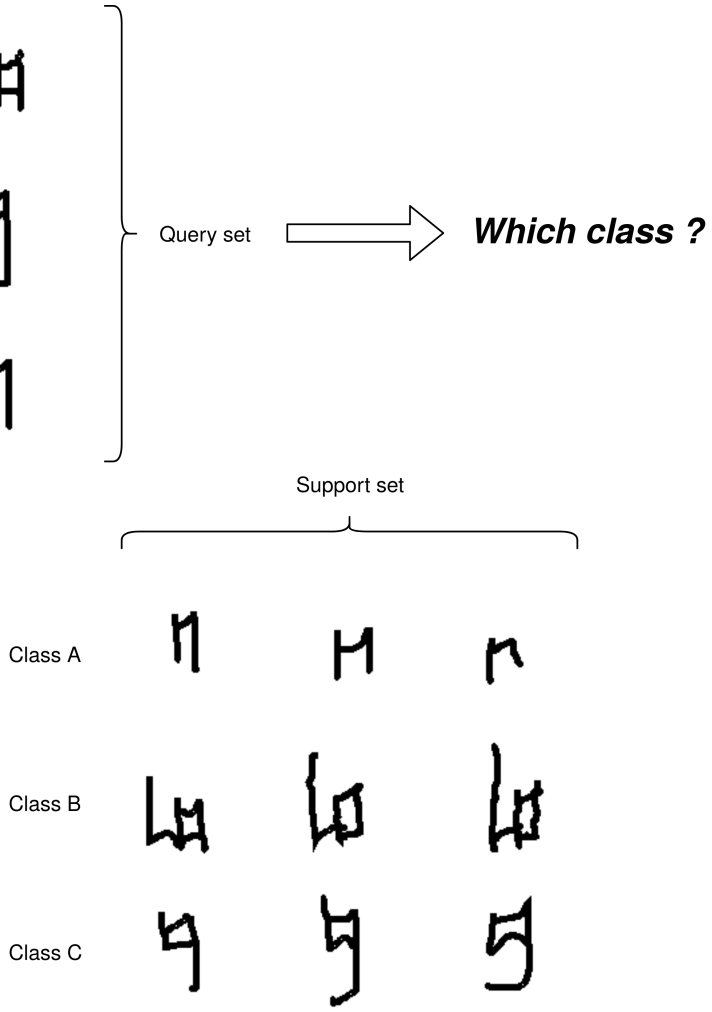

Figure 3.20: Episodic training for few-shot classification, here with 3-shot 3-way: 3 examples, 3 classes to distinguish. Examples from the Omniglot dataset [46].

An example of a promising approach is proposed by Requeima et al. [76] and Bateni et al. [9] with Conditional Neural Adaptive ProcesseS (CNAPS). CNAPS inserts Feature-wise Linear Modulation (FiLM) [69] layers in a backbone feature extractor for fast adaptation. Using episodic training and a deterministic Mahalanobis distance, the improved version SimpleCNAPS [9] trains an adaptation network (without updating the backbone) to produce adapted FiLM parameters and obtains good performance on a benchmark grouping multiple commonly used classification datasets [101]. See figure 3.21 for the principle of the FiLM layer in few-shot learning.

Note that while many interesting contributions have been made for few-shot learning, there are still significant obstacles to real world applications, including image retrieval:

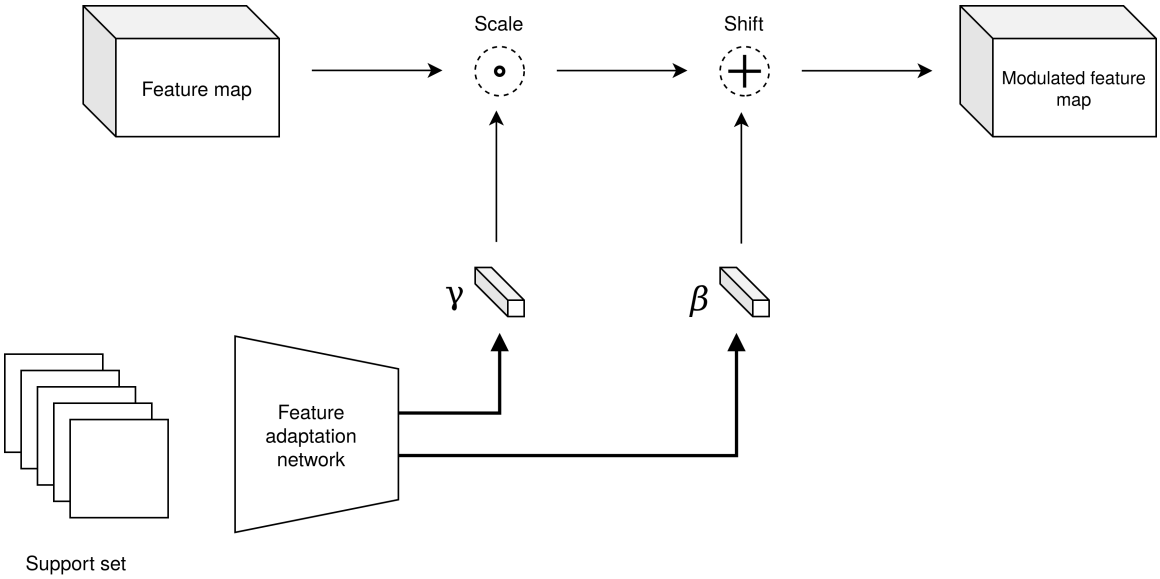

Figure 3.21: Principle of a FiLM layer applied to few-shot learning: the support set passes through a feature adaptation network, producing γ and β parameters that will respectively scale (the dot represents the Hadamard product, or channel-wise multiplication) and shift (the cross represents channel-wise addition) feature maps at different levels of a convolutional neural network.

- To our knowledge, there has been no work applying the principles of few shot learning to image retrieval (rather, few shot learning has borrowed ideas from image retrieval [100]). It is indeed a tricky situation in the sense that image retrieval does not use class information (except when evaluating), and thus defining a support set is not trivial.
- Most of the existing works have been tested on simplistic datasets with low resolution, few problematic variations and semantically easy classes, e.g. the standard MetaDataset [101] grouping ImageNet, Omniglot (character recognition) [46], VGG Flowers (flower recognition) [62], Traffic Signs (traffic signs recognition) [34] and other relatively simple datasets.
- Due to the computational overhead induced with meta-learning architectures, small feature extractors such as ResNet-18 or its even smaller version ResNet-12 are used [22] to avoid memory limitations. This naturally also limits final performance.
- Simple baselines get better results than complex meta-algorithms [22] on some setups, which indicates that there is still room for improvement.

Unsupervised and self-supervised learning

Concurrently to approaches working with limited annotated data, other works have developed strategies to learn without any annotation. Unsupervised and self-supervised learning (one might

argue that self-supervision is a form of unsupervised learning) can only rely on automated discovery of patterns in the data to conduct the given task. Some ideas with encouraging results include:

- Teaching a model how to solve Jigsaw puzzles [64] generated by selecting patches of an image, to automatically learn the important parts of an object. This is an example of a pretext task (solving puzzles) making the model learn semantically important features (shapes, relative spatial positions), that can be reused for a downstream task (in our case retrieval).

- Learning image generation models with the Generative Adversarial Networks architecture [31]. In this setup, a generative model competes with a discriminative model. The generative model tries to fool the discriminator by producing realistic fake images, while the discriminator tries to distinguish fake images from real images. Here, the discriminator provides a form of automated supervision to the generator, using only pixel data from a database of images. By reconstructing realistic images, the generative model is forced to get a visual understanding of the object. Applications to discriminative tasks have shown that the learned features do contain discriminative information [23], but for now only on the very basic MNIST [21] dataset.

- Leveraging data augmentation techniques to learn visual patterns without labels. Recall the tuplet losses presented in 3.4.3: the positive pairs that we pull together (while we push negative pairs apart) can be automatically generated with a single image on which we apply simple transformations such as cropping, distortion, blur... Chen et al. [15] achieved results similar to early supervised models on ImageNet classification with this framework, coined SimCLR (Simple Contrastive LeaRning). They report that the most efficient transformations are random cropping and color distortion, arguing that they force the model to ignore irrelevant contextual clues (the environment around the object) or the shortcut of color patterns (the histogram of colors on an image is robust against cropping, rotating or blurring).

Lastly, and considering our task of image retrieval, note that the handcrafted local descriptors such as SIFT or ORB [79] that were used before deep learning methods do not require any form of supervision (nor any learning). Combined with advanced detectors and indexing methods, they can provide competitive results [73], but we will not include them in our experiments here following early evaluations showing a clear advantage of deep local and global descriptors [30], and because they have less room for improvement.

3.4.4 Conclusion

Looking back at the methods presented in this literature review, many ideas answer at least partially to what constitutes the ALEGORIA problem but these ideas remain encapsulated in research topics and their corresponding benchmarks. A global approach for a content-based image retrieval framework

working on low-data, heterogeneous scenarios is missing.

More specifically, it seems that the current state of the art in image retrieval is still limited to relatively simple image characteristics, and has not yet catched the train of out-of-domain generalization and unsupervised learning as it is the case for parallel tasks in computer vision. Concurrently, it is not clear how existing methods would perform in a more challenging setup. Depending on the method, many different steps and hyperparameters are involved: global and local descriptors have different indexing structures, the training dataset has a poorly understood influence, post-processing steps generally bring significant improvement but vary between methods, as backbone and loss functions do... And, perhaps most importantly, the mAP measure hides all underlying phenomena behind a single value, with a higher importance given to the top results being correct rather than to the recall, *i.e.* retrieving all positive images (especially for small classes).

These factors make it hard, if not impossible, to draw any conclusion on the superiority of a given method on another. I will however mention a few preliminary pieces of evidence that were proposed in the literature, keeping in mind that they must be read with a critical view.

1. It seems that local descriptors behave better on large databases of images, supposedly because they handle visual noise and distractor images better [73].

2. While the size of the training dataset obviously matters, it might be more efficient to build a "clean" dataset (*i.e.* with accurate annotations and limited variations) rather than accumulating more images [106].

3. The choice of the loss function does not seem to play a major role. While listwise losses have claimed to be theoretically optimal for the task of retrieval [12, 77], they have not showed a clear superiority on benchmarks [65].

Concurrently, here are the ideas that seem promising in terms of enhancing generalization:

1. Diffusion can help retrieving images with different visual characteristics, provided that enough "connecting" images (images that are similar both to the query and to the target) are available.

2. Meta-learning and semi-/unsupervised learning can alleviate the need for training data, it is however unclear if or when they can perform better than supervised approaches.

3. Some approaches do propose descriptors suited for retrieval in difficult conditions, including extreme viewpoint variations, which indicates that it is possible to train robust descriptors. What is missing now is a proof that these methods can be applied to (a) an undefined, broad panel of variations (instead of domain A to domain B, N domains at the same time) (b) cases where no training data is available.

In the next chapter, I will present how state-of-the-art methods behave against the ALEGORIA benchmark, using a detailed analysis made possible by the annotations available along the usual class information, and continue with solutions allowing a gain of performance, borrowing from pararallel propositions in meta-learning and diffusion.

Chapter 4

Addressing low-data, heterogeneous image retrieval: methods and results

In Chapter 3, I presented the elements leading me to formulate the ALEGORIA problem through image retrieval, and identified the main two challenges from a computer vision perspective: the heterogeneity of visual characteristics, and the low-data setup, both consequences of working with cross-source data.

In this chapter I will present the solutions I propose to tackle this problem, and the results obtained when confronting these methods to the ALEGORIA benchmark. To not be limited to ALEGORIA, they follow a global approach, as independent as possible of the data (*i.e.* generalizable), borrowing from the broad panel of approaches in image retrieval, but also considering the cutting-edge propositions in meta-learning to see if they have applications outside of their original context.

In Section 4.1, I will begin with a necessary performance comparison of deep features on the challenging setup of ALEGORIA. The results will serve as a starting point for a new method handling the diversity of representation characteristics: Multi-Descriptor Diffusion (MD) presented in Section 4.2. Then, in Section 4.3, I will propose a second method, handling the lack of training data: retrieval-adapted CNAPS (rCNAPS) for few-shot retrieval.

4.1 Benchmarking deep features

Before proposing new ideas, existing propositions must be compared to identify their weaknesses. In Section 3.4, we covered a panel of contributions related to building deep descriptors for image retrieval. Most of these propositions are in fact small processing blocks (even if they were presented in the context of an entire retrieval framework for evaluation), inserted at different points of the processing pipeline. These blocks, plus the training dataset, can be combined or modified, leading to

a virtually infinite number of possible descriptors. Figure 4.1 shows the most important processing steps for computing a deep descriptor.

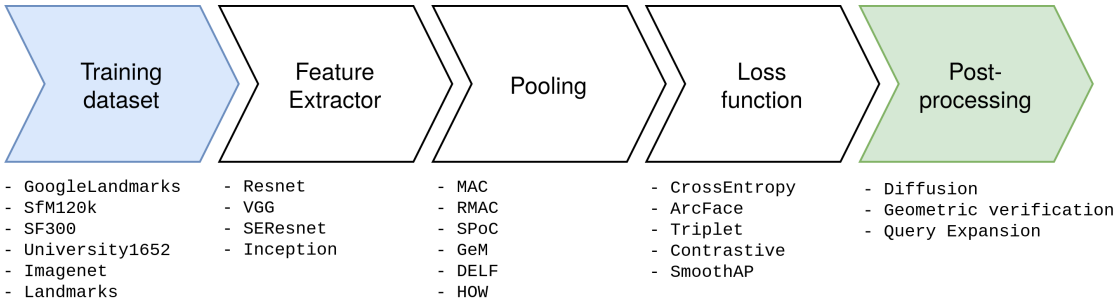

Figure 4.1: Data and processing blocks in the deep descriptor pipeline. A non-exhaustive list of possible choices is indicated below each block.

There has been extensive discussion around the choice of feature extractors, pooling operations and loss functions used to build deep descriptors, but as concluded in Section 3.4 no proper "golden standard" emerges from the literature. Ko et al. [42] extensively compared some of the combinations possible in Figure 4.1, with the same training dataset (GoogleLandmarks). They did not report a significant change of performance due to these three factors (apart from the obvious gain with bigger networks). Rather, they found that pre-processing (image resizing, cropping) and post-processing steps can heavily influence final performance.

In this section, I will not extend further the search of a potentially optimal combination of processing blocks, because we can reasonably consider that authors have already tested a substantial fraction of these combinations explicitely [42] or implicitely when trying to exceed previous performance on benchmarks. It remains the training dataset, which has a poorly understood but strong influence on the behavior of descriptors. This section will therefore compare some baselines that will, *a priori*, be the most suited to ALEGORIA, regardless of what processing blocks they use, with an emphasis on picking methods trained on different datasets.

4.1.1 Baselines

Baselines are separated in two groups: methods relying on a large training dataset with semantics close to ALEGORIA (presented in Section 3.1), and methods with no prior assumptions, *i.e.* unsupervised or out-of-domain.

Supervised baselines

We compare a panel of methods which, *a priori*, will have good performance on ALEGORIA considering that they are trained on datasets with close semantics and on an image retrieval task. We include

models that we built and trained ourselves using common processing blocks from the litterature, and models trained by authors in the corresponding publication, with provided weights.

Finetuning (**GeM - ArcFace**): using ResNet50 as a backbone feature extractor (with a linear dimension reduction layer at the end bringing final descriptor dimension to 512), GeM pooling, training datasets presented in Table 3.1, and the ArcFace loss, we fine-tuned the model weights, stopping when validation accuracy reached a maximum.

State-of-the-art methods (***[Desc] - [Loss]**): using the weights provided by authors, we test four state-of-the-art methods: GeM trained with the Triplet Loss on the multi-view University-1652 dataset [121], GeM trained with the Contrastive Loss on SfM120k [74], RMAC trained with the AP ranking loss on Landmarks [77], and the local descriptor HOW trained with the Contrastive Loss on GoogleLandmarks [97]. Global descriptors are compared with the cosine similarity, while HOW relies on the ASMK matching kernel (see Appendix B) to compute pairwise similarities.

Unsupervised and out-of-domain baselines

To assess what role annotations play in the performance of a model for image retrieval, we also include a method from unsupervised learning and an out-of-domain baseline. The first tries to directly learn accurate representation, while the second re-uses the representations learned with labels from a disjoint distribution of images. In both cases, these methods do not rely on labels relevant to the target task.

Unsupervised learning (**SimCLR**): using a reimplementation of the original method (see Section 3.4.3) with ResNet50 as the feature extractor (we modified the last layer to produce 512-dimensional descriptors), we train the model on the ALEGORIA distractor set (11k images) until convergence, and test the produced descriptors on ALEGORIA.

Pretrained (**GeM**): using ResNet50 trained on ImageNet as a backbone feature extractor (weights provided by common deep learning libraries) and GeM pooling, we produce 2048-dimensional global descriptors.

4.1.2 Metrics and evaluation setup

The evaluation routine goes as follows:

1. Descriptors are extracted for the query set (annotated images) of ALEGORIA.
2. Descriptors are extracted for the base set, which is either equal to the query set, or to the query set + distractors. We separate the two setups to study the influence of adding distractor images.
3. Query and base descriptors are compared with the similarity measure.

4. For each query, a list of results is produced by ranking similarities in decreasing order of similarity.

5. We evaluate the list of results with our proposed metrics in three evaluation setups.

The three setups for evaluating retrieval performance on the ALEGORIA benchmark are:

- Absolute retrieval performance: the average quality of result lists obtained when using all annotated images as queries, regardless of domain considerations.
- Intra-domain or attribute-specific performance: the retrieval performance obtained when using a subset of annotated images from a specific collection or with a specific attribute value. This allows a finer comparison along different representation domains and characteristics.
- Inter-domain performance: the ability to retrieve images outside of the query domain.

Absolute retrieval performance is measured with the mean Average Precision (mAP). Following the notation proposed by [12], the Average Precision for query q is defined as :

$$AP_q(P, \Omega) = \frac{1}{|P|} \sum_{i \in P} \frac{R(i, P)}{R(i, \Omega)} \tag{4.1}$$

where P is the set of positive images for query q, $\Omega = P \cup N$ is the set of all images (positives P and negatives N), $R(i, P)$ is the ranking of image i in P, and $R(i, \Omega)$ is the ranking of image i in Ω. Rankings are obtained by sorting pairwise image similarity scores (which depend on the descriptor used) in decreasing order. The mAP is computed by averaging APs over the set Q of 1858 queries. Sets P and N are usually the same for all images belonging to the same class. However on the ALEGORIA dataset, some images show objects from multiple classes. In these cases, P is specific to the query and includes all images containing one of the objects of interest.

Intra-domain performance is measured with mAP scores using the provided collection and representation attributes, on a subset of the queries. The collection- or attribute-specific mAP is defined as the mAP obtained when filtering Q with collection or attribute values, respectively. For example, the intra-domain performance for the collection "Lapie" is measured as the mAP computed on the subset of 262 Lapie queries. Similarily, the intra-domain performance for the attribute "Scale", value "Very close" is measured as the mAP computed on the subset of 242 queries with this value. Note that these various mAP scores are all computed on different sets of queries and therefore should not be compared to each other, they only allow comparison of different methods on the same setup.

Inter-domain performance is measured with a set of measures based on the position of positive images from different collections. We propose two statistical indicators based on P1, the position of the first positive image from a different collection (see Figure 4.2 for a visual representation of P1): median P1 (mP1) and the first quartile of P1 (qP1). These two measures give a rough idea of how positive images from different collections are distributed in the list of results. Note that they are

independent of in-domain performance: if the first positive image from a different collection is at position $P1$, positive images from the same collection before $P1$ are not taken into account. We also propose a new measure encoding more information about the statistical distribution of cross-domain images. First, we introduce the Average Position Deviation as the difference between the average position of positive images from different collections and the average position of all positive images. This score is computed for each query, and averaged over all queries to give the mean Average Position Deviation, mAPD, similarly to the mAP. This measure has the advantage of not depending on the number of positive images and the absolute performance of the method, and is easily interpretable: the ideal value is zero (images are retrieved regardless of their collection, *i.e.* their average position stays the same), and higher values indicate that positive images from different collections are further in the list of results.

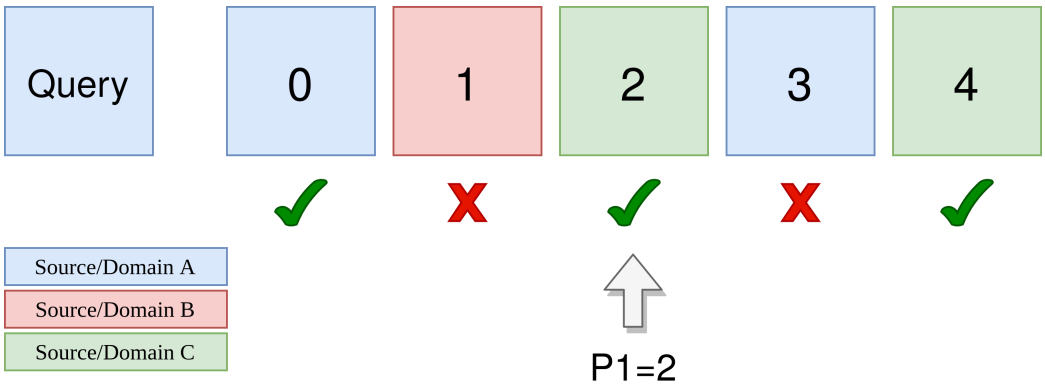

Figure 4.2: Calculation of P1, the position of the first positive image from a different domain (or collection), for a single query. Different colors indicate different domains, True/False symbols indicate if the corresponding image is positive or not for this query.

4.1.3 Pre-processing

It is well known in computer vision that simple transformations applied to test images can already help gain some performance [42]. Table 4.1 shows the effect of inserting two simple preprocessing steps with virtually no computational overhead, on our best performing descriptor: switching all images to grayscale, and cropping images with a 0.9 ratio. Motivated by the small gain in accuracy, we apply this preprocessing for all experiments in Table 4.2.

Table 4.1: Effect of preprocessing test images on absolute performance, with GeM trained on GoogleLandmarks.

Grayscale	Crop	Absolute perf. (mAP)
		23.68
✓		23.78
	✓	24.17
✓	✓	24.30

The gain of performance is easily explained:

- Switching to grayscale reduces the variance. Some information is necessarily lost when transforming an RGB image to grayscale (the raw volume of information is divided by 3), but it seems that the reduction of variance prevails over the loss of pixel information, in our setup of image retrieval. Another way of reducing variance would have been, instead, to generate information in grayscale images, *i.e.* to colorize them, so that all images are RGB. We conducted some experiments in this direction with an early version of the benchmark, but concluded that generative models lacked expressivity (at that time) and did not produce satisfying results. Appendix C presents the corresponding experiments.

- Cropping images has the double advantage of removing image borders that are found on some images (see Figure 3.6), and removing a part of images which often does not contain important information (there is a natural centering bias when taking pictures to emphasize the object of interest).

4.1.4 Results

Table 4.2: Performance comparison of various descriptors on ALEGORIA, with absolute, intra-domain and inter-domain measures. Descriptors with a star (*) are extracted using authors' provided model weights, other descriptors are from our own reimplementation and training. Best performance for each measure (column) is in **bold**. Absolute and intra-domain performance is measured with the mAP (the higher, the better). Inter-domain performance is measured with specific indicators detailed in section 4.1.2, mP1 and qP1 do not have fixed optimal values (but a lower value indicates a better resilience to visual changes), while optimal mAPD is zero (the lower, the better).

			Absolute perf. (mAP)		Intra-domain performance (mAP)					Inter-domain performance		
Method	Training dataset	Reranking	ALEGORIA	+*distractors*	MRU	Lapie	Photothèque	Internet	Henrard	mP1	qP1	mAPD
Unsupervised & Out-of-domain												
SimCLR	ALEGORIA distractors		5.77	-	10.22	7.89	5.01	4.29	7.66	134	40	**176.8**
***GeM**	ImageNet		14.25	8.6	23.63	19.92	11.28	12.98	13.93	152	23	290.9
Supervised												
GeM - ArcFace	GoogleLandmarks		**24.30**	**17.49**	27.16	**29.43**	13.45	**34.10**	**20.33**	67	10	251.0
GeM - ArcFace	SF300		14.41	8.65	26.27	17.76	14.44	9.33	13.32	99	20	256.5
GeM - ArcFace	SfM120k		15.38	9.39	22.6	17.34	12.31	15.71	12.55	143	24	283.7
***GeM - Triplet**	Univ1652		11.02	5.44	22.8	15.67	10.15	6.65	11.15	111	32	239.1
***GeM - Contrastive**	SfM120k		19.02	12.46	26.39	19.70	**14.59**	20.57	16.83	107	18	281.1
***RMAC - APLoss**	Landmarks		19.97	12.94	24.82	20.25	13.96	24.38	15.49	79	13	261.5
***HOW - Contrastive**	GoogleLandmarks		19.16	13.02	**28.11**	19.53	10.20	24.45	17.55	105	21	274.0

4.1.5 Discussion

Supervision axis

The experiment with SimCLR confirm our expected drop in performance from Figure 3.19. Even if the training datasets from supervised methods do not exactly match ALEGORIA statistics, their volume bring enough variety to compute descriptors with better accuracy and generalization ability. The global mAP with GeM trained on ImageNet, a generalistic dataset semantically very far from ALEGORIA, is ~ 9 points higher than the mAP with SimCLR, highlighting the importance of annotated training data.

Intra-domain performance disparities

Separating absolute performance from intra-domain performance allows us to better understand underlying behaviors: some descriptors comparatively perform better on a domain, even if their absolute performance is lower. For example, descriptor HOW gives a better mAP on collection MRU than GeM-ArcFace trained on GoogleLandmarks, even if its absolute performance is ~ 4 points lower.

We note that the notion of domain, as presented in the literature, stays relatively vague in its definition: it can be understood as any criterion that allows clustering of the data into separate groups. In our case, we use the term domain to refer to a collection because we are interested in matching images through different collections, but studying domain-specific performance does not inform us on the fundamental image variations that impact how descriptors behave. To better understand and visualize this, we use the available variation annotations to compute attribute-specific mAP depending on the training dataset, with GeM global descriptors and HOW local descriptors for comparison. Results are shown in Figure 4.3.

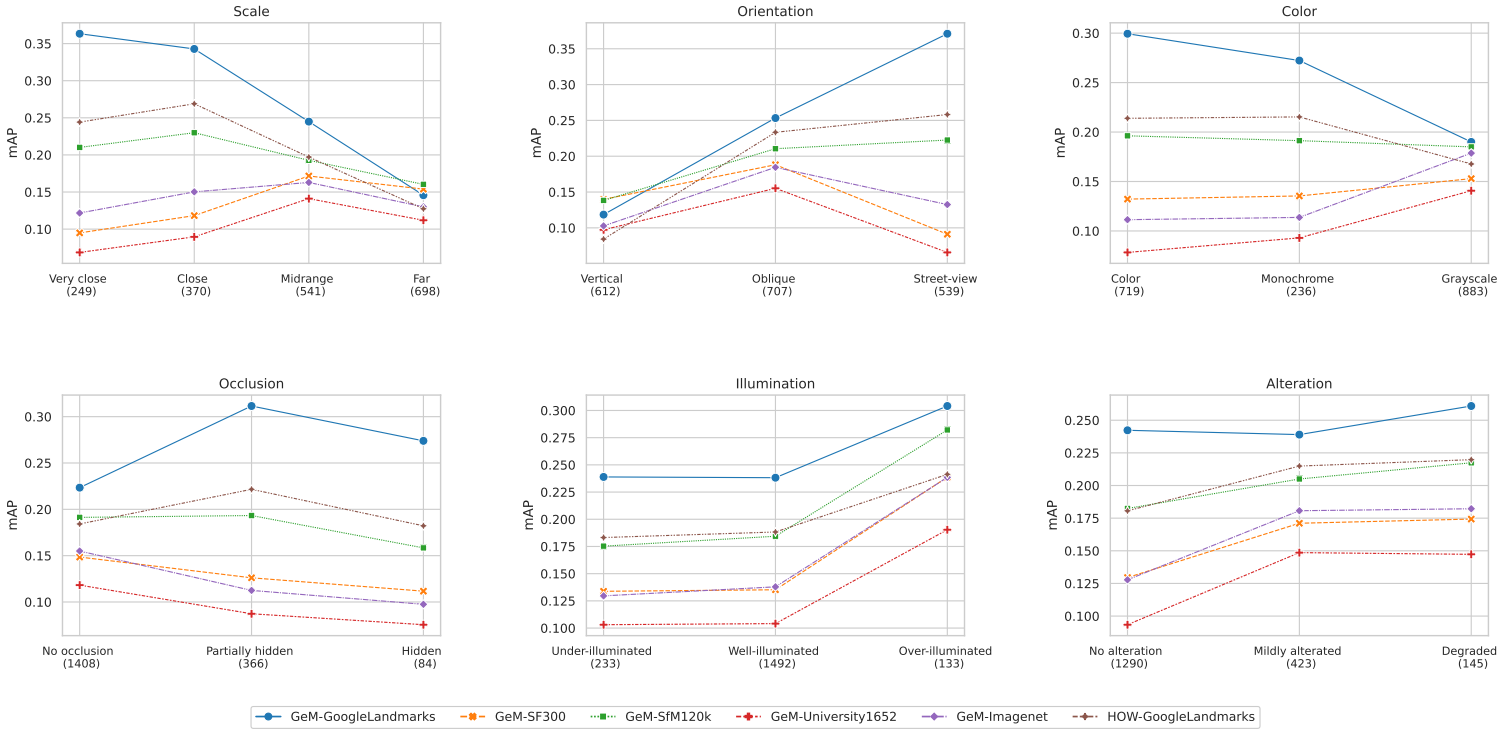

Figure 4.3: Attribute-specific performance evaluation. The number of corresponding query images is noted in parentheses. Performance should be compared between descriptors on a single value (column), because the varying query set can make absolute performance change independently from attribute variations.

We notice that GeM-GoogleLandmarks is better than other descriptors in most cases, except a particular type of images: vertical images, with very small objects. This can be explained with the main semantic of GoogleLandmarks being street-level tourist photography.

Influence of distractors

There is a major drop in performance when adding distractor images, ranging from a $\sim 28\%$ decrease for GeM-ArcFace trained on GoogleLandmarks to a $\sim 51\%$ decrease for GeM-Triplet trained on University1652. Similarly to what was observed on $\mathcal{R}$Oxford and $\mathcal{R}$Paris, we do not find that local descriptors (here, HOW) are necessarily better on larger benchmarks, indicating that the historical idea that local features are more robust is not valid anymore with deep descriptors.

Influence of the training dataset

The comparison between GeM trained on GoogleLandmarks and GeM trained on SF300 highlights the importance of the training dataset. The difference of 10 points of mAP is explained by the high volume and higher relevance of images in GoogleLandmarks. However, the method trained on SF300 is slightly better on Photothèque images, i.e. mostly vertical aerial imagery. Starting from this comparative difference, in the next section I will present a diffusion method which makes use of multiple descriptors to handle the multiple representation modalities.

4.2 Multi-descriptor diffusion

This section will present the main contribution of this thesis regarding cultural data: multi-descriptor diffusion. Section 4.2.1 describes our proposed method, Section 4.2.2 presents related methods, and in Section 4.2.3 we compare them on the ALEGORIA benchmark. We then discuss results in Section 4.2.4.

4.2.1 Method

Following eq. 3.2 and the graph diffusion framework, we propose to explore how to generalize to multiple descriptors, *i.e.* how to build and propagate on a graph from multiple similarity matrices S^1, S^2, S^3... ; with the motivation of exploiting the different properties of different descriptors.

If S^υ is the similarity matrix issued from the υ-th descriptor, and we have Υ descriptors, we first build Υ kNN graphs:

$$a^\upsilon_{i,j} = \begin{cases} 1 & \text{if } x_j \in NN^\upsilon_{k1}(x_i), \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

with $NN^\upsilon_{k1}(x_i)$ the $k1$ nearest neighbors of x_i according to S^υ. Note that $k1$ stays the same accross graphs, to avoid multiplying parameters and because it ideally corresponds to the average number of positive images for a query, a value independent from the description method. Note that kNN

graph building is independent from the descriptor used: the only input is a similarity matrix that can be obtained from a cosine similarity, an euclidean distance, a matching kernel between local descriptors...

Following [119], we make the graph undirected (*i.e.* A symmetrical) to give more importance to reciprocal pairs of images. It is done with a simple operation:

$$A^* = \frac{A + A^T}{2} \tag{4.3}$$

This way, we get:

$$a^{\upsilon}_{i,j} = \begin{cases} 1 & \text{if } x_j \in NN^{\upsilon}_{k1}(x_i) \wedge x_i \in NN^{\upsilon}_{k1}(x_j), \\ 0.5 & \text{if } x_j \in NN^{\upsilon}_{k1}(x_i) \vee x_i \in NN^{\upsilon}_{k1}(x_j), \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

We conduct a first diffusion step to refine descriptor-specific similarities. As in [119], we propagate only using the top $k2$ neighbors:

$$\boldsymbol{s}^{\upsilon}_i \leftarrow \sum_{j \in NN^{\upsilon}_{k2}(x_i)} a^{\upsilon}_{i,j} * (s^{\upsilon}_{i,j})^{\alpha} * \boldsymbol{s}^{\upsilon}_j \tag{4.5}$$

This is an exponentially weighted update rule similar to αQE (eq. 4.10): the vector $\boldsymbol{s}^{\upsilon}_i$ of similarities between image i and other images is updated by a weighted sum of neighboring similarity vectors $\boldsymbol{s}^{\upsilon}_j$. Here $s_{i,j}$ is a scalar value indicating the initial similarity between images i and j, *i.e.* the (i,j)th entry in matrix S^{υ}, while $\boldsymbol{s}^{\upsilon}_i$ and $\boldsymbol{s}^{\upsilon}_j$ are ith and jth rows respectively. Note that here we allow matrix A to take non-integer values, which differs from the standard graph notation where A is only zeros and ones, but this stays compatible with the graph view if we consider that edges contain both similarities and an additional weight in $\{0.5, 1\}$. To summarize, for each image we use the top $k1$ neighbors to define a feature vector based on similarities, and we refine this feature vector with an update based on the top $k2$ neighbors.

Matrices S^{υ} are L2-normalized, and merged :

$$S = \frac{1}{Y} \sum_{\upsilon \in [1..Y]} S^{\upsilon} \tag{4.6}$$

From this step, we follow the same logic again with merged similarities. We recompute A using similarities instead of descriptors as features (see eq. 3.4, we keep the same $k1$), and we update similarities with the top $k2$ neighboring nodes:

$$\boldsymbol{s}_i \leftarrow \sum_{j \in NN_{k2}(x_i)} a_{i,j} * (s_{i,j})^{\alpha} * \boldsymbol{s}_j \tag{4.7}$$

Using the final similarity matrix S, we evaluate on ALEGORIA. We coin this method Multidescriptor Diffusion (**MD**).

We also propose an alternative method introducing annotation information in the diffusion process: in A, we can force certain values based on external criteria. Specifically, using the domain annotations provided in the ALEGORIA benchmark, we can define an inter-domain matrix T:

$$t_{i,j} = \begin{cases} 1 & \text{if } domain(x_i) \mathrel{!}= domain(x_j), \\ 0 & \text{if } domain(x_i) = domain(x_j) \end{cases} \tag{4.8}$$

This allows us to force connections between images from different collections (positive or negative, we obviously don't use class annotation). There is intuitively a compromise to solve here between descriptor-specific performance and out-of-domain retrieval, we thus introduce a λ parameter to allow tuning. After eq. 4.4, we merge T into A with:

$$A \leftarrow A + \lambda * T \tag{4.9}$$

We will refer to this method as constrained Multidescriptor Diffusion (**cMD**).

4.2.2 Reference methods

We include two state-of-the-art diffusion methods. They cannot be directly compared to our proposed multi-descriptor methods because they only use the similarity matrix from a single descriptor, but provide an useful reference performance.

Alpha Query Expansion (α**QE**) [74]: from a kNN adjacency matrix (see eq. 3.4), we update node features using an exponentially weighted update rule (single update):

$$\boldsymbol{d}_{x_i} = \sum_{x_j \in NN_k(x_i)} (s_{i,j})^{\alpha} * \boldsymbol{d}_{x_j} \tag{4.10}$$

Graph Query Expansion (**GQE**) [119]: starting from equations 4.3 and 4.4, GQE follows a purely graph-based formulation. Node features h_i are defined as the rows of A^*, and a graph is built by finding the $k2$ nearest neighbours of each node. GQE propagates N times using the following update rule:

$$\boldsymbol{h}_i \leftarrow \sum_{j \in NN_{k2}(x_i)} (s_{i,j})^{\alpha} * \boldsymbol{h}_j \tag{4.11}$$

with L2-normalization of node features (h_i after each update). The attentive reader will notice how similar equations 4.10 and 4.11 are, but keep in mind that the difference lies in how inter-node (or inter-image) similarities are computed: αQE directly uses the cosine similarity of global descriptors, while GQE uses neighbor encoding features (each image/node is represented by its adjacency with the whole database, rows of A^*) to re-compute similarities (edge features $s_{i,j}$). The final similarity is

obtained with the cosine similarity of updated node features. There are a few differences with our proposed MD approach:

- GQE does not allow non-integer values for the adajcency matrix. With MD, we modified the update rule to include a third $a_{i,j}$ term with values in $0, 0.5, 1.0$, which allows a more fine-grained propagation.
- GQE uses multiple updates. We voluntarily did not include multiple updates, since GQE authors did not report a significant performance improvement when using multiple updates.
- GQE stays within the graph formulation: nodes are features which are updated through propagation, and the final similarity is computed with the cosine distance of nodes. With MD, we depart from the graph formulation: we directly update matrices of similarity S^v/S, which allow us to skip this last step.

4.2.3 Results

Table 4.3: Performance comparison of diffusion methods on ALEGORIA, with absolute, intra-domain and inter-domain measures. Best performance for each measure (column) is in **bold**. Absolute and intra-domain performance is measured with the mAP (the higher, the better). Inter-domain performance is measured with specific indicators detailed in section 4.1.2, mP1 and qP1 do not have optimal values, while optimal mAPD is zero (the lower, the better).

			Absolute perf. (mAP)	Intra-domain performance (mAP)					Inter-domain performance		
Method	Training dataset	Reranking	ALEGORIA	MRU	Lapie	Photothèque	Internet	Henrard	mP1	qP1	mAPD
Supervised											
GeM - ArcFace	GoogleLandmarks		24.30	27.16	29.43	13.45	34.10	20.33	67	10	251.0
***GeM - Contrastive**	SfM120k		19.02	26.39	19.70	14.59	20.57	16.83	107	18	281.1
***HOW - Contrastive**	GoogleLandmarks		19.16	28.11	19.53	10.20	24.45	17.55	105	21	274.0
Diffusion											
GeM - ArcFace	GoogleLandmarks	+ α**QE**	25.02	27.95	30.41	13.19	35.81	20.57	77	11	**250.6**
GeM - ArcFace	GoogleLandmarks	+**GQE**	27.41	30.60	45.08	15.01	37.82	23.94	78	11	284.8
GeM - ArcFace	GoogleLandmarks										
***GeM - Contrastive**	SfM120k	+**MD**	**29.17**	35.28	39.70	**17.92**	**37.53**	**26.00**	97	13	327.5
***HOW - Contrastive**	GoogleLandmarks										
GeM - ArcFace	GoogleLandmarks										
***GeM - Contrastive**	SfM120k	+**cMD**	29.10	**35.31**	**41.40**	17.87	37.35	25.60	81	12	279.5
***HOW - Contrastive**	GoogleLandmarks										

4.2.4 Discussion

Performance

The two state-of-the-art diffusion methods we tested, αQE and GQE, improved performance by respectively 0.72 and 3.11 points of mAP. Our proposed multi-descriptor method using the 3 best performing descriptors in intra-domain performance brings an improvement of 4.87 points. The constrained version does not further improve absolute performance, but changes inter-domain performance as we will detail later.

Optimal parameters in terms of absolute performance found for αQE are (α, n)= (1, 3), for GQE ($k1$, $k2$, α) = (38, 5, 1.0) ; for MD ($k1$, $k2$, α) = (15, 4, 7) and for cMD ($k1$, $k2$, α, λ) = (17, 4, 9, 0.1).

Our proposed MD and cMD methods use respectively three and four hyper-parameters. To study how these parameters influence the performance, we first put aside cross-domain performance (with the λ parameter of cMD) and evaluate the evolution of the absolute mAP when varying $k1$, $k2$ and α. Figure 4.4 shows a heatmap of the absolute mAP against $k1$ and $k2$ ($k2 < k1$), with α fixed to 7. Apart from the obviously suboptimal region of ($k1$, $k2$) $<$ (3, 3) which does not exploit enough neighbouring information, and a decreasing performance when reaching high values, there is a near-optimal zone of (10, 3) $<$ ($k1$, $k2$) $<$ (22, 20) where absolute performance is stable regardless of varying $k1$ and $k2$. This indicates that tuning these parameters should not be a problem on cases similar to Alegoria.

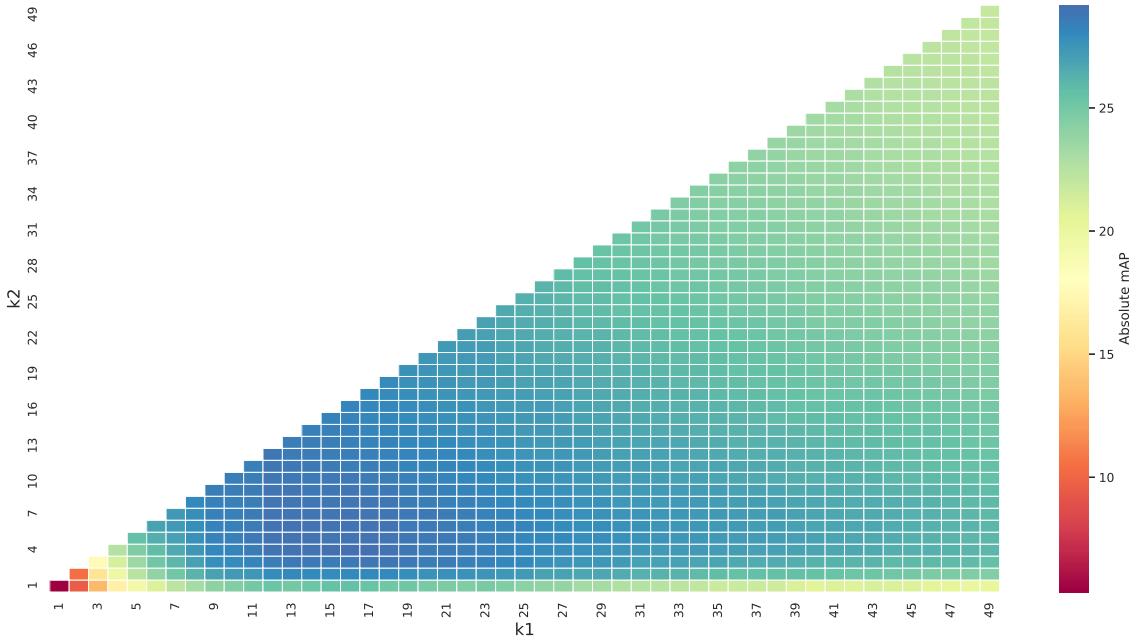

Figure 4.4: Heatmap of the absolute mAP against $k1$ and $k2$ with the MD method (best seen in color).

Similarly, Figure 4.5 shows the evolution of the absolute mAP against α, with ($k1$, $k2$) fixed to (15, 4). Again, the performance does not significantly changes when $\alpha > 4$. This is a behaviour similar to what was observed in the original proposition of αQE [74], on which our proposed MD and cMD methods draw inspiration.

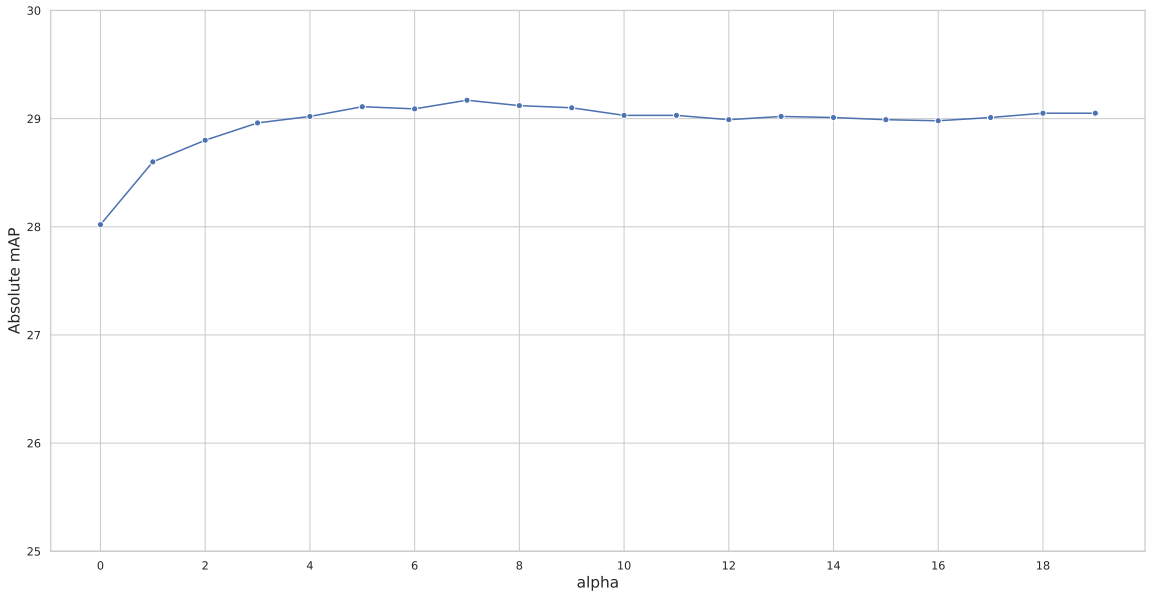

Figure 4.5: Evolution of the absolute mAP against α with the MD method.

Heterogeneity

We note that increased absolute performance, *i.e.* descriptor accuracy, is generally accompanied with an increase of the mAPD measure, indicating positive images from different collections being pushed to the end of the list. This is coherent with our assumption that descriptors are very dependent on their training statistics, meaning that they are accurate only on images with known semantics.

Diffusion does not prevent this and only reinforces dispersion of cross-domain images: the highest mAPD score corresponds to the highest absolute mAP with our proposed MD diffusion method.

To solve this challenge, we proposed the cMD method to study if it is possible to enhance cross-domain performance while keeping a reasonnable absolute performance. In particular, the λ parameter inserts control on the compromise between these two objectives. Figure 4.6 shows the influence of varying λ. We observe a bell-shaped curve for qP1 and mP1, indicators of how high in the list positive cross-domain image are, while mAPD continuously decreases with increasing λ. It seems that there is a "sweet spot" for balancing absolute and cross-domain accuracy, around 0.5-0.6.

Figure 4.7 shows some visual examples of the trade-off between accuracy and cross-domain retrieval: we observe that MD significantly improves the top 4 results with this query, but mostly retrieves images in the same collection (here, internet). Our proposed cMD variation maintains the accuracy of the results, but pushes positives from different collections higher on the list.

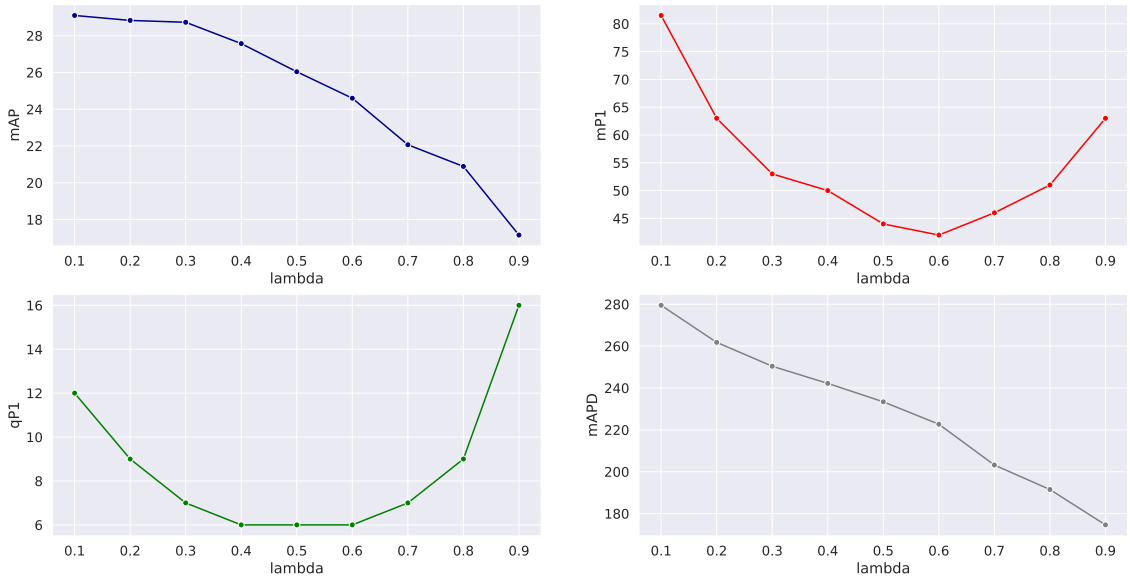

Figure 4.6: Evolution of the absolute mAP and various inter-domain measures against λ with the cMD method.

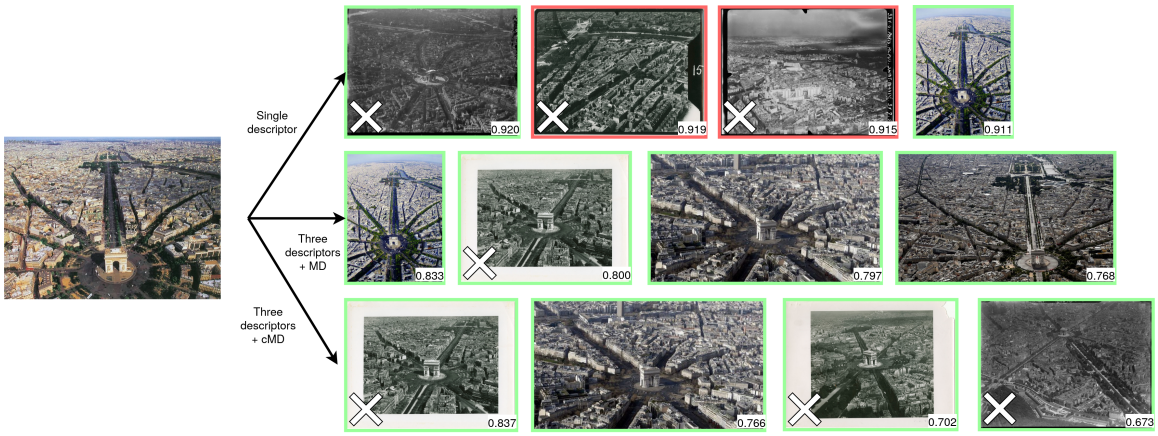

Figure 4.7: Query examples on the "Arc de Triomphe" class. First row: results using GeM - ArcFace trained on GoogleLandmarks (best performing single descriptor in Table 4.2). Second row: results using the three best performing descriptors + our proposed multi-descriptor diffusion (MD). Third row: results using the three best performing descriptors + our proposed constrained multi-descriptor diffusion (cMD). Positive results are indicated in green, negative in red. Similarity scores are indicated in the bottom right of each image. Cross-collection images (*i.e.* belonging to different collections) are indicated with a white cross on the bottom left.

Computational complexity

Our proposed MD and cMD methods are agnostic of the descriptors used, and only require as inputs the corresponding similarity matrices. In this work, we introduce and evaluate the idea of connecting images regardless of their source, and consider that the descriptors are already computed and stored for evaluating the considered diffusion methods. In an online setup, issuing an unseen query would require to compute its descriptors and similarities with the whole base. This process can be optimized: a common approach compatible both with local and global descriptors is Product Quantization [38], where descriptors are separated and binarized in sub-vectors for efficient storing ; and the diffusion process can be decoupled in offline and online steps [113] by pre-computing similarity matrices and diffusion steps. For the ALEGORIA benchmark, such scale-up methods are not necessary considering the relatively low volume of images.

Table 4.4 shows the computational overhead of the considered diffusion methods. Our proposed MD and cMD diffusion methods use the optimized GPU implementation of GQE [119], offering computation times lower than αQE (CPU) even with two additional descriptors.

Table 4.4: Computational overhead of diffusion methods on ALEGORIA. Experiments made with an Intel i7-8700K CPU (3.7GHz), 32Go of RAM and a single NVIDIA RTX2080 Ti (12Go VRAM).

Method	Computation time
αQE	147ms
GQE	57ms
MD	128ms
cMD	140ms

4.3 Few-shot retrieval

Following the ideas from semi-supervised training and few-shot learning presented in Section 3.4.3, we propose to investigate the links between few-shot learning and image retrieval, to see if image retrieval in low-data scenarii can benefit from the abundance of new methods related to few-shot learning.

First, we introduce the new problem of few-shot retrieval in Section 4.3.1. Section 4.3.2 presents a modified version of CNAPS suited for few-shot image retrieval. In Section 4.3.3 we present the evaluation setup, used to evaluate our proposed method on the ALEGORIA benchmark in Section 4.3.4. We discuss results in Section 4.3.5.

4.3.1 Few-shot retrieval formulation

In image retrieval and semi-supervised learning (of which few-shot learning is a part), we are in a setup where a defined context with given classes and examples is not available. Rather, models have to be able to differentiate samples from classes unseen at train time. In one sentence, N-way k-shot learning is formulated with: "Given k examples from an unseen class, recognize all examples fom the same class, in a pool of images from N different classes". Similarily, image retrieval can be formulated with: "Given 1 example from an unseen class, retrieve all examples from the same class, in a pool of images from an unkown number of different classes". There are two important differences between how the two tasks are formulated:

- In image retrieval, large databases are used. While few-shot learning is usually limited to $N = 5$ and $k = 1, 5$ which gives at most 25 images to separate, in image retrieval both the number of samples and the number of classes is unkown,.
- Another consequence of working with large databases is that methods need to be scalable, and for image retrieval, close to real-time execution. Few-shot learning has, for now, not been concerned with these constraints.

To bridge the gap between the two tasks, we propose a new formulation with few-shot retrieval, which is summarized as follows: "Given k examples from an unseen class, retrieve all examples from the same class, in a pool of images from an unkown number of classes". Concurrently, we propose to relax the constraint of scalability and real-time execution, considering that heterogeneity and the low-data setup are the main challenges and should be solved before reducing query times. In practice, this means that we allow different models to be used for different classes. For a given class, the feature extractor can be modified (adapted) to optimize performance on this particular class. Metrics stay the same as previously: we evaluate with absolute performance, intra-domain performance and inter-domain performance.

4.3.2 rCNAPS: a few-shot retrieval baseline

An interesting solution from few-shot learning to overcome the lack of context is fast adaptation (see section 3.4.3), which consists in quickly (*i.e.* without fully retraining) adapting a model to a given task (here, retrieving examples from a class) by exploiting external information. The external information is provided by the support set, a set of images from the target class. Fast adaptation is conducted by adding a new set of parameterized operations to the feature extractor, parameters that are produced by a feature adaptation network. Additionally, an encoder is used to give a fixed-size input to the feature adaptation network, independently of the size and order (the encoder is permutation invariant) of the support set. See Figure 4.8 for a representation of the training setup. We will refer to this method as **rCNAPS** for retrieval-adapted CNAPS.

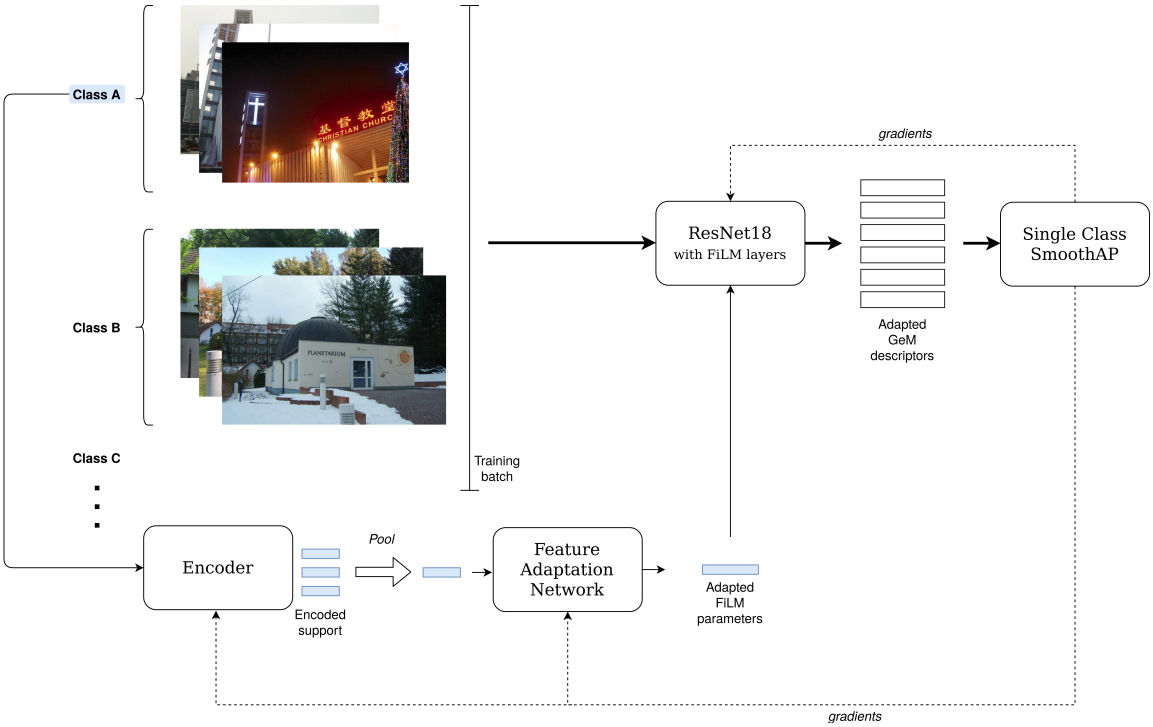

Figure 4.8: Training setup for rCNAPS with single-class optimization: at each training iteration, the model is adapted to the target class, and optimized to maximize the performance specific to that class.

We use the same networks as the "SimpleCNAPS" variation of CNAPS [76] (ResNet18 with FiLM layers inserted, set encoder, adaptation network). We adapted the training setup to be suited to image retrieval. For each training iteration, a class is randomly chosen from the training dataset, it is the target class (denoted in blue on Figure 4.8). We get k random examples from this class which are used as the support. We then build the training batch by sampling m examples from N classes, including the target class. The support is used to compute adapted parameters. The feature extractor, using the adapted parameters for its FiLM layers, computes GeM global descriptors adapted to the target class.

Concerning the loss function, we adopt the Smooth-AP loss [12], which allows to directly optimize a retrieval objective instead of a classification proxy objective. The Smooth-AP loss is computed with a differentiable approximation of the mAP, over the m examples in N classes:

$$L_{AP} = \frac{1}{N * m} \sum_{q=1}^{N*m} (1 - AP_q), \tag{4.12}$$

where AP_q is the smoothed Average Precision for query q. In our training setup, we want to maximize performance for the target class, while the performance for other classes is not relevant. In

other words, for each query in the target class, we want all images of the same class to be at the top of the list, but do not care about the order of images from other classes. We thus modify the Smooth-AP to be class specific:

$$L_{AP} = \frac{1}{m}\sum_{q=1}^{m}(1 - AP_q). \tag{4.13}$$

This way, loss values are ignored for images not belonging to the target class: we only optimize for the target class.

We train with the GoogleLandmarks dataset [106] (2nd version, "clean" subset). We choose $k = 5$ to be consistent with the usual formulation of few-shot learning, $m = 10$ to insert some intra-class variety, and $N = 6$ following the recommendation of Brown et al. [12] to aim for higher batch sizes (here we obtain a batch size of $N * m = 60$ which is the technical limitation of our training hardware with two NVIDIA RTX 2080 Ti, 12 Go VRAM). Image size is set to 320*320.

4.3.3 Evaluation setup

According to the principle of "train as you test", we measure performance on the ALEGORIA with similar setups:

A For each class, we use 1 random image as the support set, extract descriptors for the whole database, and rank queries from the target class. After iterating over all classes, we evaluate the ranking.

B For each class, we use 5 random images as the support set, extract descriptors for the whole database, and rank queries from the target class. After iterating over all classes, we evaluate the ranking.

C For reference, we extract descriptors without adaptation and rank all queries.

D For reference, we evaluate like setup A but with using noise as the support set (random values sampled from the normal distribution). This allow us to verify if the feature adaptation mechanism is used by the model.

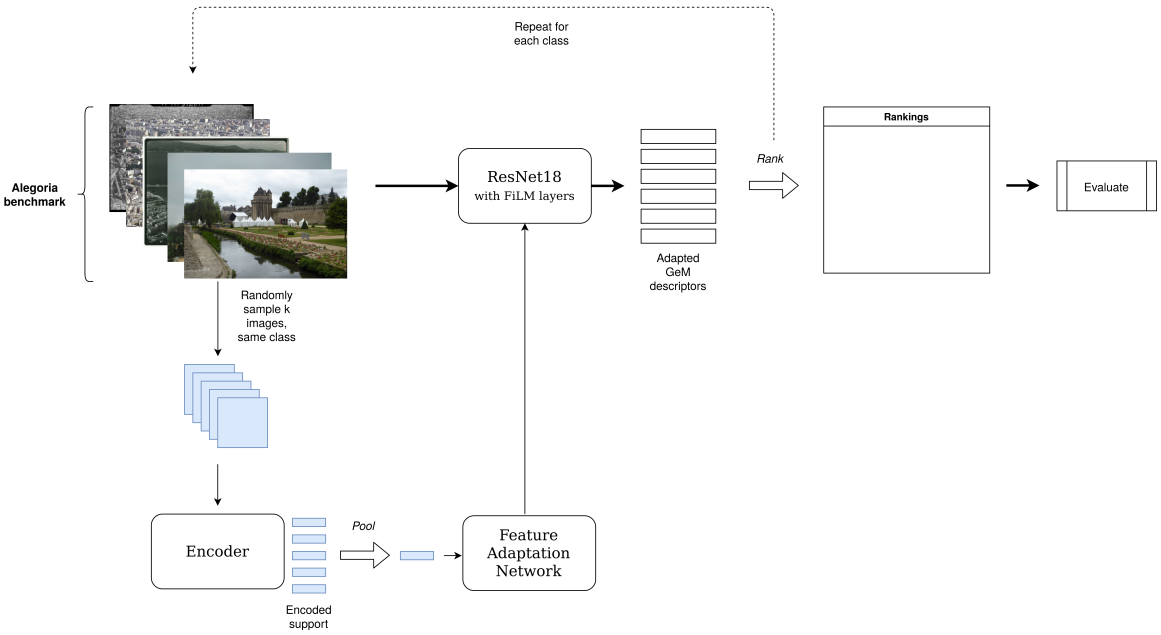

Figure 4.9: Testing setups A and B for rCNAPS. For setup A, $k = 1$, setup B, $k = 5$. For each class, the feature extractor is adapted using a randomly selected support set. With the adapted descriptors, rankings are computed for the target class and stored in the rankings list. The full rankings list is used to evaluate.
Photo credits - copyright: paulafunnell, CC BY-NC-ND 2.0

The randomness brought by sampling support images in setups C and D, and noise in setup B, makes results non-deterministic with a fixed model. We thus evaluate each setup (except setup C which is deterministic) 5 times and indicate the mean of metrics with a 95% confidence interval using Student's t-distribution.

4.3.4 Results

Table 4.5 shows the results when comparing our four few-shot retrieval setups.

Table 4.5: Performance comparison of the four evaluation setups for our few-shot retrieval method. Setup A is few-shot retrieval with 1 support image per class, B with 5 support images per class, C is a reference without the adaptation mechanism, and D a reference with noise used as support.

			Absolute perf. (mAP)	Intra-domain performance (mAP)					Inter-domain performance		
Method	Training dataset	Evaluation setup	ALEGORIA	MRU	Lapie	Photothèque	Internet	Henrard	mP1	qP1	mAPD
Reference (unadapted)											
rCNAPS	GoogleLandmarks	**Setup C**	11.53	19.82	14.95	11.58	7.99	10.41	123	24	252.5
Reference (noise)											
rCNAPS	GoogleLandmarks	**Setup D**	14.15±0.03	21.75±0.08	15.97±0.06	12.45±0.03	12.81±0.03	12.24±0.03	113±1	19±0	252.3±0.4
Semi-supervised											
rCNAPS	GoogleLandmarks	**Setup A** (k=1)	14.24±0.11	21.66±0.21	14.20±0.20	11.34±0.16	14.12±0.17	13.15±0.19	101±3	21±1	241.0±0.7
rCNAPS	GoogleLandmarks	**Setup B** (k=5)	14.36±0.17	21.90±0.06	14.05±0.24	11.40±0.04	14.25±0.04	13.21±0.02	98±1	20±0	240.3±0.2

4.3.5 Discussion

Increased expressivity

Compared to setup C, our reference without adaptation (the feature adaptation network and encoder are not used), there is already a significant gain of performance even when using noise as support in setup D. This is explained by the gain of expressivity in the model with adaptation mechanisms: the feature adaptation network and encoder add ~ 8.5M parameters to the ~ 11.2M parameters of ResNet18. Even if the input of these networks is not relevant, during the optimization process they learn to produce FiLM parameters that will provide a gain of performance on the training dataset, GoogleLandmarks, which as we have seen in Section 4.1.4 produces accurate descriptors for ALEGORIA (on some domains more than others).

Benefits of the support set

The results are consistent with what is expected: absolute performance increases with more support images, with a maximum absolute performance of 14.36 with setup B (5 support images). This is better than some state-of-the-art methods compared in section 4.1.4.

However, the improvement is mild compared to setup D, with a margin of only 0.21 points of mAP. Using intra-domain performance, we can have a more detailed analysis. Setups C and D are better mostly on images from internet and Henrard, which was also the case in Table 4.2 for the GeM-ArcFace descriptor trained on the same dataset, GoogleLandmarks. Thus, it seems that for a semi-supervised setup, the same conclusion can be drawn: the behavior of the model remains dependent on the training dataset, performing better on situations seen during training (here, internet images), and similarly (MRU) or worse (Photothèque) on unseen situations. On collections Lapie and Photothèque, even when providing relevant support images, the performance decreased, indicating the inability of the encoder and feature adaptation networks to select and provide useful information to the feature extractor.

Inter-domain performance

There is a slight improvement of inter-domain metrics when using a relevant support set. This indicates that the adaptation mechanism brings some robustness to descriptors. For setup A, it is possible that this added robustness comes from a cross-domain image used as support (the support set is randomly selected without taking into account the source of images), but setup B with 5 randomly selected images and the pooled representation from the encoder confirms that this behavior is at least partially independent from the source of images.

4.4 Conclusion

In this chapter dedicated to our proposed methods and the associated experiments, we covered how existing methods behave against the variety of visual characteristics in ALEGORIA, proposed and validated two methods, MD (and its variation cMD) and rCNAPS, and confronted them to the benchmark to make observations on what works in terms of absolute, intra-domain and inter-domain performance.

The results allow us to formulate some conjectures:

- Limiting the evaluation to absolute performance in image retrieval hides a great deal of underlying behaviors that merit consideration. By looking at intra-domain performance along our different sources, and with our proposed inter-domain metrics, we were able to make a detailed analysis which seems necessary to better compare methods, especially considering the variety in the benchmark.

- The variety of compared methods and the experiments with our proposed semi-supervised rCNAPS method indicate that the training dataset has certainly a greater influence than the description method, even when external information is provided with a support set. This calls for a systematic comparison of methods with the same training setup.

- We showed that it is possible to gain some performance with diffusion or with a semi-supervised setup, but both come at the cost of increased specificity, either with decreased inter-domain performance (MD) or decreased intra-domain performance on unseen domains (rCNAPS). With cMD, we proposed a solution alleviating this but it requires additional source information.

Taking a few steps back, the results presented in this chapter highlight the "curse" of data in deep learning: training data becomes an integral part of models through optimization, which makes building large datasets at least as important as designing new architectures. In situations where training data is not easily available, it becomes unclear if time and money must be invested to build models, or simply to annotate data. With the MD diffusion method, we showed that a promising idea is to combine descriptors instead of searching for an optimal method. In situations where there is a high variance in visual characteristics, there is disappointgly no straightforward solution. But we showed with the cMD variant that using additional data (which can be inferred through properly trained classifiers) might help.

In the next chapter, I will apply the same reasoning on a different type of data, where similarly to ALEGORIA cross-domain and few-shot classification were formulated in the litterature, and show why despite the above-mentioned drawbacks to generalizable features, image retrieval is the most relevant approach for treating high volumes of data in the era of big data.

Chapter 5

A look at remote sensing photography

Remote sensing imagery consists of images taken by sensors with a long distance to the scene, often the Earth's surface. Historically, documenting territories was done with airplanes, which was for example the case with most oblique imagery in ALEGORIA as presented in section 3.2. With the growing interest for monitoring our planet in real time, there is now also a large volume of easily accessible satellite imagery. In this chapter, we will only consider photography in the visible spectrum, but remote sensing also includes data from the infrared and ultraviolet spectrums, along with non-image data such as LIDAR (laser point cloud) or RADAR (radio wave detection). Naturally, there are some inherent specificities to remote sensing imagery, including large variations in resolution depending on the technology, the absence of depth on vertical imagery, and the high repeatability of patterns (such as roads, trees..).

Remote sensing data has applications in many territory-related fields, such as forest management, urban planning, natural disaster prevention, etc. Accordingly, there has been an extensive research effort dedicated to designing tools for these applications, and with the arrival of deep learning, to transposing the success of convolutional neural networks on (to put it simply) internet images to remote sensing images [51, 59].

Images from satellites or planes are no exception when it comes to feeding them to CNNs: the first and most important question, as concluded in the previous chapter, is "do we have enough annotated data to properly represent our task and train a model to do it ?". Looking at the literature, we noticed a variety of methods with the keywords "cross-domain" or "few shot learning" [54, 117, 91, 66, 58, 3], indicating that in some cases, the answer is "no". However, we also noted a high number of datasets especially oriented towards land use, and considering that data can be found for free from providers such as the French Mapping Agency (IGN) or the USGS (US geological survey) and crossed with geographic information systems to automatically generate annotations, it is worth questioning to what extent it is necessary to design tools for low-data scenarii in the context of remote sensing, and how image retrieval, *i.e.* a data-driven approach, can bring solutions.

In this chapter, I will begin by introducing a new dataset in section 5.1, SF300, filling a gap in remote sensing imagery: geolocalization-oriented multi-angle images. Section 5.2 presents some experiments with a baseline retrieval method on SF300, with unexpected results leading us to formulate an hypothesis concerning image-level analysis in remote sensing. In section 5.3, I will present other datasets in remote sensing image-level analysis and methods for cross-domain or few-shot classification. Section 5.4 details our proposed retrieval-inspired framework to answer to these challenges, with the experimental setup in section 5.5, results in section 5.6 compared to the state of the art, discussion in section 5.7 and conclusion in section 5.8.

5.1 SF300 dataset

In chapter 3, we saw that there was a lack of training data especially for the aerial images of ALEGORIA. To my knowledge, there is no dataset involving aerial or satellite data with the task of connecting multiple images of the same location with varying visual characteristics (see section 3.1. Taking advantage of an opportunity to access aerial imagery with geolabels, I built a large-scale training dataset, SF300. It contains multiple images of the same location with varying orientations, which is similar to ALEGORIA. I will introduce the dataset in section 5.1.1 , how it was built in section 5.1.2, general statistics in section 5.1.3 and detailed statistics in section 5.1.4.

5.1.1 Presentation

The SF300 dataset consists of 512x512 pixels images, with 308,353 images in 27,502 classes in the train set, and 21,844 images in 2,421 classes in the test set. It has been constructed using raw high-resolution images provided by the Danish Institute of Data Supply and Efficiency Improvement (SDFE) in open-access. Each class corresponds to a real-world square footprint, and is composed of a varying number of either vertical (camera pointing the nadir direction) or oblique images (camera directed with known angle) of this geographical location.

5.1.2 Dataset construction

Following a whole country aerial acquisition by planes equipped with 5-angle cameras, the source images are available at the address `https://skraafoto.kortforsyningen.dk/`. We first collected all available high-resolution (∼100MP) images in a set of selected urban and semi-urban areas. Using the provided footprint coordinates, we matched n-tuples of images covering approximately the same zone. To enhance precision, we manually aligned the images by picking a common point for all images in each tuple. We then computed the homography matrices linking pixel coordinates to real-world coordinates for each image, which allowed an automated cropping of tuples into a varying number of smaller images of fixed size. Available parameters for the source images were propagated to the

Figure 5.1: Examples of images from the SF300 dataset. The images of the same row belong to the same class.

smaller images and stored in a .csv file for each class. This process was repeated on a smaller number of other locations to create the test set. Therefore there is no common class between the train set and the test set.

In this dataset, each location is represented with different orientations, sun angles, altitudes of images (which induces variations in geographic resolutions (*i.e.* the ground coverage of each pixel), and the azimuth is known. Figure 5.1 shows some examples of images of the dataset. We also plot the number of images per class in Figure 5.2.

The 5-angle cameras induce 5 possible vertical orientations for each location (vertical + four 45 degrees on the sides).

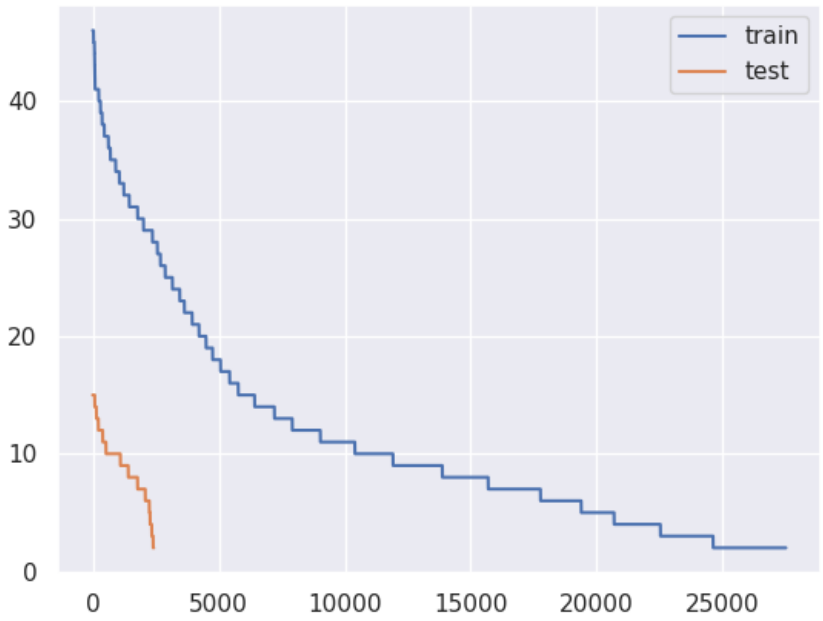

Figure 5.2: Distribution of the number of images per class in the SF300 dataset (x-axis denotes the class number ; y-axis denotes the number of images, with classes sorted from biggest to smallest class).

5.1.3 Statistics

Table 5.1: General statistics of the SF300 dataset

Item	**Value**
TRAIN	
Number of images	308353
Number of classes	27502
Min number of images per class	2
Max number of images per class	46
Mean number of images per class	11
Median number of images per class	9
Image file format	.jpg
Image dimension (width*height)	512*512
Disk usage	13 Go
TEST	
Number of images	21844
Number of classes	2421
Min number of images per class	2
Max number of images per class	15
Mean number of images per class	9
Median number of images per class	9
Image file format	.jpg
Image dimension (width*height)	512*512
Disk usage	1.2 Go

5.1.4 Attributes

The following attributes are available for each image in the dataset:

- Altitude (meters): altitude of the plane when the image was captured
- Omega (degrees): vertical orientation angle
- Phi (degrees): vertical orientation angle
- Orientation (0-5 integer): Quantized summary of Omega & Phi
- Kappa (degrees): azimuth angle
- Sun Angle (degrees): angle of the sun when the image was captured (provided with source images)

Attributes omega & phi correspond to the two angles defining the vertical orientation of the camera. After quantization, these two parameters can be reduced to a single value stating the vertical orientation of the image, as shown on Figure 5.3. 0 is vertical (omega=0° and phi=0°), and four other orientations are defined with tuples (omega=0°, phi=-45°), (omega=0°, phi=45°), (omega=-45°, phi=0°), (omega=45°, phi=0°). Value 5 is for all other orientations (rare cases of specific angle values). Table 5.2 shows the distribution of these values.

Image are all rotated to face north. The azimuth (kappa) is thus only an indication of how the plane was oriented when taking the picture but isn't visible on the images.

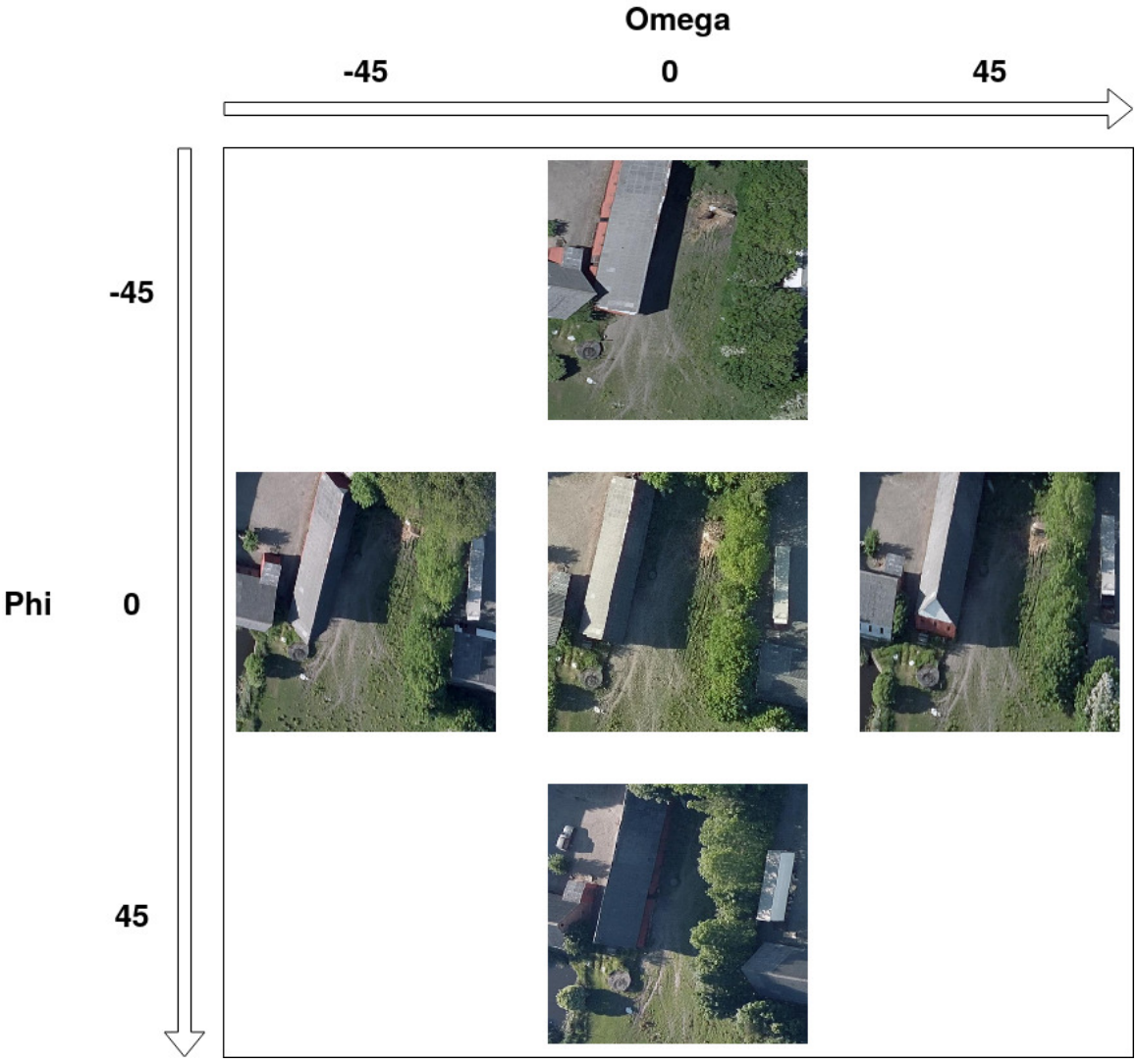

Figure 5.3: Vertical orientation cases in the SF300 dataset. Note how the four faces of the building are (partially) visible depending on the values of omega and phi. Cases where both omega and phi $!= 0$ occur rarely.

Table 5.2: Attributes distribution statistics of the SF300 dataset

Discrete attribute	Value	Train set proportion	Test set proportion
omega	0	19.53%	22.72%
	1	60.21%	55.65%
	2	18.36%	21.64%
	3	1.91%	0.0%
phi	0	66.99%	68.33%
	1	14.96%	17.52%
	2	16.18%	14.15%
	3	1.87%	0.0%
Orientation	0	27.20%	23.98%
	1	19.53%	22.72%
	2	14.96%	17.52%
	3	18.36%	21.64%
	4	16.18%	14.15%
	5	3.78%	0.0%

Figures 5.4, 5.5, 5.6, 5.7, 5.8, 5.9 additionally show graphical representation of attribute distributions.

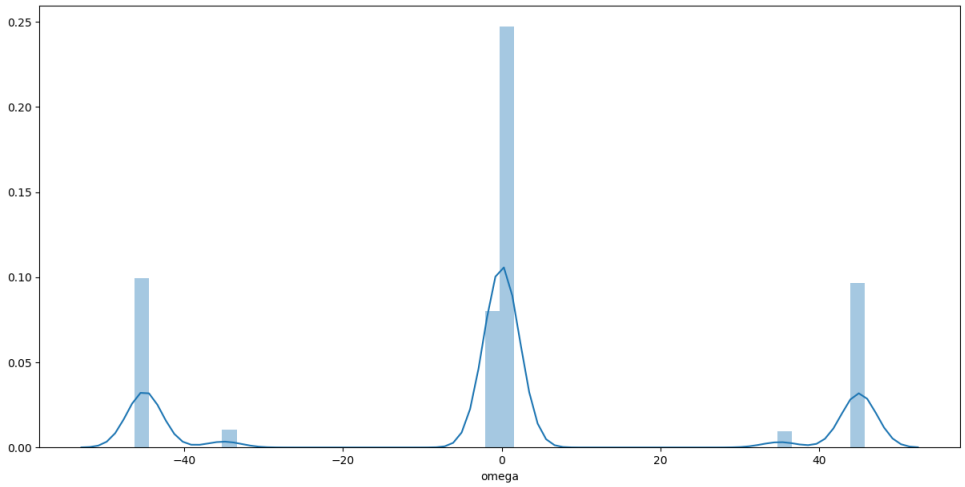

Figure 5.4: Distribution of omega values in the train set of SF300.

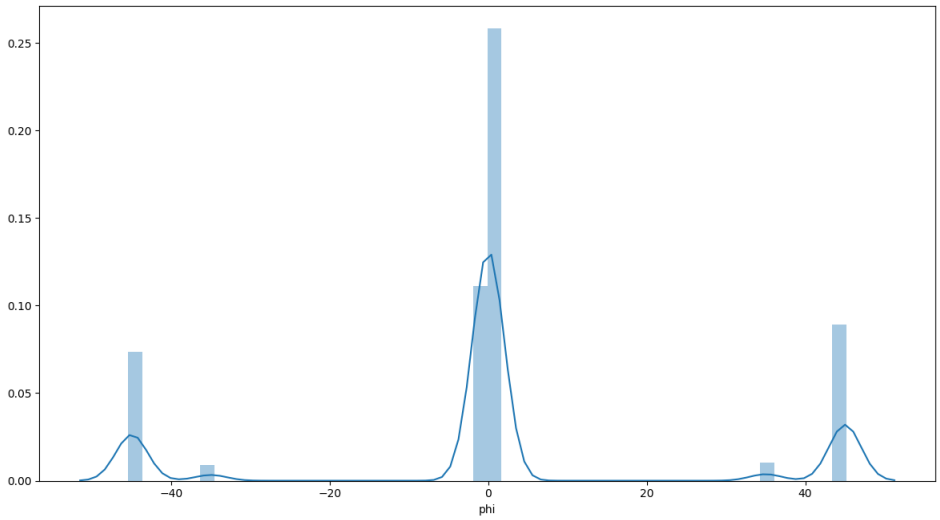

Figure 5.5: Distribution of phi angle values in the train set of SF300.

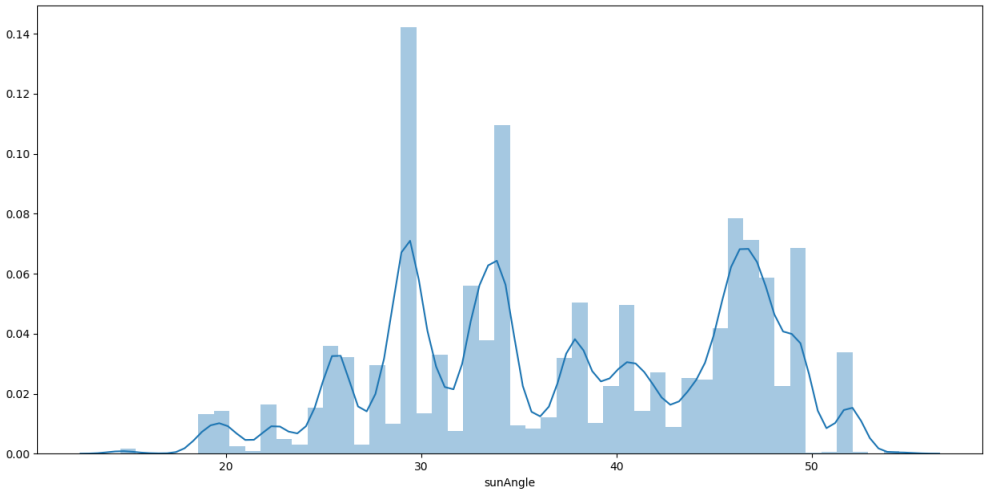

Figure 5.6: Distribution of sun angle values in the train set of SF300.

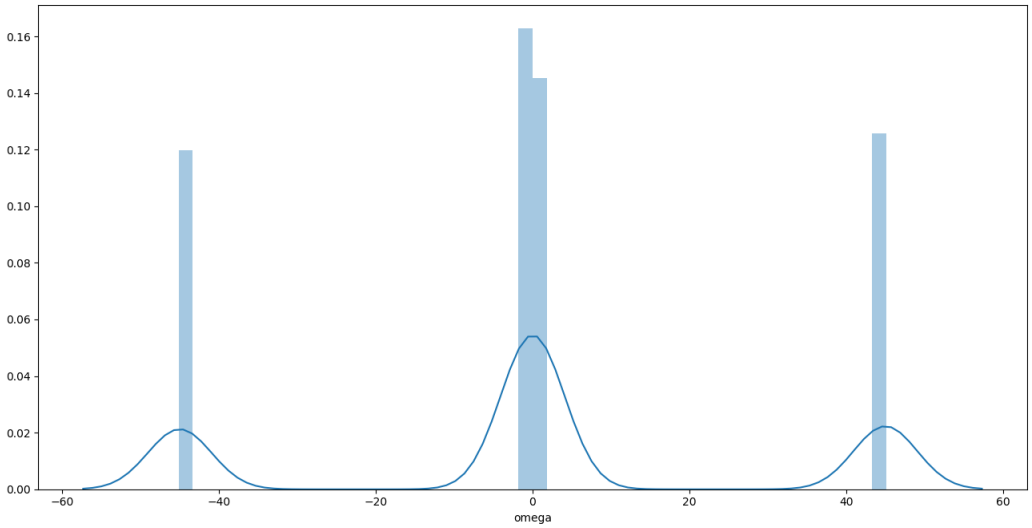

Figure 5.7: Distribution of omega values in the test set of SF300.

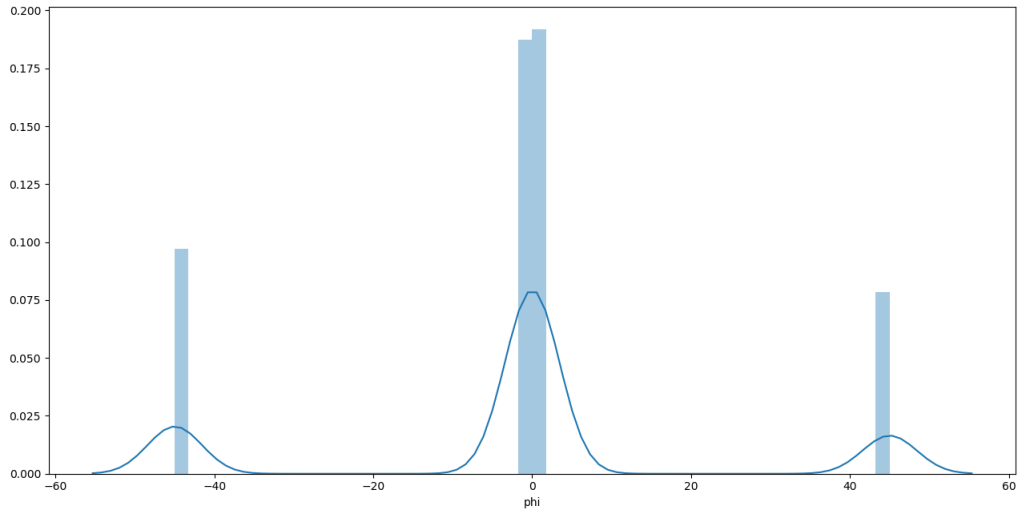

Figure 5.8: Distribution of phi angle values in the test set of SF300.

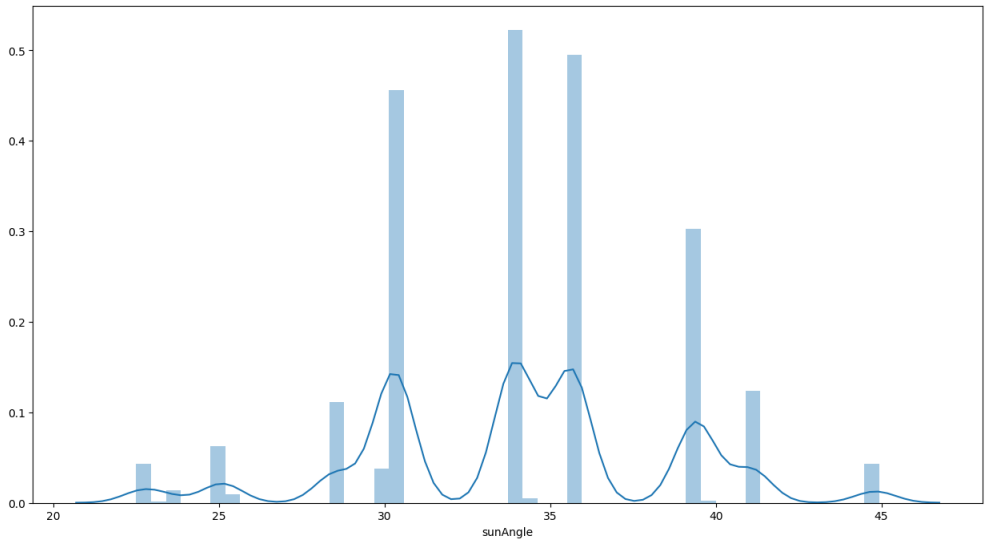

Figure 5.9: Distribution of sun angle values in the test set of SF300.

5.2 Training on SF300

A first experiment to characterize the difficulty of the SF300 dataset is simply to train with the train set and test on the test set, monitoring the evolution of the loss function and test metrics.

We use the same training setup as our GeM-ArcFace baselines in section 4.1.1. It consists of three networks shown in Figure 3.10:

- The feature extractor is any standard CNN, without the fully connected layers at the end. Here, we pick ResNet50. Deep features are pooled with the Generalized Mean, and normalized to unit norm.
- The whitening network consists of two fully connected layers, reducing the output dimension of the feature extractor (here 2048 with ResNet50) to a desired size (here we pick 512), and giving the network the ability to automatically learn an optimal separation of descriptors on the 512-dimensional hypersphere. The outputs are once again normalized to unit norm.
- The classifier is the network outputting a class prediction, but here we only use it during training, and we choose the ArcFace [20] classifier to enforce good class separability. It takes as input the normalized GeM descriptors, and assigns them to a class based on an angular proximity to learned class centroids.

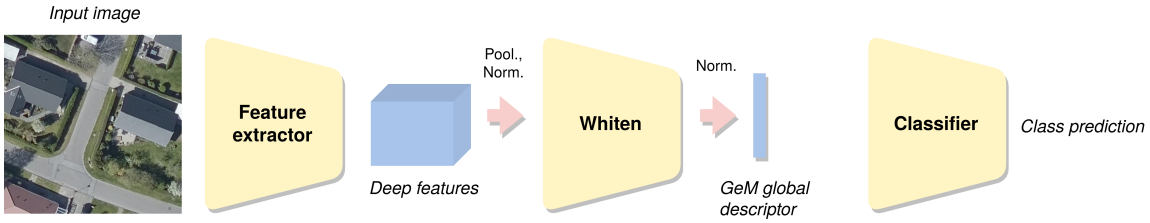

Figure 5.10: Training framework for GeM global descriptors with the ArcFace loss.

Figures 5.11, 5.12 and 5.13 show the evolution of important metrics during the training. All signals indicating that the models successfully learns accurate representations are found: the loss decreases, train accuracies increase and saturate, test peformance increases and saturates around a value of $\sim 93\%$ of mAP which can be considered near-optimal. This indicates a "clean" training dataset, *i.e.* suited for the training of regular CNNs. The fact that the test performance saturates indicates that the task is, from a scientific perspective, "solved" with a CNN. In chapter 4, we found a relatively better performance on vertical imagery with the GeM descriptor trained on SF300 than other descriptors, indicating that the learned representations can be transferred to other contexts.

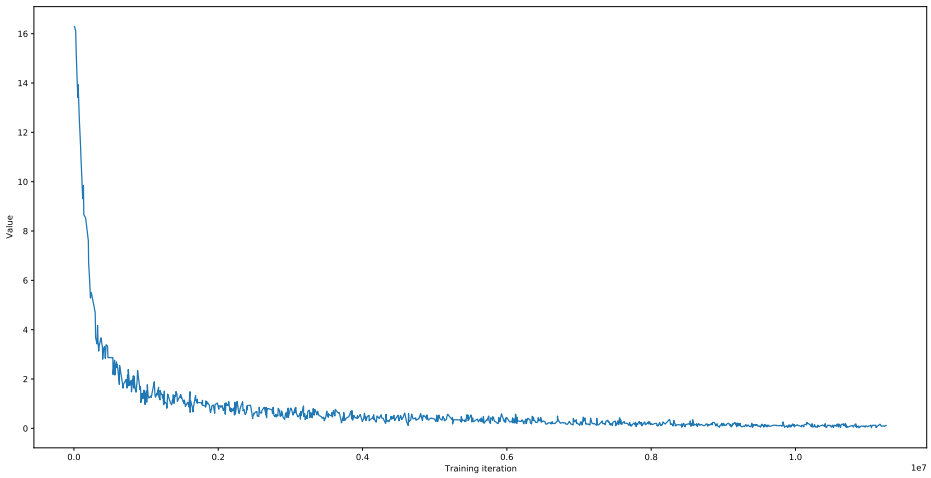

Figure 5.11: Evolution of the ArcFace classification loss during the training on SF300.

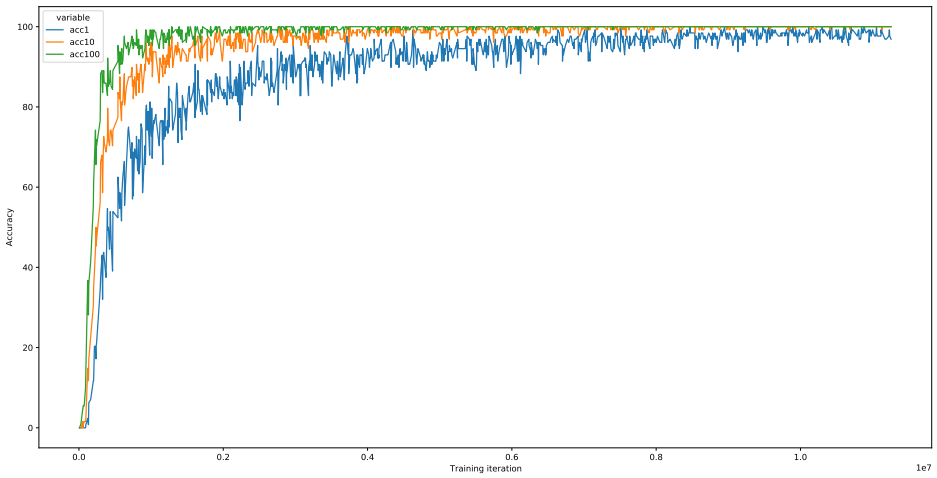

Figure 5.12: Evolution of classification accuracies during the training on SF300.

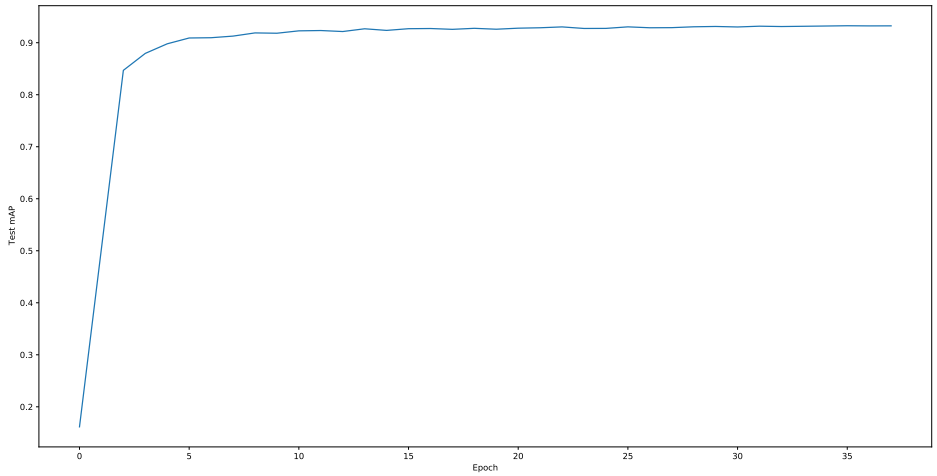

Figure 5.13: Evolution of the mAP on the test set of SF300 during training.

Similarly to ALEGORIA, having access to attribute values allow us to make a more detailed performance analysis. The 5 possible vertical orientations in SF300 can be used to define "domains", and consequently measure intra- and inter-domain performance with our proposed metrics from section 4.1.2. Even if the saturated absolute performance is already an indicator of low variance, intra- and inter-domain performance, or to use more adequate terms, intra- and inter-orientation

Absolute performance	Intra-domain performance					Inter-domain performance		
Globlal mAP	0	1	2	3	4	mP1	qP1	mAPD
93.62	96.92	92.94	91.72	92.64	92.99	2	2	5.2

Table 5.3: Detailed performance analysis for GeM-ArcFace on SF300, with absolute, intra- and inter-domain metrics. Here the term "domain" refers to the vertical orientation of images, which takes 5 possible values (rotations around two horizontal axis in both directions, and vertical)

performance, should give an idea of how robust the feature extractor is to orientation changes. Table 5.3 shows the values of our detailed performance analysis on SF300. The metrics are all near to optimal: intra-domain performance is almost saturated, with a slight bias for value 0 (completely vertical images), which is expected considering the bias towards this type of images in the training dataset. This minor bias is confirmed by inter-domain performance, with indicators close to 0, notably the mAPD of 5.2 which is excellent compared to the ~ 200 average positional deviation observed on ALEGORIA.

We did not expect such a good performance from a baseline method, on a large-scale dataset with many repetitive elements which can bring confusion, such as similar-looking buildings or roads. From a research perspective, it limits the interest of the SF300 dataset to being a training dataset for localization-oriented tasks. On the other hand, the saturation of performance gives a hint about remote sensing imagery: there is probably no visual characteristic making image-level analysis of satellite and plane images inherently different from any other computer vision data. Given enough training data and a baseline model, good performance can be reached without much overthinking. In the next sections, we will experimentally verify if this claim holds against other remote sensing datasets.

5.3 Land-use datasets and methods

Common remote sensing image-level analysis (ignoring detection or semantic segmentation applications) datasets are presented in Table 5.4. All of these datasets are designed for land-use classification (which we will refer to as RSC, remote sensing classification), *i.e.* assigning a single label (baseball field, industrial area, golf course...) summarizing the content of the image. Note that contrarily to the usual separation between training datasets and benchmarks in computer vision, these datasets are provided "as is", each with its own intrinsic characteristics (the most important being image size, spatial resolution, and class definition) and size.

Table 5.4: Overview of remote sensing image datasets (image-level analysis).

Name	Year	Image size (px)	Spatial resolution (m)	Nb. of classes	Images per class	Total size
AID [110]	2017	600	0.5-0.8	30	200-400	10,000
BCS [68]	2015	64	N/A	2	1438	2,876
PatternNet [124]	2018	256	0.062-4.693	38	800	30,400
RESISC45 [16]	2017	256	0.2-30	45	700	31,500
RSI-CB [49]	2020	256	0.3 - 3	35	609	24,747
RSSCN7 [126]	2015	400	N/A	7	400	2,800
SIRI-WHU [120]	2016	200	2	12	200	2,400
UCM [114]	2010	256	0.3	21	100	2,100
WHU-RS19 [109]	2010	600	<0.5	19	50	1,005

These 9 datasets constitute a great resource available to train and test methods, but the biggest to our knowledge (RESISC45) [1] has $31,500$ images, which is at least 10 times smaller than what is typically used in computer vision to fine-tune CNNs (ImageNet [19]: 14 million images, GoogleLandmarks [106]: 4.1 million, MS-COCO [52]: 330k). Starting from this observation, researchers have proposed way to fine-tune them efficiently with limited data [17, 28], or to artificially augment the volume of training data with data augmentation [85] or complex image generation methods such as Generative Adversarial Networks [112].

Works that considered the idea of training on a single dataset to test on another refer to their approaches as cross-domain [91, 66, 58, 3], *i.e.* that the semantics and visual characteristics of datasets are too distinct to allow direct transfer of learned knowledge. Concurrently, few-shot learning has recently gained interest in the field of remote sensing, addressing the potential lack of labeled data in real-case scenarios. Ideas borrowed from "general-purpose" have successfully been applied to land-use imagery [54, 117], but they have the major drawbacks of 1. Requiring training data with matching representation characteristics (*i.e.* same domain along the commonly adopted view in RSC), In [117] for example, 73% of RESISC45 is used to train and validate the method. 2. Using very small CNNs due to the high memory cost of episodic training. In [117] the reference performance is obtained with ResNet-12, a very small network.

[1] We could not download the RSD46-WHU [56] dataset whose access is reserved to Chinese citizens

Figure 5.14: Examples from common land-use datasets. From top to bottom: UCM, WHU-RS19, RSSCN7, AID ; corresponding classes indicated below images.

Figure 5.14 shows examples from four of the datasets presented in table 5.4. There is a very high visual similarity between these datasets. Moreover, considering that classes are defined as high-level semantic categories, it is intuitively a problem easier than the localization task of SF300: in WHU-RS19, instances of football fields, a very distinctive object, are grouped with a single label, whereas in SF300, the model must learn to differentiate locations using details that can be found in many other classes. Considering these elements, we argue that the main problem does not lie in a domain gap, but rather on a lack of volume and variety in training data, which can conveniently be solved with the abundance of training datasets, and with data-driven approaches, such as image retrieval. The next section presents a framework in this direction.

5.4 Data-driven classification with retrieval

Starting from the following observations and hypothesis:

1. Many land-use datasets are available,
2. These datasets are similar enough to be combined during training,
3. Image retrieval allows to avoid to define a context, *e.g.* for classifying images when class information is not available,

we propose to use a retrieval-based approach to conduct land-use classification, training on multiple datasets. The general framework is shown on figure 5.15: we reuse elements from chapter 4, including

the feature extractor with a pooling operation giving GeM global descriptors and the SmoothAP loss which allows optimization without using a classifier.

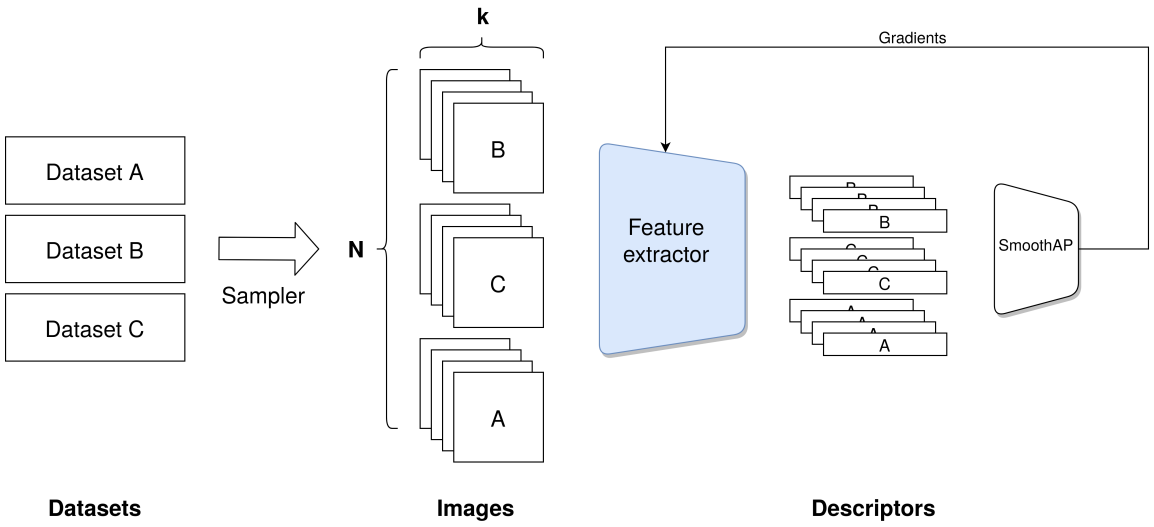

Figure 5.15: Multi-dataset training framework.

During training, we combine images from different datasets by 1. randomly sampling a dataset from the pool of datasets, 2. randomly picking one class from this dataset, 3. randomly picking k samples from this class. By repeating this operation N times, we build training batches of $N * k$ images from N different classes picked in the pool of datasets.

During testing, the few-shot classification task can be formulated as follows:

1. For each class of the dataset, select k random support examples.
2. Compute descriptors for the support, build class representations.
3. For all remaining images in the dataset, get the closest class.
4. Assign the label of prototypes to the matched images, evaluate.

To identify the closest class to each query, we use the Mahalanobis distance with estimated class covariance matrices, as proposed by Bateni et al. [9]. By inserting covariance information, the measure of distance is more precise, taking into account the distribution of support representation in the feature space. The distance between query descriptor x and class c, represented by its representation μ_c, the mean of its support descriptors, is computed with:

$$d(x, c) = \frac{1}{2}(x - \mu_c)^T (Q_c)^{-1}(x - \mu_c), \tag{5.1}$$

where Q_c is a class-specific covariance estimate, computed with:

$$Q_c = \lambda_k \Sigma_k + (1 - \lambda_k)\Sigma + I, \tag{5.2}$$

where Σ_c is the class-specific covariance, Σ is the global covariance (the covariance of all support vectors), and I the identity matrix. The weigthing factor λ_k is computed as $\lambda_k = k/(k+1)$. The rationale behind the weighted combination is to balance class information (Σ_c) and dataset information (Σ), with a higher weight given to class information when there are more support images. For $k = 1$, we have $Q_c = 0.5\Sigma_c + 0.5\Sigma + I$. For $k = 5$, we have $Q_c = \frac{5}{6}\Sigma_c + \frac{1}{6}\Sigma + I$.

Compared to how few-shot learning is usually evaluated in the literature, there are two main differences with our framework. The first is the number of classes used in evaluation. Regular few-shot learning is formulated with the k-shot, N-way notation, where N is the number of classes used during a testing episode, and k the number of support examples given for each of these N classes. In our framework, we evaluate with N set to its maximal value, *i.e.* the total number of classes in the dataset. We can refer to this setup as k-shot, all-ways. This difference makes a significative raise in difficulty, because of the higher potential for inter-class confusion. The second difference is the number of samples evaluated for each class, in regular few-shot learning, for each episode, performance is evaluated on a restricted set of queries, *e.g.* 10 [101] or 15 [117]. In our framework, we evaluate with all queries belonging to the same class. This difference should not modify the results, since it is merely an augmentation of the sample size used to compute the accuracy.

Our model is trained for 15 epochs. For each target dataset, we train on all the other datasets available. Additionally, we insert a post-processing step after computing descriptors and before computing the mahalanobis distance for classification: we use αQE, presented in section 4.2.2, with $\alpha = 3$ and $n = 10$. We do not fine-tune these parameters, following our conclusion that αQE diffusion schemes are quite robust to their modification (section 4.2.4).

5.5 Evaluation setup

We will compare our proposed framework to few-shot and cross-domain methods for land-use analysis, reporting the best results observed in the literature for each dataset. The evaluation setups differ depending on the method.

Few-shot methods in the literature are trained and evaluated on a single dataset. First, a target dataset is picked and split in meta-train, meta-val and meta-test splits. The model is trained and selected using the meta-train and meta-val splits respectively. The model is then evaluated as follows for the 5-way setup:

1. $N \in 1, 5$ support images are sampled from 5 different classes.
2. A query set (whose size is not standardized in the literature) is sampled from the same 5 classes.

3. The model assigns labels to each image in the query set.
4. Accuracy is computed for this query set.

This process constitutes a meta-testing episode and is repeated a high number (again, not standardized) of times to compute the final accuracy with 95% confidence intervals.

Cross-domain methods in the literature are trained on a source dataset and evaluated on a target dataset. A classification model (taking an image as input and producing a predicted label as output) is trained on the suorce dataset. It is then evaluated as follows:

1. A target dataset is picked and filtered to only keep common classes with the source dataset.
2. The model is used to predict labels on the filtered target dataset.
3. Accuracy is computed for these predicted labels.

Our proposed N-shot, "all-ways" method can be trained on any dataset or combination of datasets. The evaluation setup goes as follows:

1. A target dataset is picked.
2. $N \in 1, 5$ support images are sampled for each class in the target dataset.
3. A query set is constituted by taking all remaining images in the dataset.
4. The model assigns labels to each image in the query set.
5. Accuracy is computed for this query set.

There are two main sources of randomness in our experiments: dataset and class sampling during training, and support sampling during testing. Accordingly, we build 5 different models for each target dataset, and test them 5 times with a different support set, which gives 25 samples per experiment. 95% confidence intervals with Student's t-distribution are indicated.

5.6 Results

Table 5.5 shows the results of various state-of-the-art methods in land-use few-shot and cross-domain classification, compared to our proposed framework of data-driven classification.

A few important notes are necessary to ensure a proper comparison:

- Few-shot methods are trained in a meta-training setup: datasets are separated in meta-train, meta-val and meta-test splits. This means that only a fraction of the dataset is used during testing, which can be deduced in table 5.5 as the remainder after substracting indicated meta-train and meta-val ratio from 100%.

- Cross-domain methods are trained on the indicated source dataset(s). There is however often a discrepancy between classes defined in the source dataset and in the target dataset. Accordingly, methods in the literature only test on the fraction of the target dataset which contains the same classes as the source dataset. For example, if the class "Airplane" is absent from the dataset AID, images from this class will not be used when evaluating a model trained on UCM. This fraction is indicated as the test ratio in table 5.5.

- For our framework, we train on all datasets except the target dataset, and test on the target dataset. This means that we do not use a single image from the target dataset neither for training nor for validation, and test on the whole dataset.

Table 5.5: Comparison of land-use classification methods in different setups, including supervised, cross-domain, and few-shot classification, against our proposed data-driven framework with multi-dataset training for "all-ways" few-shot classification.

Dataset name	AID [110]	PatternNet [124]	RESISC45 [16]	RSI-CB [49]	RSSCN7 [126]	SIRI-WHU [120]	UCM [114]	WHU-RS19 [109]
Supervised classification								
OA (%)	97.21 [55]	-	95.17 [55]	99.66 [86]	98.89 [2]	97.83 [125]	98.93 [17]	97.50 [104]
train ratio	50%		20%	80%	50%	50%	80%	40%
Few-shot classification (5-way)								
1-Shot OA (%)	56.32±0.55[48]	-	69.46±0.22 [117]	-	-	-	57.23±0.56[48]	68.27±1.83[50]
5-Shot OA (%)	74.48±1.11[48]	-	84.66±0.12 [117]	-	-	-	76.08±0.28[48]	79.89±0.33[50]
meta-train ratio	33%		41%				33%	45%
meta-val ratio	33%		13%				33%	25%
Cross-domain classification								
OA (%)	70.94[3]	83.91[58]	77.33[58]	-	-	-	80.50[3]	-
test ratio	44%	4%	29%				62%	
source dataset(s)	UCM	RESISC45,AID,UCM	AID,UCM				AID	
Ours: few-shot cross-domain classification, "all-ways"								
1-Shot OA (%)	51.51±1.40	73.75±1.42	44.15±1.31	65.21±1.90	53.17±2.63	48.17±2.6	64.68±2.09	83.93±1.52
5-Shot OA (%)	72.58±0.94	86.35±0.92	64.80±0.78	84.57±0.52	67.42±2.00	62.83±1.50	82.08±0.91	91.07±0.65

We conducted an ablation study to see the influence of using the αQE diffusion method, results are indicated in table 5.6.

Target dataset	1-shot	$+\alpha QE$	5-shot	$+\alpha QE$
AID	48.94±1.23	51.51±1.40	70.36±0.70	72.58±0.94
PatternNet	70.11±1.26	73.75±1.42	84.27±0.82	86.35±0.92
RESISC45	40.97±1.15	44.15±1.31	61.93±0.65	64.80±0.78
RSI-CB	62.28±1.71	65.21±1.90	81.91±0.47	86.35±0.52
RSSCN7	51.93±2.45	53.17±2.63	66.93±1.77	67.42±2.00
SIRI-WHU	47.03±2.79	48.17±2.96	63.62±1.26	62.83±1.50
UCM	61.97±1.87	64.68±2.09	80.64±0.88	82.08±0.91
WHU-RS19	80.66±1.38	83.93±1.52	90.58±0.69	91.07±0.65

Table 5.6: Effect of adding αQE in our few-shot land-use classification method.

5.7 Discussion

5.7.1 Comparison to cross-domain methods

The comparison to cross-domain methods is made hard by the fact that they are tested on subsets of target datasets in the literature. On AID and UCM, our 5-shot setup obtains better performance than the state-of-the-art, using only $5 * 30 = 150$ support images for AID and $5 * 21 = 110$ for UCM, which represent respectively 1.5% and 5.2% of the datasets. By contrast, cross-domain methods perform evaluation on subsets eliminating between 38% on UCM and up to 96% on PatternNet of images in the target dataset. Compared to these methods, our data-driven approach does not require any class redefinition, and the competitive performance we obtain with multi-dataset training shows that the influence of an hypothetical domain gap between land-use datasets is probably over-estimated.

5.7.2 Comparison to few-shot methods

Similarly, the comparison to few-shot methods is made hard by the fact that they are 1. tested on subsets of the target dataset, 2. trained on subsets of the same dataset, 3. tested with episodes, *i.e.* giving the average accuracy when tested on a limited number of classes. Nonetheless, we are able to reach unprecedented performance on UCM and WHU-RS19 with margins of $\sim 7\%$ and $\sim 5\%$ respectively in the 1-shot setup, without seeing a single image of the target dataset during training.

5.7.3 Influence of diffusion

Table 5.6 shows that using a simple diffusion scheme brings a significant boost in accuracy, up to $\sim 3\%$ depending on the dataset and test setup. This experiment indicates that the diffusion principle

borrowed from image retrieval stays relevant on a classification task. On SIRI-WHU with the 5-shot setup however there is a slight decrease in performance, indicating that in some cases αQE worsens the quality of descriptor by taking into account unwanted noise. This can probably be solved by fine-tuning the parameters of αQE, which we restrain ourselves from doing to avoid using images from the target dataset (apart from support).

5.8 Conclusion

In this chapter, we explored remote sensing image analysis from a data perspective. We introduced the SF300 dataset, the first multi-orientation aerial image dataset with a task of geolocalization. Experiments on SF300 show that a modern deep learning-based method is able to successfully learn accurate and compact descriptors, encouraging us to generalize to other tasks. We thus proposed a data-driven approach inspired from image retrieval, making use of the abundance of datasets for land-use classification with multi-dataset training, and showed that this framework gives competitive results on multiple datasets against cross-domain and few-shot tasks, without using any image from the target dataset during training or restraining the evaluation setup. This should be strong evidence that image retrieval and more broadly data-driven approaches not relying on a predefined context are relevant for other tasks including classification, especially with the ever-growing volume of data in remote sensing and general-purpose computer vision. The comparison to cross-domain and few-shot tasks is also, if not a proof, a strong argument against the claim that land-use datasets are visually different enough to justify the "domain" appellation and corresponding methods such as domain adaptation or meta-learning.

Chapter 6

Conclusion

Before diving into the key takeaways of this thesis, I invite the reader to go to Appendix D for a rapid aside on two approaches I considered and later abandoned due to unconclusive results, but that are still worth mentioning in my opinion: generating new views from aerial images, and learning robustness to orientation variations with adversarial learning.

In this thesis, we first covered in chapter 2 the application context motivating this work, along with the scientific context defining what is worth working on, notably the quest for generalization which remains an open challenge in the era of deep learning. In this chapter, we already outlined an intuition: with a blurry line between algorithm and data, the factors for generalization also become unclear.

We continued with a formulation of the problem of interconnecting images across sources in chapter 3, showing how the ALEGORIA problem is a good example of the quest for generalizable features. The review of datasets and methods showed that there are countless ways of approaching the problem, with supervised methods promising high accuracy but relying on relevant annotated data, and less-supervised methods promising a decoupling of data and algorithm but sill in their early stages of development, often with low performance. We however argued that in a situation where there is little context to rely on, image retrieval has the advantage of being data-driven and thus more prone to generalization.

In chapter 4, we compared some state-of-the-art methods from supervised and unsupervised learning, and showed that a global, summarized analysis is not enough to compare methods in details. With the ALEGORIA benchmark and our proposed metrics, we were able to make a fine-grained analysis, showing that methods might perform well on a situation but badly on another, and that this performance probably mostly depends on the training dataset. To exploit the comparative advantages of the various descriptors, we proposed a multi-descriptor diffusion method, and showed that this approach is indeed able to combine their performance in a constructive way. The constrained variant highlighted the compromise between intra- and inter-domain performance. We concluded with some

preliminary results with rCNAPS, a few-shot baseline adapted to image retrieval, and reported encouraging results showing that it is possible to gain performance by exploiting on a limited set of annotated images. An application scenario would be, for example, a user already having some images of a given object and wanting to retrieve other images of the same object, specifically.

The results with the ALEGORIA benchmark showed that there are still major obstacles before claiming that deep descriptors can generalize through various domains and without training data. However, we still believe that the task of image retrieval is a serious candidate towards generalization and in the context of big data. To verify this, we studied the case of remote sensing data in chapter 5. Beginning with our proposed SF300 dataset and associated experiments, we surprisingly did not report any major difficulty, and formulated the hypothesis that deep descriptors actually have all they need (*i.e.* training data and accurate models) to perform well on this type of data. We proceeded with a new framework for few-shot classification, and obtained competitive results on an ensemble of land-use datasets.

Through this thesis, we covered ideas, methods and results, with a major common point in their formulation: the data defines a context, on which models are applied to conduct a task, not the contrary. There is an abundance of technical propositions in the literature, and they are mostly agnostic of the training dataset as long as annotations are available. Consequently, the virtually infinite number of dataset-method combinations already constitutes a solid basis for building a detailed understanding of the visual world. What is missing now are connections between images with different characteristics, if we want to move towards generalizable models. We showed in this thesis that image retrieval and more broadly data-driven approaches have a high potential in this context, and proposed some solutions connecting different sources of data either during training or testing. Intuitively, data-driven approaches converge with how we, humans, gain understanding: when presented with a new object or concept, we rely on our previous experience and similar objects or concepts that we use as references.

Going further, the works presented here can be continued by broadening the panel of dataset-methods combinations used. If I were to outline a prototype for "the" generalizable feature learning model, it would:

1. integrate a "megabase" of reference data, an ensemble of annotated datasets,
2. use a similarity criterion (as in image retrieval) to assess the relevance of available data to the target data,
3. use another criterion to assess the heterogeneity of the target data,
4. combine this information with a panel of methods, to automatically define the most suited data-method framework to conduct the given task.

The key scientific locks here are in the definition of the criterions, and the selection of "the most suited" framework. The local/global descriptor paradigms in image retrieval are an example of tools

for building criterions, but there are many more in other fields of computer vision, notably using a 2D object-oriented approach with detection [92] or segmentation, 3D or 4D (with time) modelling for a better understanding of real-world properties [25, 84], other modalities [72]... But more often than not, such approaches are not data-agnostic and thus sacrifice generalization for specificity. Regarding the automated selection of algorithm and data, meta-learning offers solutions with the idea of learning to learn, but such a model would involve many steps of which some might not be differentiable (notably the selection operation). Meta-learning is still a nascent field of research, it is not well understood if the current deep learning framework can go beyond the single task formulation (such as few-shot learning) and generalize across domains and tasks. I will end with this rough picture for the future of generalizable features, and with a quote that might sound a bit cliché but which expresses well my personal stance on the recent craze for artifical intelligence, and makes an invitation for not considering models for more than what they are, tools:

> Computers are like humans - they do everything except think.
>
> John von Neumann, 1947

Appendix A

Influence of absolute position in the AP measure

Consider a "perfect" list of results: M positive images all positioned in an arbitrary order at the top of list, before any negative image. To see the influence of the positioning of images, let us compute the variation of the mAP if image at position p is switched to a negative.

We begin with the AP formula:

$$AP = \frac{1}{M} \sum_{k=1}^{M} P(k) * rel(k) \tag{A.1}$$

with $P(k)$ the precision at rank k, *i.e.* the fraction of positive images in the list from positions 1 to k, and $rel(k)$ the relevance of image at rank k (1 if positive, 0 if negative).

If we have :

$$rel(k) = \begin{cases} 1 & \forall k \neq p \\ 0 & k = p \end{cases} \tag{A.2}$$

the AP becomes, setting N (number of positive images) to $M - 1$:

$$AP = \frac{1}{M-1}(\sum_{k=1}^{p-1} P(k) * rel(k) + \sum_{k=p+1}^{M} P(k) * rel(k)) = \frac{1}{M-1}(p - 1 + \sum_{k=p+1}^{M} \frac{k-1}{k}) \tag{A.3}$$

Figure A.1 plots some curves with different values of M. Note how, for small values of M, *i.e.* classes with few positive examples, the absolute position of positives has a high impact on the AP.

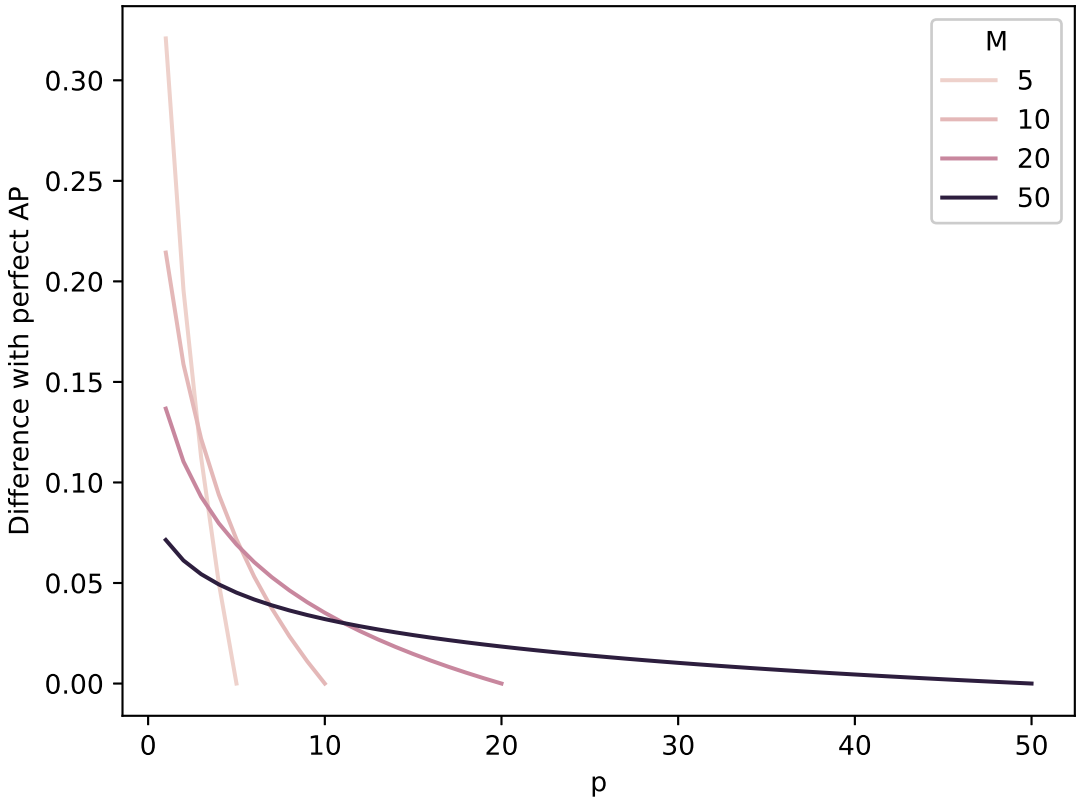

Figure A.1: Influence of inserting a negative image at position p in a "perfect" list of results, with list of length M and $M-1$ positive images in total. Reading example: for a class with 4 samples and a perfect list of results, if a negative image is inserted at position 1, the AP will decrease by 0.33.

Appendix B

Mathematical formulation of local feature matching methods

I will summarize here the mathematical framework proposed by Tolias et al. [94] which can be used to describe the BoW, VLAD, HE and ASMK matching methods.

In the visual word paradigm, image A is described by the set $S = x_1, ..., x_n$ of n d-dimensional local descriptors. The descriptors are quantized (clustered) by a k-means quantizer

$$\begin{aligned} q : \mathbb{R}^d &\to C \subset \mathbb{R}^d \\ x &\mapsto q(x) \end{aligned}$$

where $C = c_1, ..., c_k$ is a codebook comprising k vectors, the visual words. With $X_c = \{x \in X : q(x) = c\}$ the subset of descriptors in X that are assigned to a particular visual word, the family of functions that produce a similarity score between sets of descriptors has the form:

$$K(X, Y) = \gamma(X)\gamma(Y) \sum_{c \in C} \omega_c M(X_c, Y_c), \tag{B.1}$$

where function M is defined between two sets of descriptors X_c, Y_c assigned to the same visual word. ω_c is a constant depending on the visual word c, for example the IDF weighing term. $\gamma(.)$ is a normalization factor, typically computed as

$$\gamma(X) = \Big(\sum_{c \in C} \omega_c M(X_c, X_c) \Big)^{-1/2}, \tag{B.2}$$

such that the self-similarity of an image is $K(X, X) = 1$.

The Bag-of-Words (BoW) representation can be written as:

$$M(X_c, Y_c) = \sum_{x \in X_c} \sum_{y \in Y_c} 1. \tag{B.3}$$

Hamming Embeddings (HE) corresponds to:

$$M(X_c, Y_c) = \sum_{x \in X_c} \sum_{y \in Y_c} w(h(b_x, b_y)), \tag{B.4}$$

where h is the Hamming distance (difference of binary vectors) computed on b_x and b_y, the binary representations of descriptors x and y with B bits, and w a weighting function such that $w(h) = 1$ if $h \leq \tau$, and 0 otherwise.

The Vector of Locally Aggregated Descriptors (VLAD), even if is not computed by iterating over visual words in practice, is also included in this family with:

$$M(X_c, Y_c) = \sum_{x \in X_c} \sum_{y \in Y_c} r(x)^T r(y), \tag{B.5}$$

where $r(x) = x - q(x)$ is the residual vector of x.

The generalization of these approaches, the Aggregated Selective Match Kernel (ASMK) is formulated as:

$$M(X_c, Y_c) = \sigma\left\{ \psi\Big(\sum_{x \in X_c} \phi(x) \Big)^T \psi\Big(\sum_{y \in Y_c} \phi(y) \Big) \right\}, \tag{B.6}$$

where σ is a selectivity function:

$$\sigma(u) = \begin{cases} sign(u)|u|^\alpha & \text{if } u > \tau \\ 0 & \text{otherwise,} \end{cases}$$

ϕ is the same aggregated representation as VLAD:

$$\phi(x) = r(x), \tag{B.7}$$

and ψ is a normalization step:

$$\psi(x) = \frac{x}{\|x\|}, \tag{B.8}$$

In the end, ASMK is very similar to VLAD but adds a per-cell (per visual word) selectivity function that has been proven useful for handling false positives.

Appendix C

Colorizing grayscale images

The ALEGORIA benchmark includes 883 grayscale images (see Table C.1 for the whole color distribution) and 236 monochrome images (*e.g.* sepia), which are in terms of contained information similar to grayscale images. Colorizing these images seems a promising idea to reduce the gap in representation characteristics, especially considering the recent successes with generative models such as GANs [31].

Color	Grayscale	Monochrome	Infrared	Other	Total
719	883	236	11	9	1858

Table C.1: Color distribution in the ALEGORIA benchmark

The main issue in colorization is that the problem is inherently under-constrained: there are multiple colorizations that can be considered realistic or satisfying for a given image (imagine different cameras, or Instagram filters, or color variations due to the weather.. producing images with the same geometric structure but different colors). But in return, it is very straightforward to define training setups for colorization: most popular training datasets are in color, their images can simply be converted to grayscale to produce both a training sample and its ground truth. We identified two methods with available code and pretrained models, which allowed us to quickly apply them to ALEGORIA.

Iizuka et al. [35] frame the problem with dual optimization: the model produces a class label and a chrominance (color) map for training images. The part of the model used for classification plays the role of a supervisor, providing global priors (information about the semantic content) to the other part of the model which generates color information. The available model was trained on the Places365 [123] dataset, a large-scale scene recognition dataset with mostly outdoor environments. In the experiments, author show impressive results in historical photography colorization.

Zhang et al. [118] notice that optimizing with a loss averaging the difference between predicted pixel

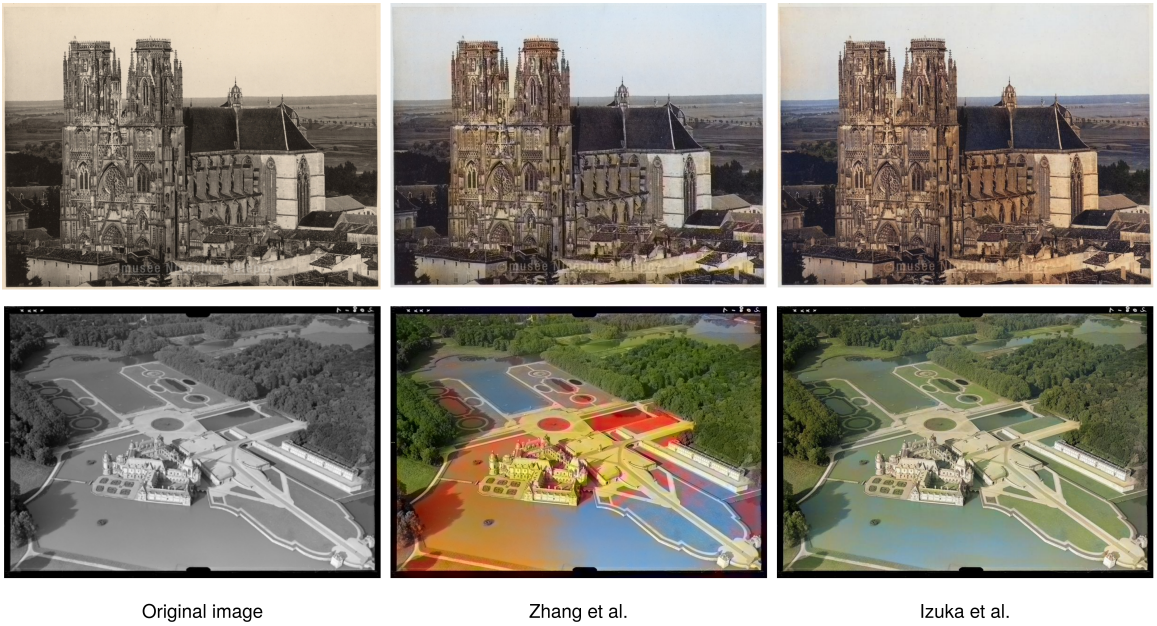

Figure C.1: Qualitative results of colorization models.
Photo credits top to bottom: IGN, Musée Nicéphore Niepce. See Alegoria rights.

color values and real values produces grayish, desaturated results. To solve this, the model outputs distributions of probabilities in the Lab color space, and a function transforms these distributions into a single value to get the final result. With this reformulation, they can fine-tune the "vibrancy" (intensity) of colors in the output image. Their results indeed show colorful images, avoiding the usual grayish veil produced by other methods. The model is trained with ImageNet.

We qualitatively and quantitavely evaluated these two methods with an early version of the ALEGORIA benchmark, containing less images and variety.

Figure C.1 shows visual results of the two colorization algorithm applied to ALEGORIA images. We observe that the algorithm of Zhang et al. [118] does produce richer and diverse colorizations, but is sometimes unable to understand the image semantics and outputs unexpected color patterns (the orange stain on the bottom image). Comparatively, Iizuka et al. [35]'s model is less audacious, but has a tendency of blurring color across geometric structures (the green water on the bottom image), like passing a damp brush after color crayons.

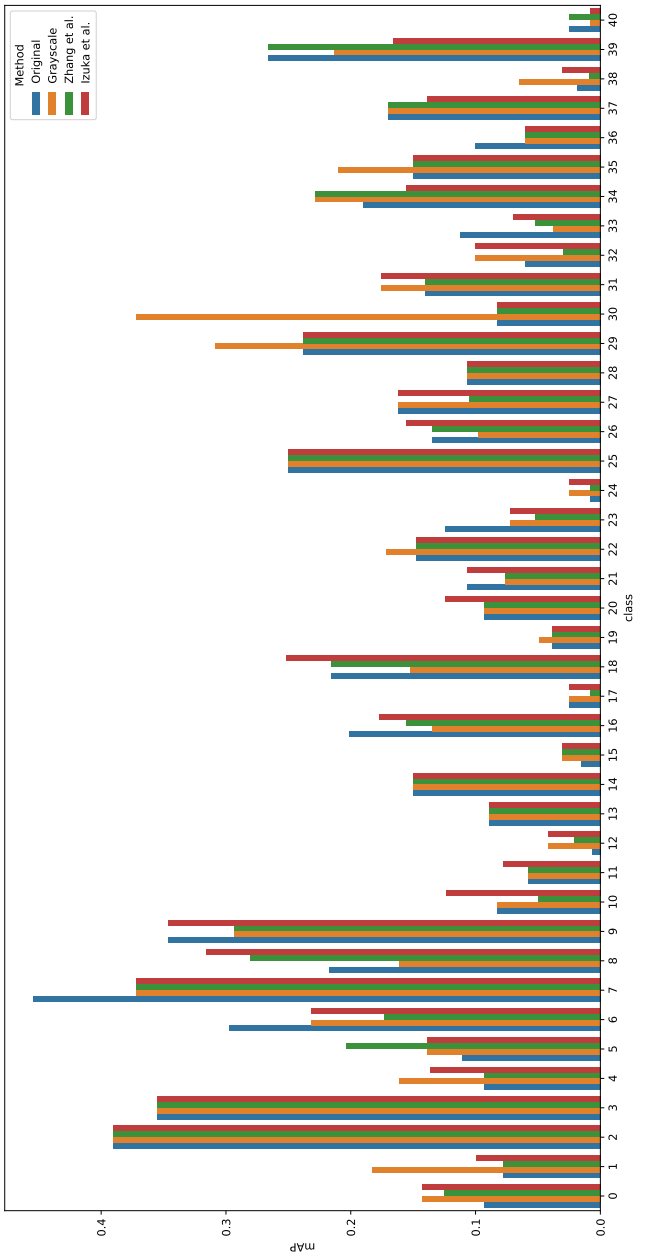

Figure C.2: Quantitative results of colorization models: classwise evaluation of retrieval performance.

Figure C.2 shows the results for a quantitative evaluation of retrieval performance with colorized images, compared with their grayscale and untouched versions, class by class. The results shown here should not be compared with the results in Chapter 4 because they were not made on the same version of the ALEGORIA benchmark (results here are with an early version, with less classes, diversity and annotations). The classwise mAP does not indicate the superiority of a given method here, except on specific classes dominated with a color pattern (*e.g.* class 30 containing only aerial grayscale images, a setup where our compared colorization algorithms are particularly bad because they were not trained with similar data).

Original	Grayscale	Zhang et al.	Izuka et al.
14.65	15.23	13.68	14.79

Table C.2: Global performance (mAP) with colorization models.

Table C.2 synthesizes the performance with a single absolute mAP value. Experiments here were made with the DELF local descriptor [63], very similar to HOW. We used the same parameters as in the authors' paper. The global mAP does not indicate a superiority of colorization methods, rather it seems better to simply switch all images to grayscale, which reduces the variance of color characteristics.

Appendix D

What did not work

Computer vision is a research field where, luckily, experiments are easy to code and run, but the downside is that with deep learning, a major part of the experimental work consists in making sure that all the parts of a model function properly and especially that the optimization process goes well. It is often hard to identify if a given experiment fails because the starting hypothesis is false, or simply because the model requires some more engineering to function properly. And sometimes, intermediate results indicate that there might be a path to follow, but the whole framework fails to produce state-of-the-art results on a full task, and has thus low interest for a rapidly evolving field which mostly cares for improved performance on standard benchmarks (with reason, otherwise the litterature would be saturated with unsound methods). However, I think that some experiments, even if they did not produce state-of-the-art results, are worth mentioning with the right context and motivations.

Generating training samples is an ambitious idea which has naturally followed the proposal of Generative Adversarial Networks (GANs) [31], an influential family of models successfully generating realistic images. The reader will probably have already experienced the astonishement after reading that the super realistic faces he sees on this internet article are generated. In cases where training data is missing, a common approach in deep learning is data augmentation, *i.e.* artificially increasing the volume of samples, either by modifying available samples with visual transformations, or by generating new samples. GANs can do both. In a traditional GAN model, there is a generator network producing fake images, and a discriminator trying to separate fake and real examples. During training, the generator is optimized to fool the discriminator, and the discriminator penalizes the generator until it is able to produce realistic samples. GANs are notoriously hard to optimize due to this "network competition" and to the high complexity of the visual world, hence the different variations that have been proposed to enhance them [5, 32, 43]. To assess 1. if GANs can generate realistic samples of aerial images, 2. if GANs can generate realistic samples of the same object under varying orientations, 3. if the generated samples are informative enough to gain some accuracy on

a downstream classification or retrieval task, we trained a GAN (with the wasserstein loss [5]) to generate fake SF300 training samples. Figure D.2 shows the results of the model after convergence, to be compared with the real samples in Figure D.1. These generated samples were picked as the most realistic across multiple trained models with varying hyperparameters and architectures. We note that the model successfully learned "high-level" visual and color patterns, such as the road in the top left image or the houses in the bottom left image. But images lack geometrical structure, object boundaries are blurry making it hard to distinguish exactly what is represented on the image. This is particularly problematic for discriminative applications such as image retrieval (visual retrieval with ALEGORIA or land-use classification): if the model fails to produce realistic images of an imaginary location, it will most certainly fail to produce visual variations of a real location. With these preliminary results and after noticing that GANs are very rarely used for discriminative applications (classification or retrieval) or with a very limited gain in accuracy [82], I decided to not pursue this idea further. However, we note that some satisfying results have been obtained in the similar task of generating human faces with varying orientations [60].

Figure D.1: Real SF300 samples.

Figure D.2: Generated SF300 samples using a Wasserstein GAN.

Learning invariance to visual variations should not be problematic: CNNs have proven multiple times that they can accomplish many tasks given enough training data and proper annotations. The difficulty comes when the invariance must be jointly learned with discriminative properties. Classification or retrieval models typically ignore the variance of training samples and assume that the model will implicitely learn invariance or at least robustness to visual variations, given enough training samples. And this hypothesis can be reasonably considered correct when the training loss converges to zero.

Going further, we explored the idea of enforcing invariance to a specific visual variation, with the idea that the resulting features would be more easily re-used for another task. Concretely, the problem was formulated as training a CNN to produce features robust to changes in vertical orientation on SF300, with the goal of applying these features to ALEGORIA later. The architecture consisted of a standard feature extractor and an orientation prediction network, taking as input the features and outputting a prediction on the vertical orientation of corresponding images. A first experiment where we optimized both networks separately confirmed that it was possible to deduce the vertical orientation of images only by looking at the features, even when the feature extractor was fully trained. It proved that some orientation information was included in the features, even if it is not relevant to the task of separating classes. In a second experiment, we optimized the feature extractor to jointly separate classes (with the ArcFace loss) and fool the orientation predictor (OP). The feature extractor (FE) has thus two terms in its loss function (+ regularization): $L_{FE} = L_{\text{ArcFace}} - \alpha L_{OP}$, where α is a weighting factor to balance the two objectives. This way, the network would produce discriminative features without including the irrelevant orientation information, using adversarial

learning similarly to GANs.

Table D.1: Results on ALEGORIA with adversarial feature learning trained on SF300

		Absolute perf. (mAP)	Intra-domain performance (mAP)					Inter-domain performance		
Method	Training dataset	ALEGORIA	MRU	Lapie	Phototheque	Internet	Henrard	mP1	qP1	mAPD
GeM - ArcFace	SF300	14.41	26.27	17.76	14.44	9.33	13.32	99	20	256.5
AdvGeM - ArcFace	SF300	13.99	25.78	18.09	13.66	9.28	13.04	114	21	225.7

Table D.1 shows the performance of the features learned with the adversarial mechanism, coined AdvGeM, trained on SF300 and tested with absolute and intra-domain performance on ALEGORIA. We picked the best performing model among multiple runs, with varying values of α and other hyperparameters. The results are very similar, with a slight decrease in absolute performance and a minor improvement of the mAPD score for the adversarial version. These observations indicate that the features might be more robust, but without maintaining enough disciminative power to give a gain in absolute performance. There are two plausible, non exclusive explanations:

1. the range of vertical variation and diversity in SF300 is not sufficient to learn robustness, an argument supported by the fact that performance saturates when training a global descriptor on SF300.

2. enforcing robustness with an adversarial mechanism is counter-productive because we constrain the model to "not do something" instead of constraining it to "do something", here learn invariant descriptors. An underlying hypothesis of the approach is that global descriptors would include more relevant information if we remove the irrelevant orientation information, but nothing guarantees that the model will not simply produce noisy descriptors instead. Moreover, the adversarial mechanism implies a weighting factor α diminishing the influence of the disciminative loss function (ArcFace), which leads to less accurate descripors. It is not clear whether the gained robustness can outbalance the loss of accuracy.

Considering the uncertainty in the validity of starting hypotheses for this project, and the preference for precision (finding positive images reliably) rather than recall (finding all positive images) in the application-oriented setup of ALEGORIA, I did not continue with this project.

Bibliography

[1] 2021 Worldwide Image Capture Forecast: 2020 – 2025. URL https://riseaboveresearch.com/rar-reports/2021-worldwide-image-capture-forecast-2020-2025/.

[2] Nouman Ali, Bushra Zafar, Faisal Riaz, Saadat Hanif Dar, Naeem Iqbal Ratyal, Khalid Bashir Bajwa, Muhammad Kashif Iqbal, and Muhammad Sajid. A Hybrid Geometric Spatial Image Representation for scene classification. *PLoS ONE*, 13(9), September 2018. ISSN 1932-6203. doi:10.1371/journal.pone.0203339. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6135402/.

[3] Nassim Ammour, Laila Bashmal, Yakoub Bazi, M. M. Al Rahhal, and Mansour Zuair. Asymmetric Adaptation of Deep Features for Cross-Domain Classification in Remote Sensing Imagery. *IEEE Geoscience and Remote Sensing Letters*, 15(4):597–601, April 2018. ISSN 1558-0571. doi:10.1109/LGRS.2018.2800642. Conference Name: IEEE Geoscience and Remote Sensing Letters.

[4] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, XX, 2017. URL https://hal.inria.fr/hal-01557234.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, August 2017. PMLR. URL http://proceedings.mlr.press/v70/arjovsky17a.html.

[6] Artem Babenko and Victor S. Lempitsky. Aggregating Local Deep Features for Image Retrieval. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1269–1277, 2015. doi:10.1109/ICCV.2015.150.

[7] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural Codes for Image Retrieval. In *LNCS*, volume 8689, 2014. doi:10.1007/978.3.319.10590.1_38.

[8] Hernan Badino, Daniel Huber, and Takeo Kanade. *The CMU Visual Localization Data Set*. 2011. URL `http://3dvis.ri.cmu.edu/data-sets/localization`.

[9] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved Few-Shot Visual Classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14481–14490, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi:10.1109/CVPR42600.2020.01450. URL `https://ieeexplore.ieee.org/document/9157183/`.

[10] N. Bhowmik, Li Weng, V. Gouet-Brunet, and B. Soheilian. Cross-domain image localization by adaptive feature fusion. In *2017 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4, March 2017. doi:10.1109/JURSE.2017.7924572.

[11] Alberto Bietti and Julien Mairal. Invariance and Stability of Deep Convolutional Representations. 2017. URL `https://hal.inria.fr/hal-01630265`.

[12] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval. In *European Conference on Computer Vision (ECCV), 2020.*, 2020.

[13] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying Deep Local and Global Features for Efficient Image Search. *arXiv:2001.05027 [cs]*, January 2020.

[14] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep Adaptive Image Clustering. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5880–5888, October 2017. doi:10.1109/ICCV.2017.626. ISSN: 2380-7504.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, July 2020. URL `http://proceedings.mlr.press/v119/chen20j.html`.

[16] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, 105(10):1865–1883, October 2017. ISSN 1558-2256. doi:10.1109/JPROC.2017.2675998.

[17] Gong Cheng, Ceyuan Yang, Xiwen Yao, Lei Guo, and Junwei Han. When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs. *IEEE Transactions on Geoscience and Remote Sensing*, 56(5):2811–2821, May 2018. ISSN 1558-0644. doi:10.1109/TGRS.2017.2783902.

[18] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, October 2007. doi:10.1109/ICCV.2007.4408891.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[20] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *CoRR*, abs/1801.07698, 2018. _eprint: 1801.07698.

[21] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. Publisher: IEEE.

[22] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A Baseline for Few-Shot Image Classification. September 2019. URL `https://openreview.net/forum?id=rylXBkrYDS`.

[23] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial Feature Learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=BJtNZAFgg`.

[24] Michael Donoser and Horst Bischof. Diffusion Processes for Retrieval Revisited. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1320–1327, June 2013. doi:10.1109/CVPR.2013.174. ISSN: 1063-6919.

[25] Anastasios Doulamis, Nikolaos Doulamis, Eftychios Protopapadakis, Athanasios Voulodimos, and Marinos Ioannides. 4D Modelling in Cultural Heritage. In Marinos Ioannides, João Martins, Roko Žarnić, and Veranika Lim, editors, *Advances in Digital Cultural Heritage*, Lecture Notes in Computer Science, pages 174–196, Cham, 2018. Springer International Publishing. ISBN 978-3-319-75789-6. doi:10.1007/978-3-319-75789-6_13.

[26] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive Methods for Real-World Domain Generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[27] Jamie Enoch, Leanne McDonald, Lee Jones, Pete R. Jones, and David P. Crabb. Evaluating Whether Sight Is the Most Valued Sense. *JAMA ophthalmology*, 137(11):1317–1320, November 2019. ISSN 2168-6173. doi:10.1001/jamaophthalmol.2019.3537.

[28] Lili Fan, Hongwei Zhao, and Haoyu Zhao. Distribution Consistency Loss for Large-Scale Remote Sensing Image Retrieval. *Remote Sensing*, 12(1):175, January 2020. doi:10.3390/rs12010175. URL `https://www.mdpi.com/2072-4292/12/1/175`.

[29] 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. ISSN 0001-0782. doi:10.1145/358669.358692.

[30] Dimitri Gominski, Martyna Poreba, Valérie Gouet-Brunet, and Liming Chen. Challenging deep image descriptors for retrieval in heterogeneous iconographic collections. *Proceedings of the 1st Workshop on Structuring and Understanding of Multimedia heritAge Contents - SUMAC '19*, pages 31–38, 2019. doi:10.1145/3347317.3357246.

[31] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680, Montreal, Canada, 2014. MIT Press.

[32] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein GANs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 5769–5779, Red Hook, NY, USA, December 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4.

[33] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi:10.1109/CVPR.2016.90.

[34] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of Traffic Signs in Real-World Images: The German Traffic Sign Detection Benchmark. In *International Joint Conference on Neural Networks*, 2013. Issue: 1288.

[35] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification. *ACM Transactions on Graphics (Proc. of SIGGRAPH 2016)*, 35(4):110:1–110:11, 2016.

[36] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Teddy Furon, and Ondrej Chum. Efficient Diffusion on Region Manifolds: Recovering Small Objects with Compact CNN Representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 926–935, July 2017. doi:10.1109/CVPR.2017.105. ISSN: 1063-6919.

[37] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial Transformer Networks. *Advances in Neural Information Processing Systems*, 28, 2015. URL `https://papers.nips.cc/paper/2015/hash/33ceb07bf4eeb3da587e268d663aba1a-Abstract.html`.

[38] H. Jegou, M. Douze, and C. Schmid. Product Quantization for Nearest Neighbor Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, January 2011. ISSN 0162-8828. doi:10.1109/TPAMI.2010.57.

[39] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, Lecture Notes in Computer Science, pages 304–317. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-88682-2.

[40] Herve Jegou, Matthijs Douze, Cordelia Schmid, and Patrick Perez. Aggregating local descriptors into a compact image representation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3304–3311, San Francisco, CA, USA, 2010. IEEE. ISBN 978-1-4244-6984-0. doi:10.1109/CVPR.2010.5540039. URL `http://ieeexplore.ieee.org/document/5540039/`.

[41] Yannis Kalantidis, Clayton Mellina, and Simon Osindero. Cross-Dimensional Weighting for Aggregated Deep Convolutional Features. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, Lecture Notes in Computer Science, pages 685–701. Springer International Publishing, 2016. ISBN 978-3-319-46604-0.

[42] Byungsoo Ko, Minchul Shin, Geonmo Gu, HeeJae Jun, Tae Kwan Lee, and Youngjoon Kim. A Benchmark on Tricks for Large-scale Image Retrieval. *arXiv:1907.11854 [cs]*, April 2020. URL `http://arxiv.org/abs/1907.11854`. arXiv: 1907.11854.

[43] Naveen Kodali, J. Abernethy, James Hays, and Z. Kira. How to Train Your DRAGAN. *ArXiv*, 2017.

[44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012. URL `http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf`.

[45] Michał Lagiewka, Marcin Korytkowski, and Rafal Scherer. Distributed image retrieval with colour and keypoint features. *Journal of Information and Telecommunication*, 3 (4):430–445, October 2019. ISSN 2475-1839. doi:10.1080/24751839.2019.1620023. URL `https://doi.org/10.1080/24751839.2019.1620023`. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/24751839.2019.1620023.

[46] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, December 2015. ISSN 0036-8075, 1095-9203. doi:10.1126/science.aab3050. URL `https://science.sciencemag.`

org/content/350/6266/1332. Publisher: American Association for the Advancement of Science Section: Research Article.

[47] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Deeper, Broader and Artier Domain Generalization. pages 5543–5551, October 2017. doi:10.1109/ICCV.2017.591.

[48] Haifeng Li, Zhenqi Cui, Zhiqiang Zhu, Li Chen, Jiawei Zhu, Haozhe Huang, and Chao Tao. RS-MetaNet: Deep Metametric Learning for Few-Shot Remote Sensing Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–12, 2020. doi:10.1109/TGRS.2020.3027387.

[49] Haifeng Li, Xin Dou, Chao Tao, Zhixiang Wu, Jie Chen, Jian Peng, Min Deng, and Ling Zhao. RSI-CB: A Large-Scale Remote Sensing Image Classification Benchmark Using Crowdsourced Data. *Sensors*, 20(6):1594, January 2020. doi:10.3390/s20061594. URL https://www.mdpi.com/1424-8220/20/6/1594.

[50] Lingjun Li, Junwei Han, Xiwen Yao, Gong Cheng, and Lei Guo. DLA-MatchNet for Few-Shot Remote Sensing Image Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–10, 2020. ISSN 1558-0644. doi:10.1109/TGRS.2020.3033336. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[51] Yansheng Li, Jiayi Ma, and Yongjun Zhang. Image retrieval from remote sensing big data: A survey. *Information Fusion*, 67:94–115, March 2021. ISSN 1566-2535. doi:10.1016/j.inffus.2020.10.008. URL https://www.sciencedirect.com/science/article/pii/S1566253520303778.

[52] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. doi:10.1007/978-3-319-10602-1_48.

[53] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5007–5015, June 2015. doi:10.1109/CVPR.2015.7299135. ISSN: 1063-6919.

[54] Bing Liu, Xuchu Yu, Anzhu Yu, Pengqiang Zhang, Gang Wan, and Ruirui Wang. Deep Few-Shot Learning for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(4):2290–2304, April 2019. ISSN 1558-0644. doi:10.1109/TGRS.2018.2872830. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[55] Yishu Liu, Zhengzhuo Han, Conghui Chen, Liwang Ding, and Yingbin Liu. Eagle-Eyed Multitask CNNs for Aerial Image Retrieval and Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(9):6699–6721, September 2020. ISSN 1558-0644. doi:10.1109/TGRS.2020.2979011.

[56] Yang Long, Yiping Gong, Zhifeng Xiao, and Qing Liu. Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2486–2498, May 2017. ISSN 1558-0644. doi:10.1109/TGRS.2016.2645610. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[57] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004. ISSN 0920-5691. doi:10.1023/B:VISI.0000029664.99615.94. URL `http://link.springer.com/10.1023/B:VISI.0000029664.99615.94`.

[58] Xiaoqiang Lu, Tengfei Gong, and Xiangtao Zheng. Multisource Compensation Network for Remote Sensing Cross-Domain Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(4):2504–2515, April 2020. ISSN 1558-0644. doi:10.1109/TGRS.2019.2951779. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[59] Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Alan Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152:166–177, June 2019. ISSN 0924-2716. doi:10.1016/j.isprsjprs.2019.04.015. URL `http://www.sciencedirect.com/science/article/pii/S0924271619301108`.

[60] Richard T. Marriott, Safa Madiouni, Sami Romdhani, Stéphane Gentric, and Liming Chen. An Assessment of GANs for Identity-related Applications. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, September 2020. doi:10.1109/IJCB48548.2020.9304879. ISSN: 2474-9699.

[61] J Matas, O Chum, M Urban, and T Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, September 2004. ISSN 0262-8856. doi:10.1016/j.imavis.2004.02.006. URL `http://www.sciencedirect.com/science/article/pii/S0262885604000435`.

[62] Maria-Elena Nilsback and Andrew Zisserman. Automated Flower Classification over a Large Number of Classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*, pages 722–729, December 2008. doi:10.1109/ICVGIP.2008.47.

[63] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-Scale Image Retrieval with Attentive

Deep Local Features. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3476–3485, October 2017. doi:10.1109/ICCV.2017.374.

[64] Mehdi Noroozi and Paolo Favaro. Unsupervised Learning of Visual Represention by solving Jigsaw Puzzles. In *ECCV*, 2016.

[65] Noé Pion, Martin Humenberger, Gabriela Csurka, and Yohann Cabon. Benchmarking Image Retrieval for Visual Localization. In *International Conference on 3D Vision*, 2020.

[66] Esam Othman, Yakoub Bazi, Farid Melgani, Haikel Alhichri, Naif Alajlan, and Mansour Zuair. Domain Adaptation Network for Cross-Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(8):4441–4456, August 2017. ISSN 1558-0644. doi:10.1109/TGRS.2017.2692281. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[67] Kohei Ozaki and Shuhei Yokoo. Large-scale Landmark Retrieval/Recognition under a Noisy and Diverse Dataset. *arXiv:1906.04087 [cs]*, June 2019. URL `http://arxiv.org/abs/1906.04087`. arXiv: 1906.04087.

[68] Otávio AB Penatti, Keiller Nogueira, and Jefersson A Dos Santos. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 44–51, 2015.

[69] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. February 2018. URL `https://hal.inria.fr/hal-01648685`.

[70] J. Philbin, O. Chum, M. Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[71] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. doi:10.1109/CVPR.2008.4587635.

[72] Nathan Piasco, Désiré Sidibé, Valérie Gouet-Brunet, and Cédric Demonceaux. Improving Image Description with Auxiliary Modality for Visual Localization in Challenging Conditions. *International Journal of Computer Vision*, 129(1):185–202, January 2021. ISSN 1573-1405. doi:10.1007/s11263-020-01363-6. URL `https://doi.org/10.1007/s11263-020-01363-6`.

[73] Filip Radenovic, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking. In *2018*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5706–5715, 2018. doi:10.1109/CVPR.2018.00598.

[74] F. Radenović, G. Tolias, and O. Chum. Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2019. ISSN 0162-8828. doi:10.1109/TPAMI.2018.2846566.

[75] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. pages 779–788, June 2016. doi:10.1109/CVPR.2016.91.

[76] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E Turner. Fast and Flexible Multi-Task Classification using Conditional Neural Adaptive Processes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7957–7968. Curran Associates, Inc., 2019.

[77] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5107–5116, 2019.

[78] Christoph Rieke. Awesome Satellite Imagery Datasets, August 2021. URL `https://github.com/chrieke/awesome-satellite-imagery-datasets`. original-date: 2018-05-01T22:13:17Z.

[79] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*, pages 2564–2571, November 2011. doi:10.1109/ICCV.2011.6126544.

[80] Avraham Ruderman, Neil C. Rabinowitz, Ari S. Morcos, and Daniel Zoran. Pooling is neither necessary nor sufficient for appropriate deformation stability in CNNs. *arXiv:1804.04438 [cs, stat]*, 2018. URL `http://arxiv.org/abs/1804.04438`.

[81] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic Routing Between Capsules. *arXiv:1710.09829 [cs]*, October 2017. URL `http://arxiv.org/abs/1710.09829`.

[82] Veit Sandfort, Ke Yan, Perry J. Pickhardt, and Ronald M. Summers. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific Reports*, 9(1):16884, November 2019. ISSN 2045-2322. doi:10.1038/s41598-019-52737-x. URL `https://www.nature.com/articles/s41598-019-52737-x`. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Diagnostic markers;Image processing Subject_term_id: diagnostic-markers;image-processing.

[83] 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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8601–8610, June 2018. doi:10.1109/CVPR.2018.00897. ISSN: 2575-7075.

[84] Grant Schindler and Frank Dellaert. 4D Cities: Analyzing, Visualizing, and Interacting with Historical Urban Photo Collections. *Journal of Multimedia*, 7, 2012. doi:10.4304/jmm.7.2.124-131.

[85] Grant J. Scott, Matthew R. England, William A. Starms, Richard A. Marcum, and Curt H. Davis. Training Deep Convolutional Neural Networks for Land–Cover Classification of High-Resolution Imagery. *IEEE Geoscience and Remote Sensing Letters*, 14(4):549–553, April 2017. ISSN 1558-0571. doi:10.1109/LGRS.2017.2657778. Conference Name: IEEE Geoscience and Remote Sensing Letters.

[86] Grant J. Scott, Kyle C. Hagan, Richard A. Marcum, James Alex Hurt, Derek T. Anderson, and Curt H. Davis. Enhanced Fusion of Deep Neural Networks for Classification of Benchmark High-Resolution Image Data Sets. *IEEE Geoscience and Remote Sensing Letters*, 15(9):1451–1455, September 2018. ISSN 1558-0571. doi:10.1109/LGRS.2018.2839092.

[87] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, September 2014. URL `http://arxiv.org/abs/1409.1556`.

[88] Oriane Siméoni, Yannis Avrithis, and Ondrej Chum. Local Features and Visual Words Emerge in Activations. *arXiv:1905.06358 [cs]*, 2019. URL `http://arxiv.org/abs/1905.06358`.

[89] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477, October 2003. URL `http://www.robots.ox.ac.uk/vgg`.

[90] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*, 2017.

[91] Shaoyue Song, Hongkai Yu, Zhenjiang Miao, Qiang Zhang, Yuewei Lin, and Song Wang. Domain Adaptation for Convolutional Neural Networks-Based Remote Sensing Scene Classification. *IEEE Geoscience and Remote Sensing Letters*, 16(8):1324–1328, August 2019. ISSN 1558-0571. doi:10.1109/LGRS.2019.2896411. Conference Name: IEEE Geoscience and Remote Sensing Letters.

[92] Marvin Teichmann, Araujo André, Zhu Menglong, and Sim Jack. Detect-to-Retrieve: Efficient Regional Aggregation for Image Search. July 2019. doi:10.17863/CAM.42014. URL `https://www.repository.cam.ac.uk/handle/1810/294927`. Accepted: 2019-07-25T23:30:43Z.

[93] Carl Toft, Will Maddern, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, Fredrik Kahl, and Torsten Sattler. Long-Term Visual Localization Revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. ISSN 1939-3539. doi:10.1109/TPAMI.2020.3032010. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[94] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. To aggregate or not to aggregate: selective match kernels for image search. In *ICCV - International Conference on Computer Vision*, Sydney, Australia, December 2013. URL `https://hal.inria.fr/hal-00864684`.

[95] 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(3):247–261, 2016. ISSN 0920-5691, 1573-1405. doi:10.1007/s11263-015-0810-4. URL `http://link.springer.com/10.1007/s11263-015-0810-4`.

[96] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular Object Retrieval With Integral Max-Pooling of CNN Activations. In *ICL 2016 - RInternational Conference on Learning Representations*, International Conference on Learning Representations, pages 1–12, San Juan, Puerto Rico, 2016. URL `https://hal.inria.fr/hal-01842218`.

[97] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum. Learning and Aggregating Deep Local Descriptors for Instance-Level Recognition. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 460–477, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58452-8. doi:10.1007/978-3-030-58452-8_27.

[98] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *CVPR*, 2015.

[99] Akihiko Torii, Josef Sivic, Toma Pajdla, and Masatoshi Okutomi. Visual Place Recognition with Repetitive Structures. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 883–890, Portland, OR, USA, 2013. IEEE. ISBN 978-0-7695-4989-7. doi:10.1109/CVPR.2013.119. URL `http://ieeexplore.ieee.org/document/6618963/`.

[100] Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. Few-Shot Learning through an Information Retrieval Lens. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 2252–2262, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4. event-place: Long Beach, California, USA.

[101] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Jordan Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples. In *International Conference on Learning Representations (submission)*, 2020.

[102] Y. Verdie, Kwang Moo Yi, P. Fua, and V. Lepetit. TILDE: A Temporally Invariant Learned DEtector. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5279–5288, 2015. doi:10.1109/CVPR.2015.7299165.

[103] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3630–3638. Curran Associates, Inc., 2016.

[104] Qi Wang, Shaoteng Liu, Jocelyn Chanussot, and Xuelong Li. Scene Classification With Recurrent Attention of VHR Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):1155–1167, February 2019. ISSN 1558-0644. doi:10.1109/TGRS.2018.2864987.

[105] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys*, 53(3):63:1–63:34, June 2020. ISSN 0360-0300. doi:10.1145/3386252. URL `https://doi.org/10.1145/3386252`.

[106] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google Landmarks Dataset v2 – A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. *arXiv:2004.01804 [cs]*, 2020. URL `http://arxiv.org/abs/2004.01804`.

[107] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-Area Image Geolocalization with Aerial Reference Imagery. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–9, 2015. doi:10.1109/ICCV.2015.451.

[108] Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and G. Brostow. Harmonic Networks: Deep Translation and Rotation Equivariance. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7168–7177, 2017.

[109] Gui-Song Xia, Wen Yang, Julie Delon, Yann Gousseau, Hong Sun, and Henri Maître. Structural High-resolution Satellite Image Indexing. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 38, 2010.

[110] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, July 2017. ISSN 1558-0644. doi:10.1109/TGRS.2017.2685945.

[111] Lingxi Xie, Richang Hong, Bo Zhang, and Qi Tian. Image Classification and Retrieval are ONE. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, pages 3–10, New York, NY, USA, June 2015. Association for Computing Machinery. ISBN 978-1-4503-3274-3. doi:10.1145/2671188.2749289. URL `https://doi.org/10.1145/2671188.2749289`.

[112] Suhui Xu, Xiaodong Mu, Dong Chai, and Xiongmei Zhang. Remote sensing image scene classification based on generative adversarial networks. *Remote Sensing Letters*, 9(7):617–626, July 2018. ISSN 2150-704X. doi:10.1080/2150704X.2018.1453173. URL `https://doi.org/10.1080/2150704X.2018.1453173`. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/2150704X.2018.1453173.

[113] Fan Yang, Ryota Hinami, Yusuke Matsui, Steven Ly, and Shin'ichi Satoh. Efficient Image Retrieval via Decoupling Diffusion into Online and Offline Processing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9087–9094, July 2019. doi:10.1609/aaai.v33i01.33019087.

[114] Yi Yang and Shawn Newsam. Bag-of-visual-words and Spatial Extensions for Land-use Classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 270–279. ACM, 2010. ISBN 978-1-4503-0428-3. doi:10.1145/1869790.1869829. URL `http://doi.acm.org/10.1145/1869790.1869829`.

[115] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: Learned Invariant Feature Transform. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 467–483. Springer International Publishing, 2016. ISBN 978-3-319-46466-4.

[116] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting Ground-Level Scene Layout from Aerial Imagery. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4132–4140, July 2017. doi:10.1109/CVPR.2017.440. ISSN: 1063-6919.

[117] Pei Zhang, Yunpeng Bai, Dong Wang, Bendu Bai, and Ying Li. Few-Shot Classification of Aerial Scene Images via Meta-Learning. *Remote Sensing*, 13(1):108, January 2021. doi:10.3390/rs13010108. URL `https://www.mdpi.com/2072-4292/13/1/108`. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[118] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful Image Colorization. *arXiv:1603.08511 [cs]*, March 2016. URL `http://arxiv.org/abs/1603.08511`.

[119] Xuanmeng Zhang, Minyue Jiang, Zhedong Zheng, Xiao Tan, Errui Ding, and Yi Yang. Understanding Image Retrieval Re-Ranking: A Graph Neural Network Perspective. *arXiv preprint arXiv:2012.07620*, 2020.

[120] Bei Zhao, Yanfei Zhong, Gui-Song Xia, and Liangpei Zhang. Dirichlet-Derived Multiple Topic Scene Classification Model for High Spatial Resolution Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 54(4):2108–2123, April 2016. ISSN 1558-0644. doi:10.1109/TGRS.2015.2496185.

[121] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A Multi-view Multi-source Benchmark for Drone-based Geo-localization. *ACM Multimedia*, 2020.

[122] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking Person Re-identification with k-Reciprocal Encoding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3652–3661, July 2017. doi:10.1109/CVPR.2017.389. ISSN: 1063-6919.

[123] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, June 2018. ISSN 1939-3539. doi:10.1109/TPAMI.2017.2723009. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[124] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. PatternNet: A Benchmark Dataset for Performance Evaluation of Remote Sensing Image Retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018. ISSN 09242716. doi:10.1016/j.isprsjprs.2018.01.004. URL `http://arxiv.org/abs/1706.03424`.

[125] Qiqi Zhu, Yanfei Zhong, Liangpei Zhang, and Deren Li. Scene Classification Based on the Fully Sparse Semantic Topic Model. *IEEE Transactions on Geoscience and Remote Sensing*, 55(10): 5525–5538, October 2017. ISSN 1558-0644. doi:10.1109/TGRS.2017.2709802.

[126] Qin Zou, Lihao Ni, Tong Zhang, and Qian Wang. Deep Learning Based Feature Selection for Remote Sensing Scene Classification. *IEEE Geoscience and Remote Sensing Letters*, 12(11): 2321–2325, November 2015. ISSN 1545-598X, 1558-0571. doi:10.1109/LGRS.2015.2475299. URL `http://ieeexplore.ieee.org/document/7272047/`.