How to Spatialize Geographical Iconographic Heritage

Emile Blettery, Nelson Fernandes, Valérie Gouet-Brunet

How to Spatialize Geographical Iconographic Heritage

Emile Blettery
LASTIG, Univ. Gustave Eiffel, ENSG, IGN, 94160 Saint-Mande, France
Ville de Paris, DAC, DHAAP
emile.blettery@ign.fr

Nelson Fernandes
LASTIG, Univ. Gustave Eiffel, ENSG, IGN, 94160 Saint-Mande, France
nelson.fernandes@ign.fr

Valérie Gouet-Brunet
LASTIG, Univ. Gustave Eiffel, ENSG, IGN, 94160 Saint-Mande, France
valerie.gouet@ign.fr

Figure 1: Examples of geographical iconographic heritage contents (images from IGN, French national Archives and Internet)[1]

ABSTRACT

This article is dedicated to the spatialization of image contents, with a focus on geographical iconographic heritage, *i.e.* digitized or born-digital image collections, acquired at variable temporal periods and showing the territory and its human-made and natural visual landmarks. We present a panorama of the current solutions (manual, semi-automatic and fully automatic alternatives) that exist to spatialize a visual content, with respect to the data available and the level of spatialization targeted. In particular, we highlight the characteristics of the approaches dedicated to geographical iconographic heritage, and in some cases, we present tests and practical feedbacks that we had the opportunity to conduct for old photographic contents in oblique aerial and terrestrial imagery.

SUMAC '21, October 20, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8668-5/21/10…$15.00
https://doi.org/10.1145/3475720.3484444

CCS CONCEPTS

• **Information systems** → **Image search**; • **Computing methodologies** → *Feature selection*; • **Applied computing** → **Arts and humanities**.

KEYWORDS

Spatialization, Localization, CBIR, Landmark retrieval, Pose estimation, Iconographic heritage

ACM Reference Format:
Emile Blettery, Nelson Fernandes, and Valérie Gouet-Brunet. 2021. How to Spatialize Geographical Iconographic Heritage. In *Proceedings of the 3rd Workshop on Structuring and Understanding of Multimedia heritAge Contents (SUMAC '21), October 20, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3475720.3484444

1 INTRODUCTION

This article is interested in the geographical iconographic heritage, *i.e.* digitized or born-digital image collections, acquired at variable temporal periods and showing the territory and its human-made and natural visual landmarks. This category of contents is extremely widespread because accounting for a large portion of what people

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

have liked to capture through drawings or amateur or professional photographs (see Figure 1), leading to a huge visual testimony of our environment that can benefit use cases and applications, ranging from historical and sociological studies up to mobile mapping scenarios, through digital tourism, landscape ecology or remote sensing. The visual representations associated with these objects of interest are extremely diverse given the various acquisition conditions (different sources, dates, viewpoints) and the evolution of landmarks over time, making their analysis still challenging today; Figure 2 illustrates the variety of such representations.

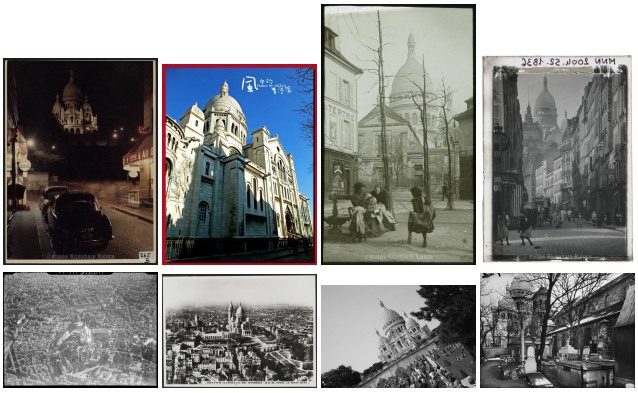

Figure 2: Sacré-Cœur Basilica, at different times, from different perspectives using various media[2]

By nature, these contents are linked to a spatial information, which can be known according to the associated acquisition system (*e.g.* GPS position capture) or post-processing (*e.g.* manual annotation), or on the contrary unknown or not precise enough for the targeted application. The older the contents, the more this spatial information may be poorly documented, or insufficiently documented for today's applications that aim at their valorization through 4D modelling applications for example. The growing interest for the valorization of heritage data and the availability of efficient digitization systems also give rise to the emergence of increasingly large volumes of digitized contents, where spatialization is an attribute that can bring structure.

The notion of "spatialization" of images can cover a wide spectrum of definitions and levels of precision - which we redefine and attempt to address in this article - from attaching to the image a simple name up to a precise position and orientation in the 3D scene. The approaches are also very numerous, as well as the data available on which it is possible to rely on. The objective of this article is to put on the table a panorama of the current solutions (manual, semi-automatic and fully automatic) that exist to spatialize a visual content, with respect to the data available and the level of spatialization targeted. When possible, we will highlight the characteristics of the approaches dedicated to geographical iconographic heritage. In some cases, we will be able to present tests and practical feedbacks that we had the opportunity to conduct for old photographic contents in oblique aerial and terrestrial imagery.

The paper is organized as follows: first, Section 2 introduces the main notions: the levels of spatialization that are usually expected, a big picture of the different methodologies of the literature and the data associated. Then Sections 3, 4 and 5 focus on the existing trends and approaches, while Section 6 concludes by providing a general synthesis and discussion on the solutions for image spatialization.

2 OVERVIEW ON METHODS AND DATA

This section is a preamble that has the objectives of clarifying what kind of spatial information can be associated with an image (Section 2.1), of proposing a classification of the existing spatialization solutions, which are numerous and multidisciplinary (Section 2.2), and finally of revisiting the data sources available to assist the spatialization process (Section 2.3).

2.1 Levels of spatialization

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

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

In the rest of the article, we will use the term "localization" to refer to one of these types of spatialization output.

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

2.2 Spatialization techniques at a glance

Different research communities, including Social Sciences and Humanities, Digital Humanities, Computer Vision, Photogrammetry, Content-Based Image Retrieval, Machine Learning and Human-Computer Interaction, address the image spatialization task. Depending on the application, the objective targeted and the data available, the spectrum of existing approaches is thus very large. To classify them, a key criterion that can be tackled is the size of the search area which has to be visited in order to determine a localization, that drastically drives the methodology to adopt:

- When the area of search is undetermined or large (in terms of spatial footprint or size of the reference data), a majority of the solutions rely on **Content-Based Image Retrieval** (CBIR) where the objective is to circumscribe the search area by efficiently determining a small set of similar images in a reference dataset of images potentially large. Depending on the targeted spatialization and research communities, nowadays there exist several variants of this problem focusing on landmark datasets, namely **place recognition** or **landmark retrieval** which perform fine-grained instance image retrieval with learning on training landmark datasets. We revisit these approaches funded on the description of the image content in Section 3.
- If an initial localization is already known (from CBIR, sensors such as GPS, metadata, experts or context, etc.) and requires to be refined, traditional solutions rely on geometrical tools from Computer Vision and Photogrammetry, with the **estimation of the pose of the camera** (*i.e.* its position and orientation) from 2D and 3D data. The most recent solutions, based on dedicated training datasets, also exploit machine learning relying on the direct **pose regression** from the data. These approaches are revisited in Section 4. Note that they are compatible with the CBIR ones which bring an initial localization, and both can be integrated in an end-to-end framework of precise localization at large scale.

The classes of techniques listed above belong to ICT and AI communities, they are computational with the capability of dealing with large volumes of data and of determining localization automatically. However, in addition to the question of their robustness when considering geographical iconographic contents, they require a human qualification of the output and remain manipulable by experts of the field, in such a way that in practice these quantitative approaches have not yet supplanted the more qualitative ones generally found in the Social Sciences and Humanities field or for general public applications, which rely on manual or sometimes semi-automatic spatialization, with the natural disadvantage of not being scalable; we revisit them in Section 5.

2.3 Available data as spatialized reference

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

2.3.1 Spatialized image datasets. In the Computer Vision and Machine Learning communities, there exist several annotated datasets dedicated to landmarks, which can apply for spatialization; let's mention Google Landmarks [47], which is one of the best known (version GLDv2 is the largest with over 5M images and 200k distinct spatial instance labels). It is relevant as training and test dataset for the CBIR and place recognition tasks mentioned in Section 2.2, assuming in practice that the localization to determine is indexed there.

For more specific or dedicated purposes and contents (neither mapped nor perennial landmarks), annotated training datasets and benchmarks are usually not available, but the CBIR task is able to address spatialized image collections that may exist in GLAM (Galleries, Libraries, Archives and Museums) which cover various iconographic contents, or in public and private mapping agencies which image territories at large scale. Here, the metadata associated with iconographic heritage are very heterogeneous, depending on the objectives and standards of the holding organizations ; for instance semantic descriptions of the content for preservation institutions and multimodal geographic descriptions for mapping agencies.

Note that maps are by definition a rich source of referencing, with an unequaled spatial (and sometimes temporal) coverage, but if there exist some automatic solutions to align a vertical airborne view with a map [19, 23], usually airborne imagery is more exploited as reference; and additionally it is more difficult to establish a link between a map and a free viewpoint image. One alternative is to rely on semantic landmark extraction (through pattern detection tasks) and to search on maps by exploiting spatial reasoning, such as in [46] where the semantic objects seen in the street-view image serve as anchors for spatialization within OpenStreetMap.

2.3.2 3D models. To gain in robustness and precision when the objective is to estimate a 3D position or a 6-DoF pose (see Section 2.2), the most recent and efficient spatialization approaches exploit all the geometrical 3D information available. Many approaches exploit 3D point clouds obtained with Structure-from-Motion (SfM) techniques, as in [50], [31] and [34], built especially on a given area for a spatialization in this area. There exist other alternatives, such as simple or sophisticated 3D building models, as well as LiDAR or RGB-D data belonging from recent scanning systems that respectively provide a 3D sparse geometrical information and a 3D depth.

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

Interestingly, thanks to public or private mapping agencies, some 3D models exist at large scale and are then usable on a much less

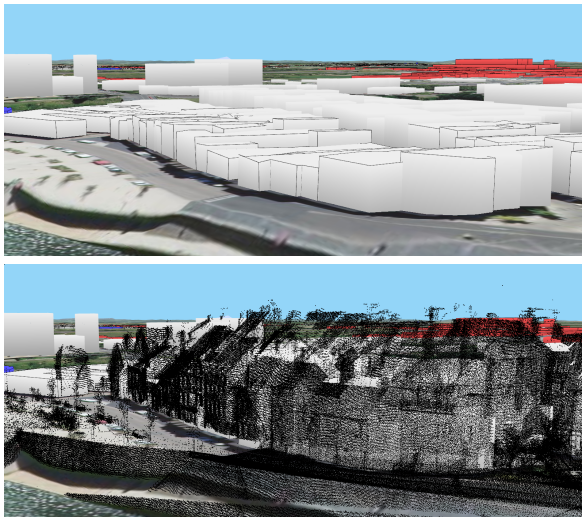

Figure 3: Examples of scalable 3D models (data from IGN). 1st row: CityGML LoD1 buildings (French "Ref3DNat" reference), available on the whole territory ; 2nd row: Superposition with terrestrial LiDAR points cloud acquired on demand at the scale of the city.

narrowed footprint than dedicated SfM clouds or public benchmarks, such as those displayed in Figure 3. Currently, their spatial coverage tends to be inversely proportional to their precision (in terms of levels of detail and localization), but this ratio is reducing with the implementation of massive and sophisticated acquisition protocols (*e.g.* aerial HD LiDAR will be available on the whole French territory in 2025).

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

To conclude this section, several public image datasets dedicated to landmarks are presented in Table 1, with a focus on the spatial coverage addressed, the type of localization data available as well as the time period covered to have an insight on their match with old contents. Whether the objective is to employ them to learn a description or as reference to spatialize a content, these datasets are numerous, but most of them are not dedicated to geographical iconographic heritage. They do not reflect correctly the heterogeneity representative of heritage contents as experimented with the Alegoria dataset [15], in which are highlighted the difficulties encountered by state-of-the-art deep features in the context of cultural heritage content retrieval.

3 SPATIALIZATION BASED ON IMAGE CONTENT DESCRIPTION

The first family of approaches, introduced in Section 2.2, relies on Content-Based Image Retrieval, which enables the possibility of retrieving images similar to a query one in a dataset, according to

Table 1: Overview on public image datasets dedicated to landmarks, exploited for training purposes or as spatialization reference (MMS stands for Mobile Mapping System).

Dataset	Number of images	Viewpoint and spatial coverage	Localization type	Time gap
Large Time Lags Locations [10]	500	Street-level 25 cities of Europe and Asia	Label	150 years
Google Landmarks Dataset v2 [47]	Over 5M	Street-level and aerial 246 countries	Label	Unspecified
$\mathcal{R}$Oxford [32]	Over 5k	Mostly street-level and some aerial Oxford	Label	Unspecified
Aachen Day-Night [35]	7712	Street-Level city of Aachen (Germany)	Label, GPS, 3D	2 years
Extended CMU-Seasons [35]	Over 110k	Street-level MMS camera areas of Pittsburgh (USA)	Label, GPS, 3D	1 year
RobotCar Seasons [35] [25]	Over 35k	Street-level MMS camera city of Oxford (UK)	Label, GPS, 3D	1 year
Kitti Vision Bench. [14]	389	Street-level MMS camera Greater Karlsruhe (city, rural areas and highways)	Label, GPS, 3D	2012
SILDa Weather and Time of Day [1]	Over 14k	Street-level and aerial London	Label	1 year
HistAerial [33]	4.9M	Vertical aerial France (sparse)	GPS	1970-1990
Alegoria [15]	13175	Street-level and aerial France (sparse)	Label	1920's-today

criteria based on the description of the image contents. When considering visual landmarks datasets, The use of descriptors adapted to fine-grained instance retrieval allows to retrieve images of the same landmark; Figure 4 illustrates this idea on aerial iconographic heritage, with here the objective of retrieving images of the same scene for spatialization with annotation. By querying a reference dataset of spatialized images, it is then possible to establish a link between the query and content-based similar spatialized contents that can be exploited to determine its localization. The popularity of such family of approaches for spatialization relies on the fact that 1/ CBIR manages image datasets at large scale, making the search area potentially very large depending on the spatial coverage of the reference, and 2/ the process being entirely unsupervised, it makes it possible to search for a landmark in a large variety of landmarks, expressed in the query (and not learned with a classification task) on public or dedicated datasets [22].

3.1 CBIR techniques

CBIR approaches share common representations based on compact image-level descriptors, especially with the help of machine learning since deep learning has given birth to powerful image descriptors. Surveys on image descriptors have been numerous, we refer the reader to [8, 13, 24, 29, 51] for hand-crafted features and more recent deep learning-based methods, as well as other modalities to help in the description of the content. [27] describes in detail CBIR approaches developed over the years, from the evolution of image representations to the specific applications (robotics or aerial imagery) via multiple techniques to improve the descriptors' robustness and discriminative power. It also points out the challenges that appearance changes raise for CBIR (*i.e.*, when the dataset considered combines multiple variations like domain, color, viewpoint, illumination, etc.). To alleviate those issues, several research avenues should be pursued. First, as in [15], focusing on image descriptors robust to multiple appearance changes. Second,

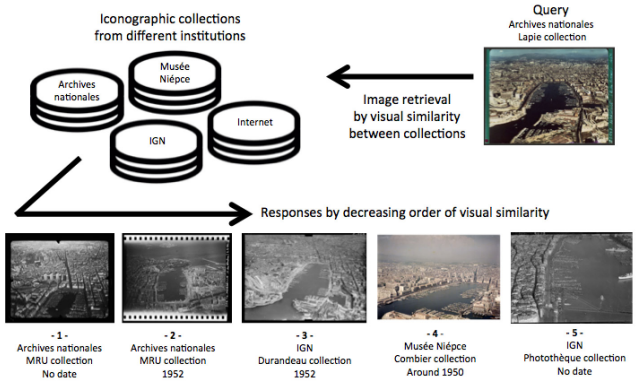

Figure 4: Example of retrieval by content, in a specialized dataset of 40 576 images, from a query showing "Le vieux port" in Marseille city (France). The images are described with the local image descriptor HOW [15, 42] and the 5 most similar responses are provided here, all showing the same area at different dates and viewpoints. The image query belongs to the collection "Lapie" from the French national Archives, while the dataset queried is distributed over different sources (French national Archives, Niépce Museum, IGN, internet) and collections in these sources (Lapie, MRU, Durandeau, Photothèque, Combier).

[30] shows that using scene geometry information during training improves the localisation accuracy in the end, which could open opportunities for CBIR focusing on heritage content. Third, [10] explores domain adaptation for CBIR dedicated to heritage content, meaning using prior learning based on common datasets and transfering it to another domain of representation via statistical representations; this strategy also exists for deep learning-based methods as described in [43]. Finally, as observed in Table 1, heritage iconography datasets are obviously under-represented, and most of the time with few representations of the same landmark. To alleviate this issue, a recent branch of research, namely few-shot learning, is investigating how to learn using a limited number of annotated data, or more broadly how to learn more efficiently (learn to learn); [44] gives a general overview of the relevant methods. It has been experimented for classical image retrieval in [45] and could be considered for heritage iconography, which indicates that there is still room for improvement.

On the specificity of landmark contents. CBIR relies on image description, but depending on the final exploitation of the retrieved images, it can be implemented differently, as it is stated in study [31]: for instance, do we want to retrieve all images showing the same landmark, as in Figure 4? Or do we want to find images in a footprint close to the one of the query, in order to estimate precisely its localization later? First, the landmark retrieval task aims at retrieving all images depicting the same landmark; it supposes for the descriptor to be very robust to viewpoint changes and occlusion such as proposed by [15]. Second, visual localization has the goal of estimating a precise camera pose, it requires similar images and limited viewpoint or illumination variations to be efficient. Hence, [31, 35] evaluate CBIR descriptors while focusing on the resulting 6-Dof pose as an evaluation metric. Finally, place recognition aims at obtaining a coarse camera position (for instance when the dataset by nature will not be suitable for visual localization), hence visual similarity needs to be balanced with approximate camera position to make sure that the retrieved camera pose are similar to the query one (for instance depicting the same side of a building); [51] gives a panorama of descriptors dedicated to place recognition.

In the following section, we go deeper in the spatialization problem by explaining how the links built with CBIR can be exploited to determine a localization.

3.2 From CBIR to spatialization

The links thus created between query and similar images can then be exploited in different ways in order to exhibit a localization, according to the data available and the objectives. We present main trends in Sections 3.2.1 and 3.2.2.

3.2.1 Metadata propagation. Propagating localization (and more generally descriptive metadata) through the built links can be seen as a label propagation problem in a network, where there exist many simple solutions, or more sophisticated ones which take the quality of the links and data into account [2].

Another alternative takes inspiration from the web of linked data [3], where information is not propagated, but data is linked in a unique and sustainable way: to each item in the collection, a unique URI (Uniform Resource Identifier) is assigned, and associated with a label and a date along the RDF (Resource Description Framework) data model. The triples formed can be extended with the Web Ontology Language (OWL), and for instance linked with property "owl:sameAs", making thus easily reachable all metadata associated with different versions of the same data (here, different iconographic representations of the same landmark), without presupposing the quality of the metadata.

3.2.2 Refinement of the localization. The small set of similar images returned by the content-based retrieval system drastically circumscribes the search area and therefore the localization prior. Exploiting the similar images' localization information allows for a first estimation or a refinement of the query's localization. Depending on the type of application, the image retrieval strategy is used either for landmark retrieval, place recognition or visual localization. The localization can be first estimated or refined using a simple position interpolation or a pose approximation [31]. However when the retrieved images are adapted for visual localization it is also possible to go deeper by exploiting geometrical approaches of pose estimation, which rely on an initial localization, such as those of Section 4. This is how many current end-to-end approaches of spatialization in the literature integrate such a retrieval step upstream, thus allowing the coverage of a large search area while targeting a precise pose estimation [30, 31, 34, 41].

4 SPATIALIZATION BASED ON AUTOMATIC POSE ESTIMATION

Once similar images have been retrieved, if their 3D pose is known, multiple methods for estimating the query image's pose are available. These methods also heavily rely on local features detected in images, and there exist two main trends of methods for estimating

the 6-DoF pose of an image once other similar images with a known pose are retrieved [31].

The first category of methods only uses similar images and their poses. A simple interpolation of the pose using similar images' poses can be a solution. This interpolation can be weighted depending on image similarity (using the image retrieval results for instance). Similarly, [41] estimates possible candidate poses by estimating the relative pose between the query image and each similar image and then fuse those candidate poses to find the best solution, (Figure 5).

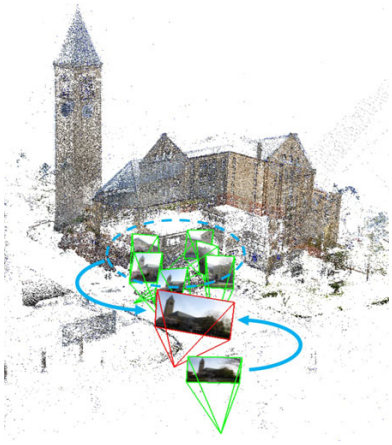

Figure 5: Pose estimation by fusing relative poses as proposed by [41] (figure from [41])

As a recent promising trend, CNN-based methods, trained on images and associated poses, directly regress the 6-DoF pose of the query image: [17] first but many after like [6] regresses the 6-DoF pose using a set of image and pose pairs. [9] and [20] train for the image retrieval step and the 6-DoF pose estimation simultaneously. From another angle, RANSAC-Flow [40] is a method that aligns two images depicting a similar view. The relative pose between points of view may eventually allow to find the absolute pose.

A second category of methods rely on 3D information, traditionally to regress the image pose solving a PnP (Perspective-n-Point) problem. This strategy implies matching 2D and 3D features between the query image and the 3D map. Those solutions often use a 3D SfM (Structure-from-Motion) model that can be constructed globally beforehand (offline) or locally on-the-fly (online) using the similar images retrieved but other input data like LiDAR or depth maps can be exploited. Today, SfM reconstruction can be performed by various softwares or libraries, for instance COLMAP [38, 39], OpenMVG [28] or VisualSfM [48, 49].

On one hand, classical methods based on a geometric (or algebraic) approach of the problem can compute the pose once the 2D-3D matches are obtained. A major difference between two sets of approaches is the previous knowledge (or not) of the camera's calibration (its intrinsic and extrinsic parameters). Hence, with a calibrated camera and 2D-3D matches (at least 3), several solvers of the PnP problem have been developed such as an efficient P3P [18] or a method accepting more than 3 matches such as EPnP [21] or PPnP [12]. An intermediary solution when the focal length of the camera is unknown is the P3Pf solver [36]. When the camera's intrinsic parameters are unknown, using the Direct Linear Transformation (DLT) [16] allows using a minimum of six matches to calibrate the camera, *i.e.* estimating its 6-DoF pose plus its intrinsic parameters (focal length, principal point, skew); this method is known as P6P. These methods also benefit greatly from the addition of a RANSAC loop [11] to select the strongest 2D-3D matches and limit the matching errors which is a common issue when dealing with very diverse content like heritage iconographic collections.

On the other hand, various CNN-based architectures have been devised to solve this problem. Some of them regress the pose by using RGB-D images taken by calibrated cameras [7]. [50] exploit images and the corresponding 3D point cloud to estimate the camera pose, creating and exploiting a "scene pyramid" containing both features and corresponding 3D coordinates, allowing for their method to be scene agnostic. Recently, [34] use a 3D model and images jointly in their network to determine the pose of the image (with a known camera calibration), camera localisation being considered as metric learning. This method appears promising as it is trained to focus on parts of the images that are important for long-term localisation. Indeed, the main issue when dealing with iconographic heritage is the matching process between keypoints as the scene may have changed, the image may be very different in color, illumination, etc. Hence devising a method able to focus on perennial objects allows for a greater certainty of the matches. Adapting this method for iconographic heritage contents could be an interesting research avenue, the main purpose being to ensure that matches are looked for in most perennial parts of the scene.

However, most methods require the knowledge of the intrinsic parameters of the camera and it may be a problem when considering iconographic heritage where the characteristics of the camera at the origin of the image may be unknown or nonexistent (*e.g.* a painting). This information is however sensitive in regard to positioning the image precisely in a 3D scene, as illustrated in Figure 6.

Evaluating the accuracy of the pose estimation methods supposes exploiting benchmark datasets where camera parameters and poses are known. Multiple datasets have been devised for this purpose and some can be found in Table 1 such as Aachen Day-Night, Extended CMU-Seasons or RobotCar Seasons. Several metrics can be used to assess a methods accuracy. From a single image perspective, position and orientation errors relative to the ground truth can be computed, as well as the mean reprojection error when matches between the image and a 3D model are available. Globally on a whole dataset, the efficiency of the method can be evaluated by applying thresholds to the position and orientation errors and thus determine how many images are correctly located. Those evaluation methods however are used on modern images for which a ground truth exist and thus hard to transfer onto heritage images.

5 MANUAL AND SEMI-AUTOMATIC SPATIALIZATION

The spatialization tools previously described represent a powerful avenue for spatialization at large scale, allowing to manage large datasets in a quantitative way. Historically, manual approaches of spatialization are naturally predominant. We revisit them briefly in Section 5.1, before considering semi-automatic solutions in Section 5.2 which can assist manual tools as well as full automatic ones.

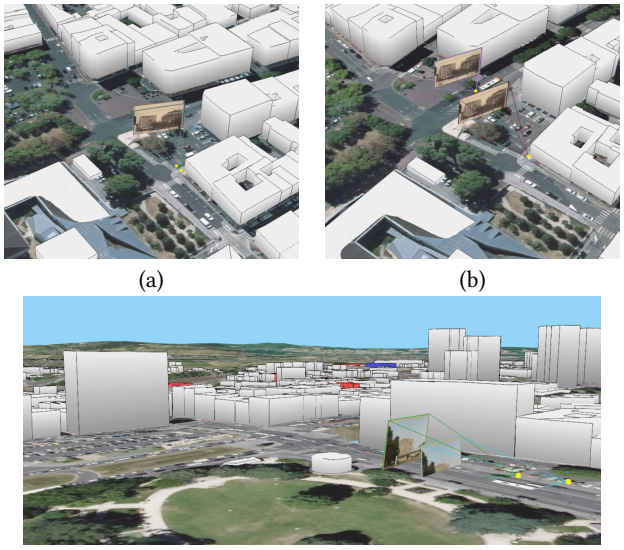

Figure 6: Illustration of the influence of the intrinsic parameters on the pose estimation : (a) Calibration estimated with DLT vs. 6-DoF pose estimated with PPnP and intrinsic parameters correctly chosen empirically (the two localizations are similar and correct) ; (b) Same estimations with different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

5.1 Manual spatialization

Manual image spatialization is mostly done through interactive web interfaces, which are certainly the oldest and most widespread tools since the development of digital humanities, as they are the most inclusive regardless of user profile and represent an excellent medium for web valorization. The last decade has given birth to the crowdsourcing paradigm via collaborative platforms and some of them exploit historical data and can even be used by researchers, to create datasets suited to their needs. Some platforms have no moderation or expert control and a limited visualisation capability (HistoryPin[3], Navilium[4]). Other solutions are dedicated to specialist users (like researchers) working on a common platform and a common dataset, but with a certainty in regard to the results displayed (UrbanHistory4D[5] [26] or Aioli[6]).

5.2 Semi-automatic 6-DoF pose estimation

Today's full automatic 6-DoF pose estimation methods have made tremendous progress with the exploitation of deep learning and various modalities, but their maturity and scalability facing practical usage remain still questionable. To improve the camera pose precision from an automatic estimation, as well as to validate it, using a semi-automatic method is still a relevant alternative, in particular during the step of local features (points) detection and matching in 2D and 3D, which is particularly sensitive to the content. It is especially the case when considering iconographic heritage where the variability of the content can be high; in particular, the manual determination of the points may ensure the selection of perennial points consistent between the old image data and the more recent 3D model, as illustrated in Figure 7. The process is then divided between the manual selection and matching of 2D and 3D features, followed by the automatic estimation of the pose. This is for example the solution chosen by Smapshot, the Swiss web-based participatory platform [4] and iTowns, the French 3D geoportal [5].

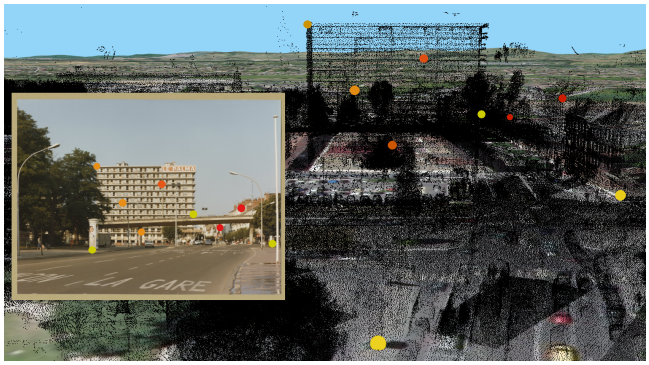

Figure 7: Interactive selection of 2D-3D pairs of points (colored bullets) in the photograph and in the 3D scene modeled with LiDAR points, as input of a 6-DoF pose estimation tool (iTowns web application [5]). In this example, we observe differences between the old photograph and the recent version of the scene (disappearance of the bridge, new buildings, roadway modification), which highlights the challenge of the points selection for the pose estimation (images from Musée Nicéphore Niepce and IGN).

By relying on a HCI system, semi-automatic methods also offer the possibility to immediately visualise the result of the pose estimation, with the possibility to evaluate it visually in its virtual environment, and potentially to refine or adjust it in a interactive loop process. When trying to develop immersive visualisation platforms, it is essential for the coherence between the estimated pose and the supposed pose to be preserved, otherwise leading to a much less pleasant experience in terms of immersiveness and interactions.

6 SYNTHESIS AND CONCLUSIONS

In this article, we have reviewed all the methods and data that seem to allow to associate a spatial information to an image, by studying more specifically the characteristics of geographical iconographic

[3] https://www.historypin.org/
[4] https://www.navilium.com/
[5] http://4dbrowser.urbanhistory4d.org
[6] http://www.aioli.cloud/

[7] https://arpenteur.bnf.fr/
[8] https://developers.google.com/maps/documentation/geocoding/overview
[9] https://nominatim.org/
[10] https://geoservices.ign.fr/documentation/services/api-et-services-ogc/geocodage-ogc

Table 2: Comparison of spatialization methods.

Type of methods	Category of method	Output	Data required	Specificities	Example
Manual	Position estimation	2D position	Images + map	The position is estimated by the user, without real visual verification	Arpenteur[7]
	Image overlaying	Relative pseudo 6-DoF pose	Images + Google StreetView	The image is overlayed, the pose is relative to a specific StreetView camera pose	HistoryPin[2]
	Pose estimation with model creation	6-DoF pose in a relative 3D model	Images	The 3D model is created as the images' poses are estimated. Exploits heritage content.	[37]
Semi-Automatic	Spatial resection within immersive platforms	6-DoF pose	Images + virtual 3D model	Time consuming	[5], [4], [26]
Automatic	Geocoding	2D position	Metadata	Require clean metadata, difficult to check results	Google[8], Nominatim[9], Geoportail[10]
	Pose fusion	6-DoF pose	Images + 6-DoF poses	Require the pose of the images, difficult to adapt to heritage content	[41]
	CNN pose regression	6-DoF pose	Images + 6-DoF poses	Require the pose of the images, difficult to adapt to heritage content	[17], [6], [9], [20]
	CNN pose regression with 3D	6-DoF pose	Images + 6-DoF poses + 3D model	Require the pose of the images, require 3D model, difficult to adapt to heritage content	[50], [34]

heritage, whether it is by diving into the literature and tools coming from computer science or from digital humanities. Table 2 makes a synthesis of the various spatialization strategies, the outputs obtained and the specificities they may have. We resume below what we can learn from this study.

The main difference between the methods relies on the size of the spatial footprint to visit and naturally on the level of spatialization targeted. Quantitative (computational) and qualitative (with a strong human involvement) strategies are not in opposition, but rather complementary: while the former offer the possibility to process large volumes of data upstream, the latter allow to obtain precise and especially qualified results. Going further than the amount of images to process, the difference between quantitative and qualitative scenarios can also be found in the reference data required to spatialize the image. Indeed, a 3D model is essential to estimate a 6-DoF pose but this model can be of various types (LiDAR, 3D mesh, SfM point cloud, etc.). The precision of the pose estimated is correlated with the richness of the model (in terms of level of detail and localization precision). Currently, their spatial coverage tends to be inversely proportional to their richness, but this ratio is reducing with the implementation of massive and sophisticated national or private acquisition protocols. For the time being, compromises must be made: [5] for instance use in their platform a simplified 3D building model with a precision of about one meter but available on all of France (similarly to the one presented in first row of Figure 3). If more detailed models are available, their coverage is still limited (small areas for SfM point clouds and at the scale of the city and on demand for LiDAR ones), and their exploitation at large scale is far from being easy because of the volume and structure of such contents.

We have also seen that several levels of spatialization get along with each other. When considering iconographic heritage, the most widespread are probably the annotations and 2D positions directly linked with the historical manual spatialization approaches. Estimating a 6-DoF pose of an heritage iconographic content is a more recent target, which takes sense with the advent of modern 4D modelling approaches that are developing in the digital twins movement and the valorization of the big data of the past. There is still room for improvement in terms of maturity of the scientific proposals, but solutions like the semi-automatic pose estimation already provide relevant concrete results directly usable, even by non experts. Further pushing the crowdsourcing paradigm via platforms for semi-automatic pose estimation could also benefit from more multimodality. Simultaneously combining and visualizing even more data types (metadata, similar images, etc.) at one glance may lead to a more efficient use of those platforms, leading in turn toward the creation of datasets for automatic methods.

ACKNOWLEDGMENTS

This work is supported by ANR, the French National Research Agency, within the ALEGORIA project, under Grant ANR-17-CE38-0014-01 and ANRT - Ville de Paris - IGN under Grant CIFRE n°2019/1841.

REFERENCES

[1] Vasileios Balntas. 2019. SILDa. https://medium.com/scape-technologies/silda-a-multi-task-dataset-for-evaluating-visual-localization-7fc6c2c56c74

[2] Hayat Dino Bedru, Shuo Yu, Xinru Xiao, Da Zhang, Liangtian Wan, He Guo, and Feng Xia. 2020. Big networks: A survey. *Computer Science Review* 37 (2020), 100247. https://doi.org/10.1016/j.cosrev.2020.100247

[3] Tim Berners-Lee. 2006. Linked Data. https://www.w3.org/DesignIssues/LinkedData.html

[4] Nicolas Blanc, Timothée Produit, and Jens Ingensand. 2018. A semi-automatic tool to georeference historical landscape images. *PeerJ* 6 (2018), 1–7. https://doi.org/10.7287/peerj.preprints.27204

[5] E. Blettery, P. Lecat, A. Devaux, V. Gouet-Brunet, F. Saly-Giocanti, M. Brédif, L. Delavoipière, S. Conord, and F. Moret. 2020. A spatio-temporal web application for the understanding of the formation of the parisian metropolis. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 6, 4/W1 (2020), 45–52. https://doi.org/10.5194/isprs-annals-VI-4-W1-2020-45-2020

[6] Eric Brachmann and Carsten Rother. 2018. Learning less is more - 6D camera localization via 3D surface regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4654–4662.

[7] Tommaso Cavallari, Luca Bertinetto, Jishnu Mukhoti, P. Torr, and Stuart Golodetz. 2019. Let's Take This Online: Adapting Scene Coordinate Regression Network Predictions for Online RGB-D Camera Relocalisation. In *Proceedings - 2019 International Conference on 3D Vision, 3DV 2019*. 564–573. https://doi.org/10.1109/3DV.2019.00068 arXiv:1906.08744

[8] Gabriela Csurka, Christopher R. Dance, and Martin Humenberger. 2018. From handcrafted to deep local features. *arXiv* (2018), 1–41. arXiv:1807.10254

[9] Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, and Ping Luo. 2019. CamNet: Coarse-to-fine retrieval for camera re-localization. In *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 2019-Octob. 2871–2880. https://doi.org/10.1109/ICCV.2019.00296

[10] Basura Fernando, Tatiana Tommasi, and Tinne Tuytelaars. 2015. Location recognition over large time lags. *Computer Vision and Image Understanding* 139 (2015), 21–28.

[11] Martin A Fischler and Robert C Bolles. 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM* 24, 6 (jun 1981), 381–395. https://doi.org/10.1145/358669.358692

[12] A. Fusiello, E. Maset, and F. Crosilla. 2013. Reliable Exterior Orientation By a Robust Anisotropic Orthogonal Procrustes Algorithm. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XL-5/W1, February (2013), 81–86. https://doi.org/10.5194/isprsarchives-xl-5-w1-81-2013

[13] Emilio Garcia-Fidalgo and Alberto Ortiz. 2014. *State-of-the-Art in Vision-Based Topological Mapping and Localization Methods*. Technical Report. https://doi.org/10.13140/RG.2.2.33178.44484

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[15] Dimitri Gominski, Valérie Gouet-Brunet, and Liming Chen. 2021. Connecting Images through Sources: Exploring Low-data, Heterogeneous Instance Retrieval. *Remote Sensing Journal (MDPI), Special Issue "Digitization and Visualization in Cultural Heritage"* (2021).

[16] Richard Hartley and Andrew Zisserman. 2004. *Multiple View Geometry in Computer Vision* (2 ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511811685

[17] Alex Kendall, Matthew Grimes, and Roberto Cipolla. 2015. PoseNet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 2015 Inter. 2938–2946. https://doi.org/10.1109/ICCV.2015.336 arXiv:1505.07427

[18] Laurent Kneip, Davide Scaramuzza, and Roland Siegwart. 2011. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2969–2976. https://doi.org/10.1109/CVPR.2011.5995464

[19] W. Krüger. 2001. Robust and efficient map-to-image registration with line segments. *Machine Vision and Applications* (2001).

[20] Zakaria Laskar, Iaroslav Melekhov, Surya Kalia, and Juho Kannala. 2017. Camera Relocalization by Computing Pairwise Relative Poses Using Convolutional Neural Network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, Vol. 2018-Janua. 929—-938. https://doi.org/10.1109/ICCVW.2017.113 arXiv:1707.09733

[21] Vincent Lepetit, Francesco Moreno-Noguer, and Pascal Fua. 2009. EPnP: An Accurate O (n) solution to the PnP problem. *International Journal Of Computer Vision* 81 (2009), 155–166.

[22] Yansheng Li, Jiayi Ma, and Yongjun Zhang. 2021. Image retrieval from remote sensing big data: A survey. *Information Fusion* 67, April 2021 (2021), 94–115. https://doi.org/10.1016/j.inffus.2020.10.008

[23] V. Gouet-Brunet M. Khokhlova, N. Abadie and L. Chen. 2021. Learning embeddings for cross-time geographic areas represented as graphs. In *the 36th ACM/SIGAPP Symposium On Applied Computing (SAC 2021) - Technical Track Geographic Information Analysis*. 564–573.

[24] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. 2021. Image Matching from Handcrafted to Deep Features : A Survey. *International Journal of Computer Vision* 129, 1 (2021), 23–79. https://doi.org/10.1007/s11263-020-01359-2

[25] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 2017. 1 Year , 1000km : The Oxford RobotCar Dataset. *The International Journal of Robotics Research* 36, 1 (2017), 3–15.

[26] Ferdinand Maiwald, Jonas Bruschke, Christoph Lehmann, and Florian Niebling. 2019. A 4D information system for the exploration of multitemporal images and maps using photogrammetry, web technologies and Vr/Ar. *Virtual Archaeology Review* 10, 21 (2019), 1–13. https://doi.org/10.4995/var.2019.11867

[27] Carlo Masone and Barbara Caputo. 2021. A Survey on Deep Visual Place Recognition. *IEEE Access* 9 (2021), 19516—-19547. https://doi.org/10.1109/ACCESS.2021.3054937

[28] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. 2016. Openmvg: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*. Springer, 60–74.

[29] Nathan Piasco, Désiré Sidibé, Cédric Demonceaux, and Valérie Gouet-Brunet. 2018. A survey on Visual-Based Localization: On the benefit of heterogeneous data. *Pattern Recognition* 74 (2018), 90–109. https://doi.org/10.1016/j.patcog.2017.09.013

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

[31] Noé Pion, Martin Humenberger, Gabriela Csurka, Yohann Cabon, and Torsten Sattler. 2020. Benchmarking Image Retrieval for Visual Localization. In *International Conference on 3D Vision*. arXiv:2011.11946 http://arxiv.org/abs/2011.11946

[32] Filip Radenovic, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. 2018. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 5706–5715. https://doi.org/10.1109/CVPR.2018.00598 arXiv:1803.11285

[33] Remi Ratajczak, Carlos Fernando Crispim-Junior, Elodie Faure, Beatrice Fervers, and Laure Tougne. 2019. Automatic Land Cover Reconstruction from Historical Aerial Images: An Evaluation of Features Extraction and Classification Algorithms. *IEEE Transactions on Image Processing* 28, 7 (2019), 3357–3371. https://doi.org/10.1109/TIP.2019.2896492

[34] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. 2021. Back to the Feature : Learning Robust Camera Localization from Pixels to Pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3247—-3257. arXiv:arXiv:2103.09213v2

[35] 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. 2018. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8601—-8610. arXiv:arXiv:1707.09092v3

[36] Torsten Sattler, Chris Sweeney, and Marc Pollefeys. 2014. On sampling focal length values to solve the absolute pose problem. In *Proceedings - IEEE European Conference on Computer Vision*. 828–843. https://doi.org/10.1007/978-3-319-10593-2_54

[37] Grant Schindler and Frank Dellaert. 2012. 4D Cities: Analyzing, visualizing, and interacting with historical urban photo collections. *Journal of Multimedia* 7, 2 (2012), 124–131. https://doi.org/10.4304/jmm.7.2.124-131

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

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

[40] Xi Shen, François Darmon, Alexei A. Efros, and Mathieu Aubry. 2020. RANSAC-Flow: Generic Two-Stage Image Alignment. *Computer Vision–ECCV 2020. ECCV 2020. Lecture Notes in Computer Science* 12349 LNCS (2020), 618–637. https://doi.org/10.1007/978-3-030-58548-8_36 arXiv:2004.01526

[41] Yafei Song, Xiaowu Chen, Xiaogang Wang, Yu Zhang, and Jia Li. 2016. 6-DOF image localization from massive geo-tagged reference images. *IEEE Transactions on Multimedia* 18, 8 (2016), 1542–1554. https://doi.org/10.1109/TMM.2016.2568743

[42] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum. 2020. Learning and Aggregating Deep Local Descriptors for Instance-Level Recognition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 12346 LNCS. 460–477. https://doi.org/10.1007/978-3-030-58452-8_27 arXiv:2007.13172

[43] Hemanth Venkateswara, Shayok Chakraborty, and Sethuraman Panchanathan. 2017. Deep-Learning Systems for Domain Adaptation in Computer Vision: Learning Transferable Feature Representations. *IEEE Signal Processing Magazine*

34, 6 (2017), 117–129. https://doi.org/10.1109/MSP.2017.2740460

[44] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)* 53, 3 (2020), 1–34.

[45] Yu-Xiong Wang, Liangke Gui, and Martial Hebert. 2017. Few-Shot Hash Learning for Image Retrieval. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*.

[46] Li Weng, Valérie Gouet-Brunet, and Bahman Soheilian. 2020. Semantic Signatures for Large-scale Visual Localization. *Multimedia Tools and Applications* (May 2020). https://doi.org/10.1007/s11042-020-08992-6

[47] Tobias Weyand, André Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2 A large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2575–2584. https://doi.org/10.1109/CVPR42600.2020.00265 arXiv:2004.01804

[48] Changchang Wu et al. 2011. VisualSFM: A visual structure from motion system. (2011).

[49] Changchang Wu, Sameer Agarwal, Brian Curless, and Steven M Seitz. 2011. Multicore bundle adjustment. In *CVPR 2011*. IEEE, 3057–3064.

[50] Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, and Ping Tan. 2019. SANet: Scene agnostic network for camera localization. In *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 2019-Octob. 42–51. https://doi.org/10.1109/ICCV.2019.00013

[51] Xiwu Zhang, Lei Wang, and Yan Su. 2021. Visual place recognition: A survey from deep learning perspective. *Pattern Recognition* 113 (2021), 107760. https://doi.org/10.1016/j.patcog.2020.107760